Humanités numériques et science ouverte
Les collectifs

Clarisse Bardiot, Esther Dehoux, Émilien Ruiz (dir.)

La fabrique numérique des corpus en sciences humaines et sociales

Centrales pour toutes les disciplines relevant des arts, des lettres et des sciences humaines et sociales, les questions relatives à l'identification, la sélection, le classement, les modalités d'exploitation et de diffusion des matériaux nécessaires à la production de connaissances ne sont pas nées avec l'ère dite « numérique ». En histoire, pour prendre l'exemple qui nous est le plus familier, le rapport à la documentation fut ainsi d'emblée au cœur des réflexions méthodologiques qui ont accompagné la construction des savoirs historiques en discipline et l'émergence du métier d'historien. Dès 1898, dans leur *Introduction aux études historiques*, Charles-Victor Langlois et Charles Seignobos formalisent les opérations qui composent ——

Septentrion
PRESSES UNIVERSITAIRES

Ce que le numérique fait aux corpus. Introduction

Clarisse Bardiot et Émilien Ruiz

1. Centrales pour toutes les disciplines relevant des arts, des lettres et des sciences humaines et sociales, les questions relatives à l'identification, la sélection, le classement, les modalités d'exploitation et de diffusion des matériaux nécessaires à la production de connaissances ne sont pas nées avec l'ère dite « numérique ».

2. En histoire, pour prendre l'exemple qui nous est le plus familier, le rapport à la documentation fut ainsi d'emblée au cœur des réflexions méthodologiques qui ont accompagné la construction des savoirs historiques en discipline et l'émergence du métier d'historien. Dès 1898, dans leur *Introduction aux études historiques*, Charles-Victor Langlois et Charles Seignobos formalisent les opérations qui composent la « méthode historique », dont la « recherche des documents », mais aussi leurs conditions de réunion, de mise en ordre et de description systématique, font partie intégrante. Ils insistent alors tout particulièrement sur l'usage des fiches dont la mobilité « permet de les classer à volonté, en une foule de combinaisons diverses », moyennant un travail de référencement précis des sources originales, l'insertion de renvois d'une fiche à l'autre, etc. Les deux historiens distinguent alors « le cas de l'historien qui classe des documents » dans la perspective de ses propres recherches, avec sa propre méthode en sélectionnant les documents qu'il estime utiles à son étude ; de celui de l'érudit dévoué à la fabrication de *corpus* (collections de documents) et regestes (qui analysent et décrivent les *corpus*) destinés à un public plus large et nécessitant un cadre de classement adapté, ainsi que des index des noms, des index des dates, des tables d'*incipits*, etc. Langlois et Seignobos n'opposent pas pour autant ces pratiques : les historiens de métier ayant tout intérêt à « l'observation régulière » de pratiques qui ne peuvent que contribuer à rendre « plus aisés et plus solides les travaux d'histoire qui ont un caractère scientifique » (Langlois et Seignobos 1992). La refondation de la discipline par les *Annales* dans les années 1930-1940 ne remet pas véritablement en cause ces fondements mais elle contribue à élargir l'éventail des matériaux mobilisables. En 1933, dans sa leçon inaugurale au Collège de France, Lucien Febvre insiste : les sources de l'historien sont composées de « textes, évidemment : mais *pas rien que les textes*. Les documents aussi, quelle qu'en soit la nature » (Febvre 1953). Marc Bloch, dans *Apologie pour l'histoire* souligne leur « caractère merveilleusement disparate » pour insister sur l'importance des inventaires, catalogues, répertoires nécessaires à leur usage, tant pour rendre hommage à ceux qui composent de « pareils ouvrages » qu'afin d'affirmer l'importance de l'apprentissage de leur maniement (Bloch 1959). Dès lors, pour les fondateurs des *Annales*, il est indispensable de « négocier perpétuellement des alliances nouvelles »

entre disciplines, qu'elles soient « proches ou lointaines », qu'il s'agisse d'échanger des « notions » ou, surtout, des « méthodes », ce qui ne devait pas manquer de conduire au regroupement de « travailleurs d'éducation diverse s'unissant en équipes pour joindre leurs efforts » (Febvre 1953).

3. Revenons au début des années 2020, branchons un ordinateur, connectons-nous à Internet et accédons à la salle des inventaires virtuelle des archives nationales avec notre navigateur web, tout en sauvegardant, sur un disque dur externe ou dans le *cloud* des photographies de documents iconographiques, des enregistrements audios d'entretiens, pensons au passage à synchroniser les métadonnées de notre bibliographie sur les serveurs de Zotero. Entreprenons de croiser les informations de cet ensemble documentaire, interrogeons-nous sur la meilleure façon de l'organiser, de le décrire et de l'exploiter. Relisons maintenant le début de cette introduction. L'ensemble des questionnements que se posaient les fondateurs des sciences sociales aux tournants des XIX^e^ et XX^e^ siècles reste d'une frappante actualité. Mais le cadre est radicalement différent puisque, entre-temps, l'avènement des technologies permettant de convertir tout type d'information en données pouvant faire l'objet de traitements informatiques et d'une transmission quasi-instantanée, a donné naissance à ce que l'on appellera ici, par commodité, l'ère numérique.

4. Quelle que soit la discipline (art, lettres, histoire, sciences sociales…), nous sommes donc de plus en plus confrontés à des corpus et à des archives numériques, qu'ils soient numérisés ou nativement numériques. Pour étudier les phénomènes culturels, sociaux, politiques, économiques, nous nous appuyons sur des traces numériques, que celles-ci viennent du Web, des réseaux sociaux, des publications numériques ou encore de fichiers rassemblés sur des disques durs. Si le terme de « corpus » s'est généralisé ces dernières années, c'est dans une acception qui tend parfois à brouiller la distinction que pouvaient faire les fondateurs des sciences historiques. Bien souvent, on demande au chercheur, avant tout projet, analyse ou démonstration, de « définir son corpus ». Mais cet ouvrage collectif entend inscrire les réflexions sur les transformations récentes dans un contexte plus large : celui de l'histoire longue de la mise en données des SHS, bien antérieure à l'ère numérique. Ne serait-ce qu'en mentionnant le fait qu'un tournant majeur a eu lieu dans les années 1970, avec l'amorce d'une démocratisation de l'usage de l'informatique dans la recherche en sciences humaines et sociales, et plus largement d'une informatisation de la société. En histoire, cela a conduit à de profonds renouvellements de l'usage des méthodes quantitatives dont témoigne la création de revues telles que *Le Médiéviste et l'ordinateur* (Bourlet *et al.* 1979) et *Histoire & Mesure* (« Editorial » 1986). Dès son premier éditorial, cette dernière constate que « l'ordinateur a profondément transformé les pratiques de nombreux historiens ».

5. Certaines expérimentations dont les résultats furent publiés à l'aube des années 1970 sont ainsi à l'origine de pratiques que l'on associe volontiers aux humanités

numériques aujourd'hui. Ne prenons qu'un exemple en citant les recherches fondatrices d'Antoine Prost sur le vocabulaire des proclamations électorales. Mobilisant les méthodes de la linguistique et de la lexicologie au service d'une problématique pleinement historienne sur les mentalités politiques des années 1880, cette recherche fondée sur « un gros travail de collationnement et de perforation » pour le « passage en ordinateur » des « données » réunies, conduisit les historiens à collaborer avec une équipe de statisticiens du laboratoire de Jean-Paul Benzécri. Entreprise collective, cette recherche historique put ainsi tester « des méthodes nouvelles et encore incertaines » telles que… l'analyse factorielle des correspondances (Prost et Rosenzveig 1971 ; Prost 1974 ; à mettre en regard de Guaresi 2018).

6. L'irruption d'Internet, et surtout le développement du Web, ainsi que les vastes entreprises de numérisation menées, en particulier par les institutions patrimoniales dont l'objectif n'est pas seulement la recherche mais aussi la préservation et le rayonnement de leurs collections ainsi que l'élargissement de leurs publics, ont donné à ces questionnements une nouvelle ampleur depuis la fin des années 2000. Il faut toutefois se garder de lire ce mouvement comme la succession de deux périodes, celui du temps de la numérisation des corpus puis celui des corpus nativement numériques. D'abord parce que la numérisation se poursuit aujourd'hui et a encore de belles heures devant elle tant la tâche est colossale. À cela s'adjoint la nécessité de collecter et de préserver les traces nativement numériques, qui ont leur propre spécificité et font surgir de nouvelles problématiques de conservation dans les bibliothèques et centres d'archives ; qu'il s'agisse de formats, d'accessibilité, de pérennité, d'authenticité ou de stockage. L'élargissement du champ des possibles à l'ère numérique offre autant de ruptures que de continuités avec la période qui précède l'apparition du Web. Ensuite, parce qu'en certaines circonstances, il semble utile de considérer tout corpus numérique déjà constitué comme « nativement numérique », qu'il soit composé à partir de matériaux réellement nativement numériques (production textuelle, visuelle, audiovisuelle, etc. n'ayant pas d'existence analogique initiale) ou produit à partir de documents numérisés auxquels les utilisateurs ont accès sans en être les concepteurs.

7. En effet, aujourd'hui, les chercheurs ont de plus en plus accès à des corpus numérisés qu'ils n'ont pas constitués. Pour eux, de tels corpus sont donc, en quelque sorte, nativement numériques, ce qui n'est pas sans conséquences sur les modalités d'exploitations empiriques des matériaux ainsi réunis. L'ère numérique change donc en partie la donne, et parfois profondément, tout en posant des questions anciennes. Elle soulève notamment, de manière renouvelée, des interrogations sur l'accès, l'origine, le cadre juridique ou encore la fiabilité des sources. Ainsi celle de l'accès est-elle loin d'être résolue, malgré les encouragements à la publication des corpus, au développement de plans de gestion des données ou encore de la publication de données ouvertes selon les principes

FAIR (*Findable, Accessible, Interoperable, Reusable*)[1]. Comment, par exemple, accéder aux données des entreprises privées telles que Facebook ? Comment constituer un corpus de tweets au-delà des limites imposées par Twitter ? De surcroît, il faut aussi poser la question de l'exploitabilité de telles données : de quoi un corpus extrait de publications sur les réseaux sociaux numériques de ce type est-il réellement représentatif ?

8. Parfois, le seul fait d'accéder à des corpus autrefois presque inaccessibles peut être présenté comme la découverte d'un véritable eldorado. Alors qu'il fallait parfois une vie de chercheur pour constituer un corpus, la mise à disposition de ce travail permettrait ainsi aux générations suivantes de se confronter directement aux questions de traitement et d'analyse en faisant l'économie de la mise en données. L'exemple de la presse est ainsi très parlant ; compte tenu du temps nécessaire au dépouillement et à l'analyse d'un quotidien, national ou régional par exemple, il n'était pas rare d'en faire la source principale, voire unique, de mémoires universitaires. Aujourd'hui, les opérations de numérisation de la presse du XIX^e^ siècle et du début du XX^e^ siècle permettent d'en faire une source complémentaire à d'autres explorations documentaires puisqu'il est possible de procéder à des recherches plein texte à partir de mots-clés, sur des collections de titres nationaux, locaux, régionaux, politiques ou culturels. Ce n'est toutefois pas sans poser de questions méthodologiques : comment les titres ont-ils été choisis pour faire l'objet de cette mise en corpus ? La numérisation a-t-elle été réalisée en plein texte pour l'ensemble ? L'océrisation des contenus a-t-elle été contrôlée ? A-t-elle été corrigée ?

9. L'accessibilité aux corpus déjà constitués redouble ainsi les questions relatives à leurs modalités de fabrication. Qui en sont les auteurs (d'autres chercheurs, des bibliothécaires, des entreprises privées) et, de là, quels objectifs ont présidé à la sélection de ce qui serait ou non numérisé ? Quelles techniques d'indexation ont été appliquées ? etc. On retrouve ainsi des mises en garde anciennes qui ouvraient cette introduction. Comme le souligne Laura Putnam, si le tournant numérique a considérablement élargi et accéléré l'accessibilité aux documents, notre capacité à lire les sources avec précision, à les analyser et à en comprendre le sens « ne peut pas s'accélérer comme par magie » (Putnam 2016). En outre, on ne saurait négliger le fait que la mise en ligne de corpus peut aussi conduire à une forme de désintérêt de la part de chercheurs particulièrement attachés au caractère inédit de leur documentation, soucieux d'avoir la primeur d'un accès aux sources exploitées. Claire Lemercier rappelle ainsi que les efforts considérables qui ont été réalisés en matière de numérisation des sources de l'état civil ne se sont pas accompagnés d'une recrudescence de travaux relevant de la démographie historique ou de l'histoire sociale des populations (Lemercier 2014).

1. Pour plus d'informations sur les données FAIR, cf. https://doranum.fr/enjeux-benefices/principes-fair/.

10. À la croisée de questionnements ontologiques et épistémologiques, cet ouvrage prolonge la réflexion entamée à l'occasion du colloque DHNord 2019 organisé par la Maison européenne des sciences de l'homme et de la société (MESHS) de Lille, en partenariat avec le Centre de recherches interuniversitaires sur les humanités numériques (CRIHN) de Montréal. L'un des enjeux actuels est celui de l'évolution de la définition du terme corpus, et des termes qui lui sont associés, tels que collections, archives, données, sous l'impulsion des mutations engendrées par le numérique. Si l'on s'en tient à une définition actuelle et générique, un corpus, qu'il soit ou non numérique, est un ensemble d'éléments circonscrits (ces éléments pouvant être de différentes natures, des manuscrits aux données), assemblés en fonction d'un projet de recherche (ce qui sous-tend un cadre méthodologique, épistémologique et herméneutique), au contraire de la collection, qui appartient à une institution et dont la raison d'être n'est pas toujours, ou pas uniquement, la recherche. Une collection peut susciter une infinité de corpus. De même, un corpus peut rassembler des items en provenance de différentes collections – et c'est d'ailleurs sans doute là l'une des grandes promesses du numérique : rassembler sur son écran d'ordinateur, à l'intérieur d'un dossier virtuel sur son disque dur ou sur un *cloud*, des éléments de différentes provenances, numérisées ou nativement numériques, qu'il serait bien difficile, sinon impossible, de réunir dans un espace réel. Le choix d'intégrer ou d'exclure un élément dans un corpus procède toujours d'une mise en regard des autres éléments qui y sont ou non déjà présents, tout en étant intimement lié aux questions de recherche posées par celui ou celle qui le compose. Le seul fait d'assembler des fragments permet de contextualiser chacun d'entre eux au regard des autres. Série d'éléments non linéaires, le corpus se prête particulièrement bien aux pratiques hypertextuelles. Paradoxalement, avec le numérique et les possibilités d'augmentation, d'agrégation qu'il offre, notamment avec le Web sémantique, le corpus semble aujourd'hui infini, au risque de rendre le concept même inopérant.

11. Si cette introduction est rédigée par deux historiens, qui ont donc un rapport au corpus spécifique, cet ouvrage souhaite interroger de façon transversale l'impact du numérique sur les corpus et archives en SHS, de leur constitution, de leur fiabilité ainsi que de leur enrichissement, jusqu'à l'enjeu de leur accessibilité et de leur publication. L'ère numérique change notre rapport aux corpus et aux archives, mais quels sont les contours de cette transformation ? Tel est l'objet de ce recueil de textes (de ce corpus de contributions scientifiques serait-on tenté de dire) qui interroge ce que le numérique fait au corpus et aux archives. La question est d'autant plus importante que ces derniers sont de plus en plus accessibles, tout en étant le fruit de toute une succession de manipulations que l'on peut regrouper en trois grandes étapes : constitution, traitement et interprétation. Même si les frontières sont loin d'être étanches, ce volume est essentiellement consacré à la première étape. Celle-ci est traversée de trois opérations fondamentales, la numérisation, la description et la publication, qui structurent les trois grandes parties de cet ouvrage. Quels modes

d'accès et de lecture des sources le numérique permet-il face à des corpus soulevant des problèmes de taille ou de contraintes juridiques ? Quels sont les enrichissements possibles, que ce soit de manière automatique ou manuelle ? Pour une recherche basée sur des hypothèses ou sur des données (*hypothesis driven versus data driven*) ? Afin de privilégier une perspective quantitative ou qualitative ? Peut-on maîtriser son corpus dans un contexte numérique ? Quels sont les apports ainsi que les limites de l'édition numérique de corpus ? Et enfin, quels invariants et particularités observe-t-on en régime numérique face à la diversité des corpus ?

12. La constitution de corpus numériques, à commencer par la numérisation, implique de nombreux choix qui auront une incidence, non seulement sur les éléments constitutifs du corpus, mais aussi de manière plus large sur son accessibilité, ses usages, sa publication, voire sa réutilisation. Ces questions suscitent aujourd'hui de nombreux débats dans et entre les disciplines. L'ouvrage entend ouvrir des pistes de réflexion à destination des chercheurs, doctorants et étudiants, en faisant dialoguer différents initiatives et projets de recherche récents dans le champ des humanités numériques ainsi que divers retours d'expérience permettant d'interroger nos pratiques professionnelles dans l'archivage numérique et la recherche en SHS. Que signifie constituer un corpus à l'ère numérique ? Quelles pluralité et complémentarité d'acteurs retrouve-t-on ? Loin de faire consensus, ces questions portent en elles une définition même de la manière dont la recherche se pratique aujourd'hui.

13. Comme le rappelle Juliette De Maeyer[2], toute une nouvelle génération de chercheurs n'a « jamais connu de corpus ou d'archives en dehors du numérique » ou encore, comme le souligne Damon Mayaffre[3], « tous les linguistes de corpus travaillent aujourd'hui sur corpus numériques ». « Le goût du corpus numérique »[4] scrute nos « routines numériques "discrètes" », les signaux faibles qui ont le sceau de l'évidence, ceux de la pratique concrète du chercheur au travail transformée par la nouvelle « boîte à outils » à disposition qui pourtant nous conduisent à repenser nos pratiques en profondeur. Car numériser, ce n'est pas uniquement faciliter l'accès. C'est d'abord une succession de transformations, de remédiations, qui peuvent y compris amener à considérer l'OCR, comme le fait Juliette De Maeyer « comme une nouvelle édition du texte » et dont il importe de documenter le processus, les méthodes, les choix, ou encore les freins, en particulier juridiques. Numériser, ce n'est pas créer un « reflet fidèle » de l'artefact numérisé (que ce soit un tableau, une estampe ou un texte), ni du fonds dont il provient – bien souvent il s'agit d'une sélection dont le choix ne relève pas toujours de logiques scientifiques[5]. C'est augmenter, enrichir les informations sur

2. Voir son chapitre dans cet ouvrage : « Traverser les corpus de presse numériques : un travail d'artisan ? ».
3. Voir son chapitre dans cet ouvrage : « Les corpus numériques textuels (re)spécifiés ».
4. Voir le chapitre éponyme de Frédéric Clavert et Caroline Muller dans cet ouvrage.
5. Voir le chapitre d'Odile Gaultier-Voituriez dans cet ouvrage : « *Archelec*, les archives électorales françaises de la Ve République, du papier au numérique : reflet fidèle ou distorsion ? »

l'objet de départ[6], comme le formalise le « millefeuille informationnel » d'Antoine Courtin[7] ; ce qui, parfois, ne va pas sans une certaine forme de bricolage[8]. Ce travail d'enrichissement se poursuit lors de l'analyse du corpus, par exemple lors de la comparaison entre différents items : l'un des enjeux actuels consiste à pouvoir reverser ces nouvelles connaissances issues du travail de la recherche vers les collections qui ont permis la constitution du corpus[9]. Cela implique éventuellement d'en passer par une formalisation et la création d'ontologies[10] afin de décrire précisément un domaine ou encore de favoriser l'interopérabilité. Car les corpus ne sont pas toujours homogènes, loin s'en faut, et l'exploitation de corpus hétérogènes soulève des questions spécifiques[11]. De même pour les corpus nativement numériques, et en particulier les archives du Web[12], confrontés à l'épineuse question de l'authenticité des documents, d'où l'appel à une diplomatique numérique[13].

14. La réutilisation du corpus fait désormais partie des questions qui sont posées d'entrée de jeu, avant même sa constitution, tout comme l'intégration et le respect de normes et de standards, en particulier les principes FAIR ou encore les formats de description des objets du corpus. Les TGIR (très grandes infrastructures de recherche) jouent à cet endroit un rôle fondamental[14]. Comme le montre Emmanuel Château-Dutier pour les institutions muséales[15], l'enjeu de l'ouverture est non seulement une question technique et juridique mais aussi sociale et politique.

15. De la « fièvre corpus[16] » au « tout est corpus[17] », le phénomène n'est pas sans conséquences sur les sujets et objets de recherche : ce qui est numérisé, ce qui est accessible, a aujourd'hui bien plus de chance de « faire corpus », et

6. Voir le chapitre de Solenn Huitric dans cet ouvrage : « Enrichir un corpus de sources numérisé en histoire de l'éducation. Le cas du *Bulletin administratif de l'instruction publique* ».
7. Voir son chapitre dans cet ouvrage : « Pour un regard à 360 degrés sur les corpus visuels : pratiques de mise à disposition et de réutilisation ».
8. Voir le chapitre de Juliette De Maeyer dans cet ouvrage : « Traverser les corpus de presse numériques : un travail d'artisan ».
9. Voir le chapitre de Johanna Daniel : « Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique » et celui d'Antoine Courtin dans cet ouvrage.
10. Voir le chapitre d'Éric Kergosien et Mathilde Wybo dans cet ouvrage : « Les technologies du Web pour la valorisation d'un patrimoine industriel textile en mouvement dans les Hauts-de-France » ; ainsi que celui d'Amélie Daloz : « Méthodologie de validation et d'enrichissement d'une ontologie minière fondée sur le CIDOC CRM ».
11. Voir le chapitre d'Alix Chagué, Manuela Martini, Victoria Le Fourner et Éric Villemonte de la Clergerie dans cet ouvrage : « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non-uniforme ? ».
12. Voir le chapitre de Francesca Musiani dans cet ouvrage : « Archivage du Web, un enjeu de gouvernance (d'Internet) ».
13. Voir le chapitre de Marie-Anne Chabin dans cet ouvrage : « La méthode diplomatique face à l'information numérique ».
14. Voir le chapitre de Céline Alazard, Jean Vigreux et Serge Wolikow dans cet ouvrage : « Le traitement numérique des sources : la construction des corpus et des instruments de recherche comme enjeu pour la mise à disposition des données » ; ainsi que celui de Renaud Limelette : « Du corpus archivistique au corpus numérique : les soubassements du Web sémantique. L'exemple des sources relatives au parlement de Flandre ».
15. Voir son chapitre dans cet ouvrage : « Le musée comme service d'information. Pour une politique des interfaces muséales ».
16. Voir le chapitre de Damon Mayaffre dans cet ouvrage : « Les corpus textuels numériques (re)spécifiés ».
17. Voir le chapitre d'Antoine Courtin dans cet ouvrage : « Pour un regard à 360 degrés sur les corpus visuels : pratiques de mise à disposition et de réutilisation ».

de devenir objet de recherche, que ce qui ne l'est pas, y compris pour des chercheurs issus de disciplines a priori éloignées de la « cible » initiale[18] ; ou comment les politiques de numérisation des institutions influent sur la recherche, ses sujets, ses objets, ses méthodes. Autrement dit, il faut aussi avoir conscience des biais de la numérisation à l'encontre des objets d'études ; bien souvent les chercheurs qui commencent par l'étude d'un corpus numérique poursuivent leur projet aussi en le complétant par un corpus non numérique qui pour rejoindre le premier corpus devra à son tour être numérisé. Le corpus est un organisme vivant : il peut s'étendre, se scinder, être augmenté, mis à jour, dans une relation interdépendante avec les questions de recherche[19]. Celles-ci découlent autant du corpus qu'elles permettent de le constituer. Le corpus est une construction pour l'interprétation, ne serait-ce qu'au travers des premières étapes de saisie et de codage (Lemercier et Zalc 2008). Comme le rappelle Damon Mayaffre, il est « un lieu, lui-même construit, où s'échafaude le sens, où se scénarise l'interprétation[20] ».

16. En filigrane, les auteurs de cet ouvrage, au travers des différentes expériences qu'ils ont menées, ne cessent de rappeler que constituer un corpus c'est se positionner avant tout sur ce qu'est la recherche. Comment « configurer » le corpus pour qu'il puisse être cet organisme vivant qui évolue avec les questions de recherche ? Comment faire en sorte qu'il ne se fige pas et qu'il reste bien un corps, organique, vivant, évolutif ? L'un des enjeux dans le cadre de projets d'envergure comme *ENCCRE* sur l'*Encyclopédie* de Diderot et D'Alembert est alors que la constitution, la publication numérique du corpus et la création d'interfaces de consultation et de travail *ad hoc*, puissent permettre à tout chercheur d'y poser ses propres questions de recherche. Ce qui importe alors c'est non pas la versatilité du corpus en tant que tel mais son interface : les corpus interfacés doivent permettre la coexistence de différentes démarches scientifiques (de ce point de vue on pourrait considérer qu'ici le corpus se rapproche de la collection). L'objectif est que le « laboratoire virtuel de recherche » ainsi constitué grâce au corpus interfacé suscite de « nouvelles dynamiques de recherche[21] ».

17. Certes les corpus et archives numériques induisent souvent un changement d'échelle, mais ce n'est pas un critère suffisant comme le montre Frédéric Glorieux[22]. Ils entraînent une évolution des méthodes de travail qu'il importe d'examiner en tant que telle. Les chercheurs sont amenés à acquérir de nouvelles compétences, en

18. Voir le chapitre d'Agathe Sanjuan dans cet ouvrage : « Programme des registres de la Comédie-Française : un corpus numérique en extension » ; ainsi que celui d'Andrea Del Lungo et Karolina Suchecka : « Le projet *eBalzac* : construire une bibliothèque hypertextuelle des sources intellectuelles ».

19. Voir le chapitre de Solenn Huitric dans cet ouvrage : « Enrichir un corpus de sources numérisé en histoire de l'éducation. *Le cas du Bulletin administratif de l'instruction publique* ».

20. Voir son chapitre dans cet ouvrage : « Les corpus textuels numériques (re)spécifiés ».

21. Voir le chapitre d'Alexandre Guilbaud dans cet ouvrage : « L'*Édition numérique collaborative et critique de l'Encyclopédie de Diderot et D'Alembert (ENCCRE)*, comme prototype d'un laboratoire virtuel de recherches sur l'*Encyclopédie* et les Lumières ».

22. Voir son chapitre dans cet ouvrage : « Le corpus de tous les livres depuis les débuts de l'imprimerie, tous comptes faits… »

particulier sur le numérique en tant que média[23] ou sur les « bases des pratiques métiers[24] », mais aussi comme nombre des chapitres qui constituent cet ouvrage le montrent, à travailler au sein d'équipes multidisciplinaires (archivistes, conservateurs des bibliothèques ou des musées, informaticiens, ingénieurs d'études, chercheurs d'autres disciplines). Ce sont très souvent des projets longs, ce qui pose des questions institutionnelles, notamment de financement récurrent. D'un point de vue général, cette dernière dimension est d'autant plus prégnante que se pose aussi la question, éthique et citoyenne, de l'opportunité qu'il y aurait à accorder des financements publics à des projets pharaoniques, perçus comme innovants parce que numériques, mais dont l'utilité sociale et scientifique n'est pas toujours évidente.

18. Si les mutations abordées par les contributions de cet ouvrage affectent les pratiques de recherches en sciences humaines et sociales, elles constituent aussi un véritable défi pour la formation des étudiants de toutes les disciplines concernées. Pour conclure l'ouvrage, nous avons choisi d'aborder cet enjeu à partir de réflexions historiennes sur l'enseignement des humanités numériques et la façon dont nous préparons collectivement les futures générations de chercheuses et de chercheurs à faire face aux transformations à l'œuvre aujourd'hui[25].

23. Voir le chapitre de Juliette De Maeyer dans cet ouvrage : « Traverser les corpus de presse numériques : un travail d'artisan ».

24. Voir le chapitre de Johanna Daniel dans cet ouvrage : « Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique ».

25. Voir le chapitre d'Émilien Ruiz dans cet ouvrage : « Former "au numérique" en sciences humaines et sociales ? Propositions d'un historien ».

Constituer des corpus, du papier au numérique

Traverser les corpus de presse numériques : un travail d'artisan ?

Juliette De Maeyer

Introduction

1. Que fait le numérique aux corpus et aux archives ? Cette question peut sembler difficile à évaluer pour quelqu'un qui, comme moi, a vécu ses premières incursions dans le milieu de la recherche dans les années 2000. Disons-le clairement : je n'ai jamais connu de corpus ou d'archives en dehors du numérique. Il y a toujours eu, quelque part, une base de données, un écran, un moteur de recherche ou un logiciel de gestion des références bibliographiques. Comment prendre la mesure de ce que le numérique fait à nos pratiques de recherche, quand on baigne à ce point dedans ?

2. On voit toutefois un certain changement d'échelle, au cours des dix dernières années : les bases de données s'étoffent, les collections numérisées ne cessent de croître, les outils et les savoir-faire des humanités numériques s'organisent, l'infrastructure numérique de recherche se solidifie… Des indices qui laissent penser que le rapport des chercheurs en sciences humaines et sociales au numérique se professionnalise. C'est dans ce contexte que je souhaite, dans cet article, mettre en valeur les embûches et les mérites des corpus artisanaux, c'est-à-dire ceux que l'on bricole à petite échelle, que l'on bidouille avec les moyens du bord.

3. Le bricolage des corpus numériques est une approche explicitement revendiquée par certaines chercheuses et certains chercheurs, par exemple pour pallier l'incomplétude de certaines sources (Lécossais et Quemener 2018) ou pour appréhender les contours d'une interface de programmation (Clavert 2017). En m'appuyant sur des exemples tirés de mon propre sous-champ disciplinaire, l'histoire des médias et du journalisme, je montrerai que les approches artisanales ne se font pas toujours par défaut (en l'absence de connaissances techniques sophistiquées, ou de sources numériques accessibles facilement), mais qu'elles peuvent avoir une valeur en soi et révéler des choses devenues invisibles dans les corpus industriels (j'adopterai ce mot pour désigner le contraire des corpus artisanaux). En tirant parti de deux projets de recherche récents, je tâcherai de montrer que les corpus artisanaux résistent, qu'ils sont parfois remplis d'erreurs, difficiles à dompter. Cela conduit de temps à autre à des constats d'échec (qu'il est important de raconter), mais c'est peut-être aussi la grande qualité des corpus artisanaux : en nous obligeant à passer à travers de nombreuses étapes de remédiation et de transformation, ils nous rappellent de penser la matérialité des médias, et ses conséquences épistémologiques.

Corpus de presse numérisés

4. De nombreuses communautés de recherche ont largement bénéficié des diverses initiatives de numérisation de journaux, anciens ou plus récents. Les corpus de presse s'offrent désormais de manière beaucoup plus accessible, et avec la reconnaissance optique de caractères*[1] s'ouvre aussi la possibilité de travailler le texte de manière renouvelée, et de « feuilleter la presse ancienne par Giga Octets » (Gaillard 2018). La somme accumulée des campagnes de numérisation de la presse menées par les bibliothèques patrimoniales, par les musées ou par certaines entreprises privées permet actuellement aux chercheurs qui s'intéressent au journalisme de travailler sur des corpus de grande ampleur, avec des méthodes et des outils qui font la part belle à la « lecture distante », c'est-à-dire des méthodes et des outils computationnels qui permettent de révéler des récurrences, des schémas, des tendances dans de grands corpus de textes – transformés en « données » analysables, quantifiables, et traitables par des algorithmes* (Moretti 2005). Parmi les initiatives ambitieuses qui tirent parti de grandes quantités de journaux numérisés, on peut entre autres citer l'ANR *Numapresse* (Langlais 2018) ou le projet *Viral Text* (Cordell et Smith 2017).

5. Toutefois, lorsqu'elle n'a pas lieu de manière systématique, la numérisation des corpus de presse provoque inévitablement des effets, voire des distorsions. Au Canada, par exemple, il n'y a pas de projet centralisé et systématique de numérisation de la presse qui serait similaire, dans son ampleur, au travail effectué par Gallica en France. La numérisation se fait alors au gré d'initiatives dispersées (Kheraj 2014). Conséquence inévitable : certains titres sont plus numérisés que d'autres, et l'accès facilité à ceux-ci va de pair avec une visibilité amplifiée, parfois déséquilibrée. Ian Milligan a ainsi montré que les journaux canadiens qui ont été numérisés et mis à disposition dans des bases de données au début des années 2000 ont gagné une présence accrue dans les thèses de doctorat en histoire. Par exemple, le *Globe and Mail*, un quotidien basé à Toronto à la distribution nationale, était mentionné 58 fois dans les 67 thèses en histoire publiées en 1998, telles que recensées dans la base de données Proquest. Ce nombre a bondi à 708 dans les 69 thèses de 2010, et une croissance comparable est visible dans les travaux de recherche publiés par la *Canadian Historical Review* (Milligan 2013). D'autres journaux, d'importance ou de portée similaire, ne connaissent pas cette croissance fulgurante : les mentions de ceux dont les versions numérisées ne sont pas aisément disponibles en ligne restent stables à travers la même période.

6. Certains corpus échappent donc encore à la numérisation à grande échelle, et nous aurions tort de les délaisser simplement parce qu'ils sont numériquement inexistants ou algorithmiquement inaccessibles. Quand on ne veut pas se laisser dicter ses objets par les priorités de numérisation d'institutions sur lesquelles nous avons

1. Les mots et expressions suivis de * renvoient au glossaire de fin d'ouvrage.

très peu de contrôle, nous sommes alors parfois amenés à numériser nous-mêmes, en bricolant (Blandin et Garcin-Marrou 2018).

Un bricolage raté

7. C'est la voie que j'ai essayé d'emprunter dans un projet qui porte sur l'évolution du copier-coller dans la presse québécoise du XX^e^ siècle, et qui avait pour but d'identifier les fragments de texte reproduits d'un journal à l'autre (qu'il s'agisse de dépêches d'agences de presse, de communiqués et de déclarations de figures publiques, ou d'emprunts directs à des concurrents). Je voulais retracer ce phénomène – déjà bien documenté pour la presse du XIX^e^ siècle (Cordell 2015 ; Schuh 2017) – dans la presse moderne et de plus en plus professionnalisée (Schudson 1981 ; Charron et De Bonville 2004) du XX^e^ siècle, sur un terrain qui n'avait pas encore été beaucoup défriché, à savoir les journaux québécois.

8. La presse au Québec a fait l'objet de campagnes de numérisation, notamment par Bibliothèque et Archives nationales du Québec (BANQ), mais celles-ci sont encore en cours et les collections numérisées ne le sont que partiellement. J'avais donc échafaudé la stratégie suivante : numériser un échantillon choisi des journaux, à raison d'une semaine (sélectionnée aléatoirement) par décennie, entre 1920 et 1990. Une ponction raisonnée, bien plus modeste que l'échelle des données « massives », mais néanmoins suffisante pour se prêter à l'exercice de la lecture distante – le repérage de fragments de texte reproduits dans différents journaux est typiquement le genre d'exercice pour lequel nos capacités humaines ne suffisent pas et qui montre bien la complémentarité et la valeur ajoutée de certaines analyses computationnelles (Smith, Cordell et Mullen 2015).

9. Poursuivre ce travail de numérisation *ad hoc*, c'est aussi expérimenter à quel point la numérisation ne va pas de soi. Entre le journal et le fichier numérisé que je peux traiter avec des méthodes algorithmiques, il y a une suite d'étapes, de passages et de transformations non négligeables, qu'on ne voit pas (et c'est très heureux) quand on utilise les données déjà numérisées du journal *comme si c'était le journal*.

10. Quelles sont ces transformations ? D'abord, il faut souligner que le journal lui-même brille par son absence : l'objet journal, avec ses pages de papier fragiles et usées par le temps, son encre qui tache les doigts et son format encombrant, je n'en ai touché aucun. Les archives de presse sont souvent conservées sous forme de microfilms, et les ressources que j'avais à disposition pour élaborer mon échantillon étaient les collections de journaux sur microfilms de l'université de Montréal. Il a donc fallu convaincre les bibliothécaires et conservateurs d'accepter de me confier certaines précieuses bobines, afin de pouvoir les apporter à une société spécialisée dans la numérisation de documents. Un an plus tard (mes quelques bobines de microfilm n'étaient apparemment pas prioritaires pour cette société de numérisation qui a

plutôt l'habitude de faire affaire avec des ministères, des compagnies d'assurance ou des hôpitaux), je récupérais une dizaine de DVD qui contenaient chacun des fichiers PDF – un fichier par bobine de microfilm, ce qui correspond à 150 à 200 pages de journal. Voilà un découpage physique et temporel radicalement différent de celui de l'objet-journal (un numéro par jour), et même de ce que j'avais imaginé comme l'échantillon idéal pour mon étude (une semaine de publication).

11. Arrêtons-nous un instant sur ce matériau. Chaque page du fichier PDF est une image d'une image, un fac-similé d'un fac-similé, la reproduction numérique d'une reproduction analogique d'une page du journal. Mais ce n'est pas uniquement une image d'image, puisque le fichier PDF contient aussi une couche de texte, c'est-à-dire l'interprétation de l'image faite par un logiciel de reconnaissance optique de caractères (OCR) qui convertit l'image d'un texte en une chaîne de caractères compréhensible par un ordinateur. Voilà donc, aux termes de toutes ces transformations, le journal métamorphosé en texte interprétable par une machine. Lisible par une machine, certes, mais au prix d'une transformation en quelque chose d'illisible et d'inintelligible. Car, en ouvrant ces fichiers PDF, il avait bien fallu que je me rende à l'évidence : la plupart étaient simplement inutilisables, l'OCR ayant produit plus de bruit que de chaînes de caractères reproduisant effectivement les mots imprimés dans le journal (figure 1).

12. Que l'OCR ne soit pas parfait n'était pas une surprise. Nous savons que la reconnaissance optique de caractères

Figure 1. Erreurs d'OCR : extrait de la version océrisée du journal *Le Droit*, 28 mai 1921

Crédit : Juliette De Maeyer

est une technique qui produit des erreurs. C'est particulièrement le cas pour les journaux qui posent des défis supplémentaires (Langlais 2019 ; Holley 2009) à cause de leur structure éditoriale en colonnes, des espacements variables, des polices rares, ou de l'état de conservation du papier ou des microfilms. Dans leur rapport sur la reconnaissance optique de caractères des documents historiques, Smith et Cordell (2018, 5) avancent que la proportion de mots erronés dans les corpus de jour-

naux numérisés du XIXe siècle, en anglais, peut « excéder 40 % », et que c'est pire encore pour les sources plus anciennes ou dans d'autres langues. Face à cela, une réaction typique des chercheurs est simplement d'abandonner les corpus pour lesquels l'OCR est trop « sale » laissant donc des pans entiers de la recherche en friche.

@franklinfordbot : tirer parti des erreurs d'OCR

13. Dans mon projet portant sur l'évolution du copier-coller dans la presse, les transformations liées à la numérisation et les erreurs d'OCR sont un problème, quelque chose qu'il faudrait résoudre[2] pour pouvoir procéder à l'analyse. À peu près au moment où je butais contre un mur de bruit d'OCR dans ce corpus, j'entamais un autre chantier dans lequel les mêmes éléments – transformations matérielles liées à la numérisation, erreurs dans la reconnaissance du texte – allaient devenir productives, et un terrain fructueux pour penser la relation entre histoire et études des médias.

14. Cet autre projet, mené en collaboration avec Dominique Trudel[3], porte sur l'œuvre et la vie de Franklin Ford, un journaliste et théoricien des médias américain, né au Michigan en 1849 et mort à New York en 1918. Parmi les raisons pour lesquelles ce personnage, jusque là plutôt marginal et obscur, est digne d'intérêt, il y a son rôle dans la naissance de la recherche sur la communication et les médias aux États-Unis (Trudel et De Maeyer 2017), son inscription dans le développement du pragmatisme, notamment par son association avec le philosophe John Dewey dans les années 1890, ainsi que sa vision avant-gardiste (et jamais réalisée) d'une réforme complète du paysage médiatique, vision qui était à la fois technologique et politique. Les théories de Ford, pour les résumer très brièvement, avaient pour objectif de réorganiser la collecte, le traitement et la distribution des nouvelles, de manière à ce que chaque information, chaque « fait », puisse rejoindre son public au moyen des technologies adéquates (telles que le télégraphe, le téléphone, ou encore le chemin de fer). Sous l'appellation de « triangle de l'intelligence », Ford envisageait trois façons de produire, centraliser et distribuer les informations :

- l'information d'intérêt public à un public général, grâce à plusieurs journaux quotidiens
- l'information spécialisée à des secteurs d'activités économiques spécifiques, grâce à des publications portant sur certains domaines ou secteurs
- l'information personnalisée grâce à des bureaux d'information auprès desquels des individus pourraient obtenir des rapports en fonction de leurs besoins en information

2. Au moment d'écrire ces lignes, le « problème » n'est toujours pas résolu mais la vivacité des débats et de la recherche autour de l'OCR (Holley 2009 ; Gupta, Jacobson et Garcia 2007 ; Magallon, Béchet et Favre 2018) et la mobilisation des communautés de chercheurs en humanités numériques autour de ces questions me donnent bon espoir de réussir un jour à extraire quelque chose de ces documents.

3. De nombreux aspects du présent article doivent beaucoup au travail de Dominique Trudel, je lui suis donc grandement redevable.

15. À terme, c'est une utopie de « gouvernement par l'information » que Ford imaginait : une fois son système mis en place, la société se régulerait d'elle-même car la bonne information se trouverait toujours au bon endroit, auprès des bonnes personnes.

16. Pour explorer la contribution exacte de Ford à différents champs (histoire de la recherche en communication, histoire du pragmatisme, histoire des utopies informationnelles), il fallait donc rassembler son œuvre et en délimiter les contours – ce qui n'avait pas été fait auparavant, l'intérêt que portent les historiens des médias et du journalisme à Ford étant généralement limité à un épisode précis des années 1890 (Pinter 2003). C'est donc là que la question de l'archive et du corpus surgit : nous rassemblons des documents hétéroclites, il y a là-dedans des articles de presse, des livres, des opuscules publiés à compte d'auteur, beaucoup de lettres issues de la correspondance de Ford avec divers intellectuels, politiciens ou journalistes. Le corpus est disséminé dans de nombreux lieux différents, virtuels et physiques. Alors nous fouillons des bases de données de périodiques et des catalogues de bibliothèques (d'abord à l'aveugle, puis en suivant des traces de plus en plus précises, au fur et à mesure que nous identifions des éléments de sa biographie et de son parcours professionnel). Quand les instruments de recherche sont bien faits, nous convainquons parfois de très aimables bibliothécaires et documentalistes de numériser quelques pages d'un dossier précis. Le « goût de l'archive à l'ère numérique » (Clavert et Muller 2017) passe aussi par de nombreux courriels échangés avec des archivistes et des documentalistes, et par le remplissage de formulaires pour demander des reproductions, qui nous arrivent par mail quelques jours ou semaines plus tard. En quelques rares occasions, nous avons plongé les mains dans les boîtes et les dossiers, et pu dépouiller l'archive au sens de Farge (1997).

17. Les objectifs de ce projet, tels que je les ai exposés jusqu'ici, s'accommodent très bien d'une approche de « lecture rapprochée », basée sur un travail de constitution d'une archive et de dépouillement des documents. Mais ce n'est pas tout : il y a aussi une expérimentation méthodologique, qui est une mise en abîme des théories de Ford. Puisque celles-ci portent sur la circulation de l'information et le rôle des (nouvelles) technologies (de la fin du XIX[e] siècle) dans la « socialisation » des faits au sein de la société, nous tentons de mettre en pratique les théories de Ford, et de faire circuler son œuvre, en utilisant les technologies d'aujourd'hui. Cela prend la forme de bots[4], des programmes qui permettent de publier de manière automatisée des contenus sur des plateformes numériques, telles que Twitter ou Reddit. Chaque bot a pour but d'incarner une des branches du « triangle de l'intelligence » de Ford (information générale, spécialisée et personnalisée), et publie donc des extraits de l'œuvre de Ford, ce qui constitue une nouvelle éditorialisation de celle-ci.

18. Ainsi, @franklinfordbot (De Maeyer et Trudel 2018) est en fonction depuis mars 2017 et correspond à la

4. Cf. https://twitter.com/franklinfordbot et http://www.franklinford.org/

branche généraliste du triangle. Il diffuse des extraits de l'œuvre de Ford sur Twitter, à destination d'un large public. Le bot publie de manière automatisée des phrases extraites des écrits de Ford, choisies au hasard, à des intervalles également aléatoires. Ces tweets, et l'intervention du hasard, ont une valeur heuristique : ils sont de nouveaux points d'entrée dans l'œuvre de Ford, qui attirent souvent notre attention sur des aspects de l'archive qu'une lecture linéaire n'aurait pas nécessairement remarqués (De Maeyer et Trudel 2018). Ils ont aussi une valeur pédagogique et de mobilisation des connaissances, en « socialisant » l'œuvre de Ford sur une plateforme grand public.

19. Cette expérimentation qui consiste à transformer l'archive en tweets est également une réflexion pratique sur l'ancien et le nouveau, une des préoccupations des historiens des médias (Gitelman 2006 ; Lesage et Natale 2019) ainsi que de l'archéologie des médias (Parikka 2018 ; Mak 2014) qui invite à resituer nos médias actuels dans des temporalités élargies. La fascination collective pour la catégorie des « nouveaux » médias nous fait parfois oublier que même les « vieux » médias ont commencé par être « nouveaux ». Les tweets de @franklinfordbot sont donc une intervention qui met en exergue la permanence de cette catégorie : ils parlent de « nouveaux » médias (le télégraphe, le téléphone) sur un « nouveau » médium (Twitter), faisant se télescoper près d'un siècle de « nouveauté ».

Une expérience de remédiation

20. La transformation de l'archive en tweets devient également une expérience de remédiation, au sens de Bolter et Grusin (1999). Pour ces auteurs, la remédiation est l'opération par laquelle les « nouveaux » médias remodèlent (« *refashion* ») les anciens médias, ou l'acte d'incorporer les médias précédents dans un nouveau médium. C'est bien ce que fait @franklinfordbot : il incorpore d'anciens médias, ceux qui composent l'archive de Ford, et les remodèle dans un flux de tweets qui ont à la fois tout et rien à voir avec les documents « originaux ». Tout à voir car le « contenu » est le même (les phrases sont bel et bien les mots assemblés par Ford dans ses livres, ses articles, ses lettres), mais rien à voir car ces mots s'inscrivent dans des régimes d'existence médiatique radicalement différents.

21. En faisant cela, @franklinfordbot incarne aussi un des aspects importants de la « double logique » de notre culture de la remédiation identifiée par Bolter et Grusin, c'est-à-dire de simultanément multiplier et effacer les médias (1999, 5). L'effacement est clair dans l'interface d'une plateforme telle que Twitter, qui promet l'interaction directe entre ses utilisateurs. Dans la petite fiction que constitue @franklinfordbot, c'est comme si Franklin Ford parlait directement à ses abonnés. Mais pour en arriver à cette impression d'immédiateté, d'absence de médiation, il faut passer par la multiplication des couches médiatiques.

22. Reprenons ces transformations, une par une. D'un côté de la chaîne de transformations, il y a un document écrit par Ford, choisissons par exemple le *Draft of action*, une plaquette de 58 pages imprimée à compte d'auteur (probablement par Ford lui-même) à Ann Arbor, Michigan, en 1893. Des copies de ce document sont conservées dans plusieurs bibliothèques sur la côte Est des États-Unis (notamment à l'université du Michigan, à la Detroit Public Library et à l'université Brown). Une de ces copies a été numérisée, dans des circonstances que nous ne connaissons pas : c'est une collègue connaissant notre intérêt pour Ford qui nous a envoyé le document, par courriel. L'OCR transforme l'image du document en texte, c'est-à-dire en une chaîne de caractères lisible par l'ordinateur – dépouillée de tout son contexte et de tout son paratexte. Nous stockons celle-ci dans une base de données, agrémentées des métadonnées pertinentes (telles que la date et le lieu de publication) qui proviennent de notre lecture rapprochée des documents. Un script, c'est-à-dire du code écrit dans le langage de programmation Python*, découpe ce texte en phrases, puis choisit une phrase au hasard. En utilisant l'API* de Twitter, ce même script publie la phrase (parfois divisée en plusieurs morceaux pour respecter la limite de caractères de la plateforme) sur le compte Twitter de @franklinfordbot, et nous voici à l'autre bout de la chaîne de transformations.

Montrer la remédiation : « etaoin shrdlu » et le bruit de l'OCR

23. Pour les abonnés de @franklinfordbot, ces mutations sont invisibles. Pourtant, elles sont assez importantes. En chemin, certaines choses sont perdues (le grain du papier sous les doigts, la mise en page, le contexte de chaque phrase) et d'autres sont gagnées (les métadonnées* qui permettent de géolocaliser les tweets sur le lieu de publication du document, la possibilité de collecter les réponses aux tweets comme des annotations). Améliorations ou dégradation ? Peu importe, finalement, cela dépendra toujours des objectifs poursuivis par la recherche. Ce qui me paraît essentiel ici est de souligner qu'il y a des métamorphoses, qu'elles sont nombreuses, mais très peu visibles dans le produit final.

24. C'est ici qu'intervient à nouveau le « bruit » produit par l'OCR. Nous avons en effet délibérément décidé de conserver ces « erreurs » dans le texte publié par @franklinfordbot, car il s'agit de traces de ces métamorphoses successives. Faire tweeter @franklinfordbot avec des erreurs d'OCR, c'est donc rappeler que ce n'est *pas* Franklin Ford qui nous parle et nous écrit. Entre l'archive et le tweet, il y a toute une distance, parcourue à travers de nombreuses médiations qui sont façonnées par la matérialité de chaque couche médiatique.

25. Toutes ces étapes par lesquelles il faut passer pour faire un bot ne sont pas de simples opérations techniques. Cordell (2017) propose ainsi de comparer l'interven-

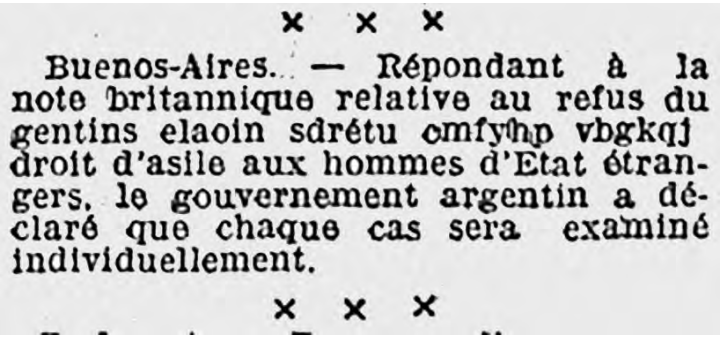
tion de l'OCR à celle des typographes. Compris comme un procédé de composition, l'OCR « recompose le texte dans des fichiers .txt ou .xml plutôt que sur une galée » (Cordell 2017, 196). Ce qui est produit par l'OCR doit alors être considéré comme une nouvelle édition du texte, au même titre qu'un nouveau travail de composition typographique. Certes, la nouvelle édition *ressemble* aux autres éditions, mais elle peut contenir des variations, et les chercheurs ont à leur disposition tout un appareillage bibliographique pour rendre compte précisément de ces différences. En ajoutant que l'OCR compose dans une langue qu'il ne comprend pas – tout comme certains typographes étaient capables de composer dans une langue étrangère – Cordell (2017) nous met donc en garde contre la tentation d'attribuer des qualités *humaines* aux uns, *automatiques* aux autres.

26. Ce n'est pas la première fois qu'un dispositif de composition qui allie humains et machines laisse des traces dans le texte sous forme d'un charabia de lettres mal assorties. Shoemaker (2019) propose de faire un parallèle entre l'intervention de l'OCR et celle des typographes et, plus précisément, des opérateurs de Linotype (une machine de composition au plomb, grâce à laquelle les opérateurs peuvent utiliser un clavier pour composer la forme imprimante d'une ligne de texte). Parfois, quand les linotypistes faisaient une erreur dans une ligne, ils la terminaient rapidement par une suite de caractères génériques correspondant aux lettres des deux premières colonnes du clavier, « etaoin shrdlu » sur les claviers anglais, « elaoin sdrétu » sur les claviers français. Cette ligne était ensuite destinée à être écartée, mais certains « etaoin shrdlu » et « elaoin sdrétu » passaient à travers les mailles du filet, et finissaient par être imprimés dans le journal (figure 2). De telles erreurs deviennent précieuses pour les historiens des médias, car elles font apparaître le travail des linotypistes, autrement invisible. Il est d'ailleurs assez amusant de voir que ces erreurs sont à leur tour réinterprétées par l'OCR dans la presse numérisée – avec des effets étranges de recomposition : là où il y a des « etaoin shrdlu », l'OCR reconnaît d'autres mots ou, au contraire, il en détecte là où il n'y en a pas (Shoemaker 2019).

Penser les médias et leur matérialité

27. Nos corpus sont donc le fruit de très nombreuses transformations, et cela n'est pas propre au numérique. Mais l'utilisation croissante de corpus et des archives numérisés en SHS multiplie potentiellement les couches de remédiation, alors même que l'immédiateté du numérique efface l'épaisseur matérielle des médias précédents. Confrontées à cela, les humanités numériques ne peuvent pas faire l'économie de penser les médias, leur matérialité et les conséquences épistémologiques de celle-ci.
28. Pour penser tout cela d'un point de vue théorique, on peut s'inspirer d'une conception des « médias » redevable aux travaux d'Harold Innis ou de Marshal McLuhan, qui postulent la primauté des médias dans la production de connaissance, en insistant sur le rôle central des médias dans la gouvernance des sociétés humaines,

depuis l'invention de l'écriture, mais également sur le rôle des médias dans le développement des savoirs (on notera que leur définition de « médias » va bien au-delà des traditionnels médias de masse). Pour McLuhan, les médias sont d'abord des épistémologies ; ils sont l'infrastructure même de nos existences (Peters 2015). Puisque nos connaissances et nos existences dépendent des médias, elles sont aussi façonnées et changées par les qualités matérielles des médias.


× × ×

Buenos-Aires. — Répondant à la note britannique relative au refus du gentins elaoin sdrétu cmfyhp vbgkqj droit d'asile aux hommes d'Etat étrangers, le gouvernement argentin a déclaré que chaque cas sera examiné individuellement.

× × ×

Figure 2. « Elaoin sdrétu » imprimé dans l'*Alliance Française*, 3 septembre 1943

Bibliothèque nationale de France, https://gallica.bnf.fr/ark:/12148/bpt6k768770n

29. Élargissons ici quelque peu le propos de John Durham Peters qui affirme que, dans la mesure où ils travaillent avec des médias d'enregistrement et de transmission ayant leurs propriétés matérielles spécifiques, « les historiens sont nécessairement des chercheurs en études médiatiques » (Peters 2008, 20). L'écriture de l'histoire est inséparable des médias d'enregistrement dont elle dépend, qu'il s'agisse de mise en boîte d'archives ou de mise en base de données de sources numérisées. Les historiens ont d'ailleurs développé de nombreux outils théoriques et pratiques pour penser ce rapport aux sources. Plus largement, ce sont tous les chercheurs en SHS qui se retrouvent parfois être des chercheurs en études médiatiques, à partir du moment où ils mettent les mains dans des corpus, qui sont toujours nécessairement *médiatisés*.

30. Pour penser la matérialité des médias d'un point de vue pratique, on peut alors adopter une démarche réflexive dans la fabrication de nos corpus artisanaux. C'est ce que @franklinfordbot cherche à faire, à son échelle modeste. Il n'y a évidemment pas de solution universelle à imposer ici : chaque corpus aura ses spécificités, chaque recherche ses manières créatives de réfléchir à la remédiation de ses archives. Dans beaucoup de cas, d'ailleurs, ces questions ne se poseront pas. Les corpus artisanaux peuvent très bien coexister avec des corpus industriels, qui continueront à envisager ces enjeux sous l'angle de l'efficacité, et ne voir dans les erreurs d'OCR qu'un problème à résoudre. Mais, dans certaines circonstances, nous pouvons bricoler des corpus, pas nécessairement parce que les modes de production industriels nous sont inaccessibles, mais parce que l'artisanat est un rapport à la matière, aux choses et à la connaissance qui lui est propre. De la même manière qu'un ébéniste est capable de mettre en valeur les défauts du bois qu'il travaille, nous pouvons alors chercher à apprécier les aspérités de nos matériaux.

Le corpus de tous les livres depuis les débuts de l'imprimerie, tous comptes faits…

Frédéric Glorieux

> *Le Congrès* est peut-être la fable la plus ambitieuse de ce recueil ; son thème est celui d'une entreprise tellement vaste qu'elle finit par se confondre avec le cosmos.
> (J. L. Borges, *Le Livre de sable*)

Introduction

1. En 2009, Google a publié Google Books Ngram Viewer[1] (GBNV), un outil qui projette la courbe de fréquence d'un ou plusieurs mots sur 4 siècles d'imprimés, à partir des index de leur moteur de recherche dans leurs livres numérisés. La masse de données est inédite, 100 milliards de mots. Avec le slogan initial « *Explore cultural trends* », cette proposition reprenait des idées de Google Trends (2008), une application montrant les mots les plus fréquemment cherchés, et donc, avec une grande valeur publicitaire. C'est avec cet esprit que la firme prétendait révolutionner l'histoire culturelle. Nous allons montrer que la puissance informatique et la masse des corpus ne suffisent pas à être concluants dans les humanités. En programmant un autre Viewer sur les mêmes données, nous en montrerons les biais et les béances, rappelant que l'essentiel dans les disciplines de la culture est la probité critique.

1. Cf. https://books.google.com/ngrams/

50 ans de statistiques lexicales, une tradition à continuer

Muller et Brunet

2. L'enthousiasme pour les données lexicales chronologiques est ancien. Étienne Brunet (1936-…), dès les années 1970 depuis Nancy (INALF), a démontré comment la base de textes littéraires informatisés pour la rédaction du *Trésor de la langue française* (70 millions de mots à l'époque, mille fois moins que Google Books) peut devenir une mine de réflexion sur l'histoire de la langue et de la société (Brunet 1981), en généralisant les méthodes statistiques que son maître Charles Muller (1909-2015) avait mises au point sur Corneille (Muller 1967). Avec des moyens techniques autrement plus limités que les nôtres, cette époque avait l'ambition théorique des pionniers. Leur rigueur statistique, leur finesse linguistique, et l'intelligence de leur jugement, reste le meilleur guide pour s'y retrouver dans les annonces publicitaires tonitruantes d'une grosse capitalisation boursière. Brunet n'a pas manqué de s'intéresser au GBNV, cette enquête

reprend le dossier où il l'a laissé en 2012 (Brunet 2012) et 2014 (Brunet et Vanni 2014).

Google, un *campus*

3. Le GBNV est accompagné d'une publication dans *Science* (Michel *et al.* 2011) dont est tiré le graphique ci-dessous (figure 1), montrant que la « pizza » est désormais plus populaire que le « steak ». Tous les graphiques ne sont pas aussi anecdotiques, on retrouvera par exemple des invitations aux études sur le genre, mais l'équipe ne comporte pas d'historien ou de sociologue.

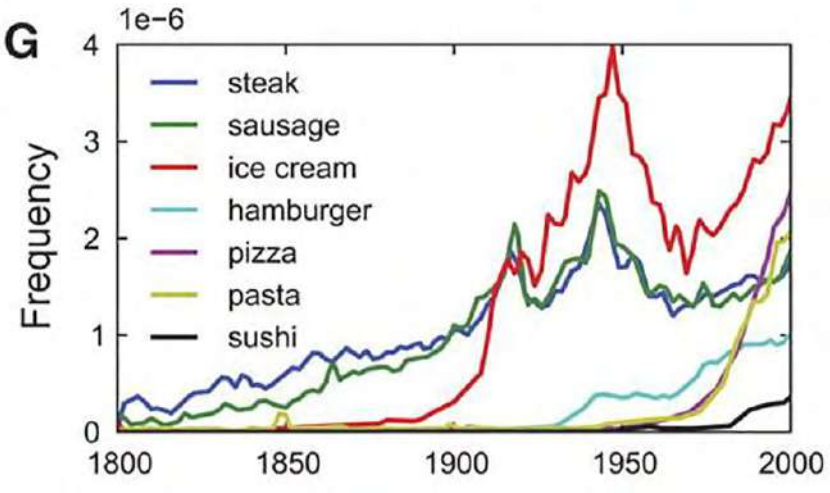


Figure 1. *Culturomics provides quantitative evidence for scholars in many fields. Historical Gastronomy.*
(Michel *et al.* 2011, fig. 5)

4. La publication scientifique n'est pas un accident pour Google mais une partie de sa production industrielle. Depuis l'origine, la firme se pense comme un « campus », pas très loin de l'université des fondateurs (Stanford). Ils affichent 5 650 publications entre 1998 et 2019[2], le rythme annuel ne cesse d'augmenter (3 en 1998, 306 en 2010, 714 en 2018). Les sujets les plus travaillés sont l'intelligence artificielle (1763 articles), les algorithmes (800), le traitement de la langue (549) ou les interfaces (528). Aucune science humaine n'est explicitement désignée, l'article ci-dessus est rangé dans la catégorie « *Information Retrieval and the Web* » (recherche d'information et Web, 269 articles). Cette publication était donc marginale dans la production de Google.

Google Books, un projet à l'arrêt

5. Vers 2010, les chercheurs du texte ont pu s'inquiéter à raison que la numérisation aboutisse à une privatisation des sources de la culture. Depuis, le coût de la mémoire baisse toujours, d'autres acteurs non commerciaux peuvent réunir des collections de taille importante (BNF, Wikisource, archive.org, Hathi Trust…), et Google a donné beaucoup de fichiers aux fondations américaines non commerciales. Bien loin d'une dystopie orwellienne où une firme aurait pu reconstruire l'histoire selon ses intérêts, on mesure plutôt que les humanités relèvent pour eux du budget publicitaire, ou de la bienfaisance. Ce projet Books, de numérisation des

2. Cf. https://ai.google/research/pubs/

livres, a été voulu par un fondateur de la firme (Larry Page), il tourne maintenant au ralenti depuis 2013, officiellement pour des raisons juridiques sur les droits d'auteur dont Google s'accommode pourtant très bien pour les images et les vidéos, plus sûrement parce que le retour sur investissement est faible.

Une bibliothèque sans bibliothécaire

Catalogage défectueux

6. À la sortie du GBNV, la communauté scientifique française a alerté sur des risques d'erreur, notamment avec les défauts dans la reconnaissance optique des caractères*, techniquement pardonnables et en progrès, ou le peu de représentativité pour le français des bibliothèques partenaires, en général nord-américaines, car la plupart des établissements français avaient refusé la proposition d'une numérisation gratuite de leurs fonds (à part la bibliothèque municipale de Lyon). Mais un biais important n'est pas indépendant de la volonté de Google.

7. Prenons par exemple le mot « Bacbuc », plutôt rare, il a un pic étrange vers 1730, quelques sursauts au XIX^e^ siècle, et un tapis mou au XX^e^. Ce mot ne se trouve que dans Rabelais (« la dive Bacbuc », « bouteille »). Chaque pic dans Google Books correspond à une nouvelle édition de Rabelais, enregistrée à sa date de publication, et non au siècle dont la langue du texte témoigne ; on trouvera ensuite des commentaires de Rabelais, et des dictionnaires. Le problème est similaire pour l'anglais, il suffit de tester « *heberon* », un hapax de Shakespeare, qui ressort à chaque nouvelle édition jusqu'aujourd'hui. Il y a donc beaucoup de textes en double dont l'année ne correspond pas à l'état de langue.

Contexte historique de l'édition, et des rééditions

8. Il est normal qu'un catalogue enregistre un livre à sa date de publication. Même une bibliothèque très centralisée comme la BNF n'arrive pas encore en 2019 à relier toutes les éditions d'une œuvre à un titre normalisé et daté du vivant de l'auteur. Toutefois, avec les dates de l'auteur principal, de mieux en mieux renseignées avec les rééditions, il est possible d'observer la proportion de titres dont l'auteur est mort à la date de publication. À partir des données data.bnf (la partie vérifiée du catalogue de la BNF), on peut établir que 10 à 20 % des titres, selon les années, sont publiés après la mort de leur auteur (figure 2, la ligne pointillée des titres des vivants est sur l'échelle de droite, 10 fois supérieure à l'échelle de gauche). Cette proportion n'est qu'un ordre de grandeur qui dépend de ce qui est pris en compte (par exemple, que faire des collectivités auteurs comme les congrégations religieuses ?). On notera que les rééditions chutent pendant les guerres et reprennent en force par temps de paix pour renouveler le stock des éditions scolaires et savantes. Cette croissance séculaire du nombre de titres et de pages, tranchée de guerres et de

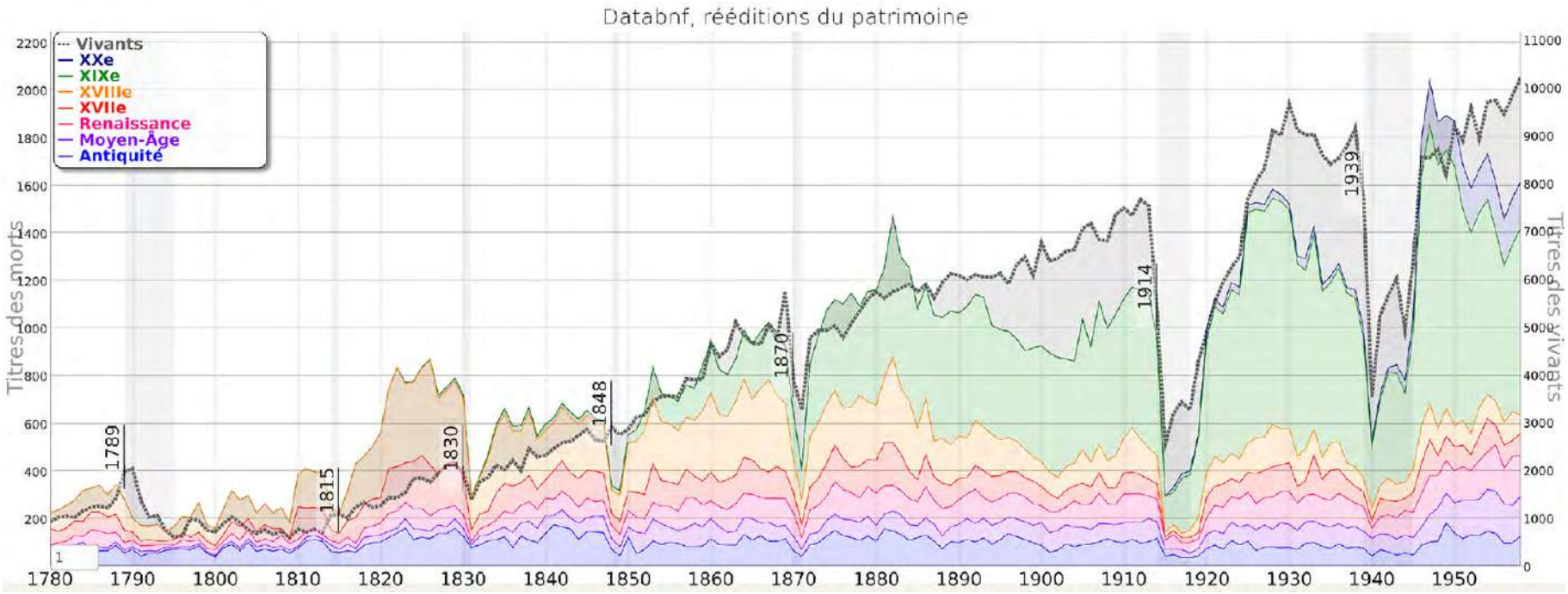


Figure 2. Catalogue de la Bibliothèque nationale de France, projection chronologique du nombre de titres par an à la date de publication, distingués selon la date de naissance de l'auteur

Titres XVIIe siècle = auteur né entre 1580 et 1680. Données data.bnf.fr

révolutions, forme le paysage sur lequel se projettent les fréquences. Toute statistique lexicale doit tenir compte du contexte historique.

9. Rappelons que la BNF reçoit le dépôt légal*, donc tous les titres publiés, et qu'elle a l'obligation légale de les conserver, contrairement à une bibliothèque municipale, ou une bibliothèque universitaire nord-américaine, qui peut désherber pour libérer ses rayons de livres jamais empruntés. Il est donc probable que le corpus Google Books comporte beaucoup de doublons des classiques de toutes années, et une plus forte proportion de livres anachroniques que sur ce graphique. Toutefois, la réédition d'un classique, ou d'un auteur, peut être considéré comme témoignant des intérêts d'une époque. C'est un matelas de mots dont il faudra tenir compte avec prudence.

Le dictionnaire de l'OCR

Volumétrie

10. Les données à télécharger sont lourdes, pour les seuls *1-gram* (les mots simples), il faut 3 Go de fichiers compressées, 15 Go décompressés, environ 2 000 *Recherche du temps perdu* de Proust pour seulement des listes de mots et leurs effectifs, sans aucun texte. Les listes comportent 9 663 037 *mots* différents. *Le Trésor de la langue française*, le plus gros dictionnaire du français, ouvre 100 000 entrées, dont sont tirées 600 000 flexions par les différents accords des noms et des adjectifs ou la conjugaison automatique des verbes[3]. Cette mécanique produit des mots dont on ne trouvera aucune citation d'auteur, comme « nous abalourdîmes » (ces mots *existent* désormais sur Internet par les réplications de ce lexique et les conjugueurs). Le dictionnaire orthographique de LibreOffice (le même que la plupart des logiciels libres) se suffit d'environ 500 000 lignes pour les mots de la langue et les noms propres les plus utiles[4]. Ce lexique permet en plus une lemmatisation grossière, par exemple pour grouper toutes les formes graphiques d'un verbe. Mais d'où viennent les 9 autres millions de *mots* de Google ?

3. Morphalou 2.0, https://www.ortolang.fr/market/lexicons/morphalou.
4. Cf. https://grammalecte.net/home.php?prj=fr.

Big data, une compétition épuisante

11. Cette livraison révèle la *culture* industrielle de l'entreprise. Les index sont donnés sans nettoyage, la société n'a pas craint pour son image, les fichiers en restent là depuis 2012, sans plus personne pour s'en soucier. Ce manque de soin se couvre d'une objectivité techniciste, les données sont le produit brut des algorithmes sans le biais d'un jugement humain. Donner 20 fois trop de *tokens** tient de la générosité, de la prudence scientifique de ne pas préjuger de ce qui pourra être utile un jour, mais aussi, d'un énorme gaspillage énergétique qui étouffe l'ordinateur personnel d'un particulier. Habituer les utilisateurs à ne plus ranger, à ne plus attendre, permet de les attacher aux produits d'une puissance informatique que seules quelques firmes peuvent concentrer. Le groupe préfère investir dans des centrales électriques qu'employer des documentalistes ou des lexicographes. Cette stratégie est à l'inverse de ce qui a conduit le projet *Frantext*. Les chercheurs n'y ont pas traité n'importe quelles données mais les monuments de leur langue, avec soin. Les promoteurs européens de l'intelligence artificielle et des grosses données devraient repenser au mirage de la guerre des étoiles de Reagan qui acheva d'épuiser l'URSS dans un effort militaire impossible. Pour être fiables, les réseaux neuronaux ont besoin d'être entraînés. Le travail qui n'est pas fait en amont dans le choix des données d'entraînement, ce qui demande de la qualification, est réalisé en aval, par des travailleurs peu qualifiés du clic (Henneton 2017), diminuant la finesse des effets que l'on peut en tirer. Il

est possible de faire mieux à moindre coût, à condition de travailler autrement.

Pourquoi seulement 10 millions de mots différents ?

12. « BrUn », « ch'altri », « pl.6 », « va1n »… le bas du tableau des fréquences est rempli d'incidents d'OCR, le phénomène statistique est inévitable. Les 50 mots les plus nombreux (« de », « le », « la »…) représentent 50 % des occurrences en contexte, ce sont les plus courts, avec peu d'erreurs ; les mots rares sont beaucoup plus nombreux, avec plus de lettres[5], ils s'exposent à plus d'aléas dans la reconnaissance de leurs caractères. Cette croissance exponentielle devrait en réalité aboutir à un dictionnaire bien plus grand, mais la liste a été tronquée à 40 occurrences au minimum sur 400 ans (avec parfois des années vides).

Les mots des images

13. Beaucoup de ces grappes de lettres viennent aussi des graphiques ou des photos, les numérisations Google en masse ne sont pas *zonées* (titres courants ou tableaux ne sont pas sortis du flux de texte). On trouve aussi des pages à l'envers, ou des mains d'opératrices (oui, généralement des femmes), qu'elles n'ont pas pu retirer à temps d'un scanner qui tourne à 1 000 pages à l'heure. L'avantage du numérique, c'est que ces *petites mains*[6] n'ont pas été coupées par la machine, mais ce ne sont pas des algorithmes qui ont tourné les pages.

Une expérience syntaxique prometteuse mais brouillonne

14. La tête de la liste de fréquences a d'autres types d'erreurs, dont notamment les entrées d'un étiquetage grammatical parfois confus. L'apostrophe, les majuscules, et certains accents sont étrangement mal traités alors que quelques lignes de programme suffiraient pour éviter ces bévues (21e « _X_ » ; 25e « _NOUN » ; 32e « l'_NOUN » ; 92e « '_DET » ; 115e « La_DET » ; 119e « qu'il_ADJ » ; 256e « 1_DET » ; 468e « Etat » ; 961e « État »…). Cependant, l'automate grammatical a aussi inscrit des distinctions plausibles, et précieuses, comparez par exemple « politiques_ADJ », « politiques_NOUN », « politiques_NOUN *_ADJ ». Le substantif est proprement distingué, les bi-grammes (couples de mots) révèlent des associations technocratiques en explosion depuis 1945 : politiques publiques, économiques, nationales, sociales, agricoles, culturelles, monétaires… Mais avant d'analyser les étiquettes grammaticales et les *n-gram**, il faut assurer les données simples. On se limitera aux mots simples uniquement composés de lettres, sans étiquettes, présents dans un dictionnaire

5. Le nombre de lettres des mots rares est une conséquence de la loi de l'entropie de Shanon (1948).

6. Wilson, Andrew Norman. 2012. *ScanOps*. Toronto, Canada : Art Metropole. http://www.andrewnormanwilson.com/ScanOps.html.

de référence. Lexique, le correcteur orthographique du logiciel libre, a l'avantage d'enregistrer beaucoup de noms propres vérifiés. Parmi les mots non reconnus, on trouve principalement :

- des abréviations et initiales : « M. », « p. », « A. », « J., P. », « Cf. »...
- des mots étrangers : « *the* », « *of* », « *and* », « *to* », « *und* », « *by* », « *that* », « *University* », « *Press* »...
- des graphies anciennes : « avoit », « étoit », « enfans », « tems »...
- des chiffres romains : « XV », « IX », « XVI^e^ »...
- des erreurs de segmentation : « luimême », « peutêtre », « GrandeBretagne », « EtatsUnis »...
- des erreurs d'accents : « siege », « priviléges », « Liége »...

25 % d'occurrences inutilisables

15. Sur environ 100 milliards d'occurrences déclarées (100 Go), 6 Go d'espaces sont enregistrés comme des *mots*. Sont-ils dans le décompte global ? On ne sait pas. La fréquence des mots anglais, comme « *the* », « *of* », « *and* », « *to* », indique qu'il ne s'agit pas seulement de notations bibliographiques éparses mais de textes rédigés, que l'on peut évaluer à 1,3 Go, qui croissent tout au long du XX^e^ siècle. La reconnaissance automatique de la langue d'un texte est pourtant très fiable, on peut supposer qu'il s'agit de volumes bilingues. Au final, 70,8 Go correspondent à un mot correctement orthographié selon Lexique, sans aucun traitement ; 4,9 Go peuvent être récupérés avec des règles déterministes et fiables (par exemple : si ce mot avec une majuscule initiale n'est pas au dictionnaire, est-ce qu'il s'y trouve en minuscules ?).

16. Les biais de l'OCR sont importants, mais ils sont homogènes. Google a procédé à 2 campagnes de reconnaissance des caractères, en 2009, et en 2012, avec d'ailleurs de réels progrès ; ce qui assure un même niveau pour tout le corpus. Depuis le remplacement des caractères Garamond par la Didot (début du XIX^e^ siècle), les formes des lettres sont assez stables. Un progrès dans la reconnaissance est observable avec l'apparition de l'impression *offset* (procédé photographique sans usure des caractères au plomb), mais ces biais sont homogènes sur une année, et de toute façon le nombre de mots augmente avec le temps, par l'effet du nombre de livres publiés. Ce bruit ne semble pas interdire des conjectures globales, en étant prévenu des risques sur certains mots (par exemple « celte » est beaucoup plus fréquent que « Gaule », parce que c'est une erreur d'OCR fréquente sur le mot « celle »).

Sémantique des courbes

Unité, le rang

17. Google présente ses effectifs de mots en pourcentages de son total (figure 3), avec nécessairement beaucoup de décimales après la virgule. Mais ces proportions ne sont pas comparables avec d'autres corpus, non seulement parce que le protocole de comptage n'est pas documenté (les locutions comme « parce que » sont-elles comptées comme un ou deux mots ? les ponctuations sont-elles considérées comme des *tokens* ?...), mais surtout, parce qu'on ne sait pas ce que sont les mots qui ne sont pas au dictionnaire. Aussi, nous préférerons donner un indice d'importance d'un mot en donnant son rang dans l'ordre des fréquences. « De », « le », « la » sont les numéros 1, 2, 3 ; plus le rang est grand, plus le mot est rare. Cette grandeur est moins dépendante du total, elle est par ailleurs corrélée à la fréquence par la loi de Zipf, fondamentale en statistiques lexicales : log(rang)=K*log(effectif)

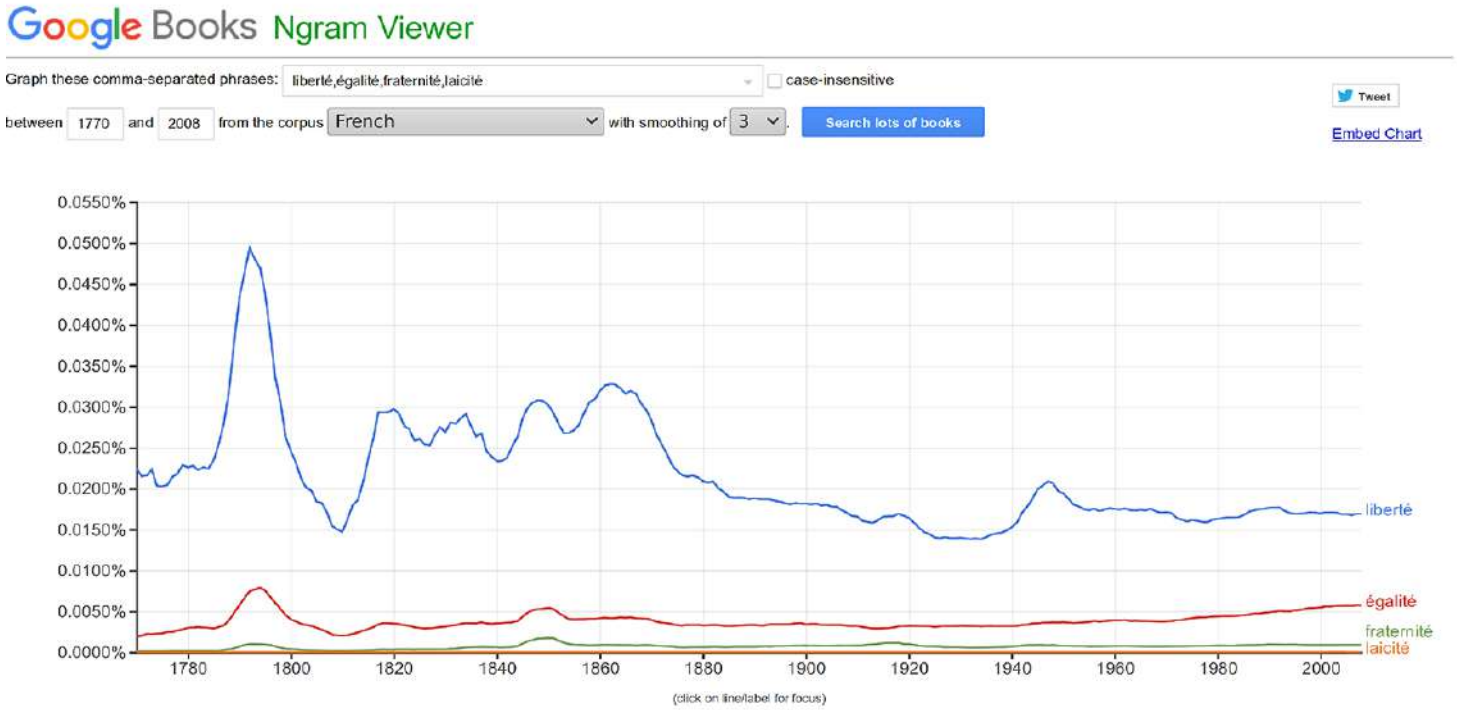


Figure 3. Google Books Ngram Viewer, corpus français 2012, 1770-2008, « liberté, égalité, fraternité, laïcité »
Crédit : Frédéric Glorieux

Échelle logarithmique

18. Afin de pouvoir comparer des grandeurs d'ordre très différent, il est conseillé d'utiliser une échelle logarithmique où par exemple la pente d'une baisse de 50 % aura le même angle pour une valeur de départ de 10, 100, 1 000 ou 10 000. Notre œil animal est très sensible aux alignements. Enfin, les courbes sont affectées d'une agitation brownienne d'autant plus élevée que le mot est rare et ancien (moins de livres).

Lissage

19. Un coefficient de lissage est généralement appliqué, donnant l'impression d'un phénomène continu, alors que les données réelles sont des points éparpillés. Une mauvaise formule peut araser les pics et masquer des ruptures historiques, ou créer des faux points d'inflexion. Il faut trouver une formule visuelle qui laisse les points à leur emplacement exact sur l'échelle, et ne présenter des lignes que comme des interprétations, ou une marge d'erreur.

Histoire longue et événements

20. L'image ci-après (figure 4) met en pratique cette réflexion sur la sémantique du graphique à deux variables. Il suscite des interprétations plus riches que les courbes de Google. « Liberté » est un mot fréquent, ses variations ont plus d'ampleur en valeurs absolues que « égalité » ou « fraternité », mais l'échelle logarithmique permet de remarquer de fortes corrélations entre eux, ainsi qu'avec l'histoire. L'échelle des dates est spécialement adaptée à l'édition française. Dès la révolution de 1789, les trois mots de notre devise préoccupent les publicistes. La chute après Robespierre ne résulte pas seulement d'une fatigue de la Révolution, mais aussi d'une baisse du nombre de titres publiés, en raison de l'effort de guerre et de la censure napoléonienne. 1848 est un nouveau pic révolutionnaire et éditorial, mais l'écho est moindre qu'en 1789, le nombre de pages imprimées est en forte croissance avec la révolution industrielle, et les valeurs de la « liberté » se compliquent (le *libéralisme* bourgeois prend conscience de lui-même). Ensuite, la « liberté » et la « fraternité » auront des sursauts patriotiques pendant les guerres mondiales, mais pas l'« égalité ». Le petit pic de 1989 ne vient pas de la chute du mur mais de l'opportunisme éditorial pour le bicentenaire de 1789.

21. Le mot « laïcité », que certains voudraient maintenant accoler à notre devise, permet de tester l'effet visuel dans les basses fréquences. Les occurrences sporadiques avant la III^e^ République sont des scories d'OCR dans la littéra-

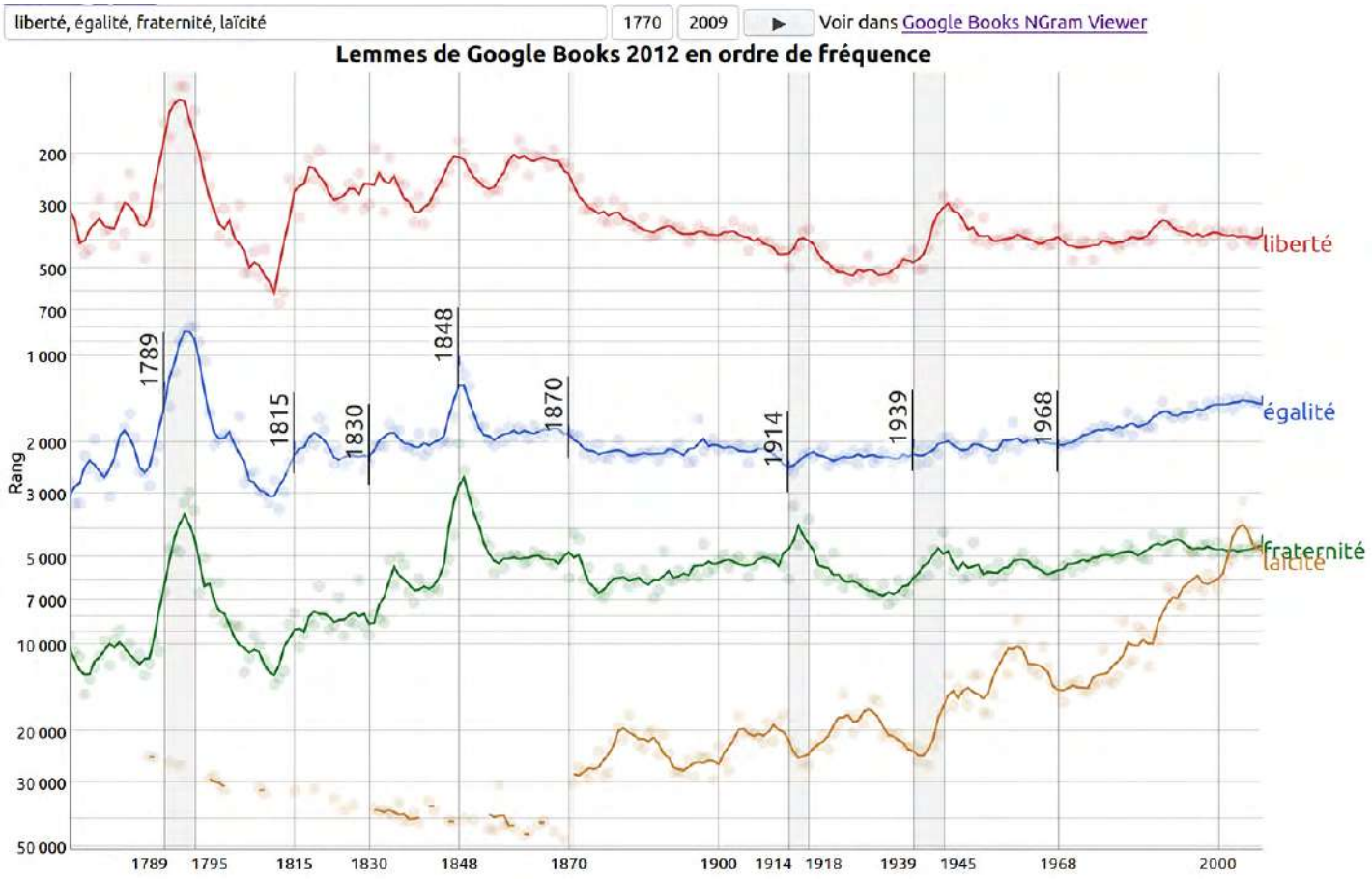


Figure 4. Données Google Books 2012, projection logarithmique de la fréquence chronologique de 4 lemmes « liberté, égalité, fraternité, laïcité »

Crédit : Frédéric Glorieux

ture religieuse, la notion telle que nous l'entendons apparaît après 1870 et croît jusqu'à aujourd'hui. Elle s'oublie un peu pendant les guerres comme toutes les questions de société.

22. On peut aussi essayer « travail », « famille », « patrie » ; « guerre » et « révolution » ; ou « socialisme », « communisme » et « anarchiste » ; l'expérience confirme que l'écrit en français est sensible à l'histoire, qu'il annonce parfois

les événements, et plus souvent les suit, avec l'inertie du temps de l'écriture. Le corpus peut ainsi répondre à des curiosités, mais que contient-il vraiment ? Quels sont les mots caractéristiques d'une année, d'une période ?

Spécificités chronologiques

23. L'interface précédente permet d'observer les variations d'un mot sur le corpus, mais la lexicométrie française nous a appris à être plus exigeants. Comme il est courant par exemple de comparer le vocabulaire d'*Atilla* relativement aux autres pièces de Corneille, quels seraient les mots qui caractérisent la Seconde Guerre mondiale, ou la présidence Mitterrand ?

Le temps n'est pas un tout borné

24. Les données manquant de fiabilité, il est imprudent d'appliquer une formule statistique sans réflexion. L'année d'édition d'un livre est une unité pertinente. Grouper à l'avance des années, par décennie par exemple, permettrait de pré-calculer des ensembles mais écraserait les événements de l'histoire dont nous avons vu le retentissement. Il n'est pas possible de comparer une partie à un tout dont on ne connaît pas les bornes. Faut-il caractériser 1944 à 1800-2000, 1789-1990, 1870-1962, ou seulement 1939-1945 ? Et pourquoi ne pas comparer 1914-1918 à 1939-1945 ? Il faut que l'utilisateur puisse toujours lui-même définir librement deux ensembles et trouver des formules statistiques qui relèvent les spécificités de l'un par rapport à l'autre.

1929-1938	1939-1945
102 % international #180	1614# Québec 103 %
84 % convention #425	1807# Montréal 126 %
82 % conférence #566	1808# Hitler 288 %
106 % congrès #740	2201# ravitaillement 113 %
97 % tumeur #2053	2247# canadien 87 %
86 % rapporteur #2725	3018# armistice 84 %
90 % pacha #2881	3169# Laval 82 %
81 % inoculation #3235	3513# Vichy 167 %
116 % syphilis #3559	3685# nazi 320 %
85 % tchécoslovaque #3783	3891# Pétain 207 %
92 % armateur #3805	4318# Roosevelt 83 %
103 % Serbie #4172	4863# hitlérien 129 %
90 % avarie #4296	5100# gazogène 119 %
121 % aéronef #4430	5633# totalitaire 255 %
129 % syphilitique #5523	5733# Gaulle 503 %
141 % connaissement #6107	5775# Churchill 145 %
84 % vinicole #6154	5881# interné 126 %
152 % abordage #6319	6313# blindé 120 %
120 % rouble #6678	6530# tank 97 %
133 % uretère #7296	6531# artisanal 118 %

Figure 5. Données lexicales Google Books Ngram, comparaisons chronologiques, différence de rang > 80 %

Statistiques des rangs

25. Entre la croissance séculaire du nombre de titres et les creux des guerres, les années sont inégales, il faut donc éviter que les effectifs de 2009 par exemple, écrasent 1940. Puisque les fréquences ne sont pas fiables, on en

restera à une arithmétique des rangs. Quel est le top 50 des mots de 1954-1962 ? Nous dirons que c'est la moyenne des rangs de chaque année. C'est discutable, mais on peut s'en faire une intuition, et donc avoir un rapport critique avec le résultat.

26. Pour comparer des palmarès, on peut signaler les plus grosses montées, et les plus grosses chutes. Toutefois, comme vu plus haut avec l'échelle logarithmique, une chute de 10 à 100 est relativement beaucoup plus importante qu'une chute de 5000 à 20000, il faut garder les proportions. Toutefois, même avec des coefficients judicieusement réglés, des différences importantes de rang n'ont pas la même valeur pour le haut ou le bas du tableau. En un premier temps, il est donc prudent de garder l'ordre des rangs, et n'utiliser les différences de rangs que comme un seuil de filtre. On peut alors proposer une liste des mots caractéristiques de la Seconde Guerre mondiale relativement à l'avant-guerre.

Un fatras n'est pas un corpus

27. Selon les données lexicales Google Books, le rang du mot « Hitler » est 3 fois plus haut en 1939-1945 qu'en 1929-1938. Les mots spécialement plus présents pendant la guerre qu'avant, comportent : « Hitler », « ravitaillement », « armistice », « Laval », « Vichy », « nazi », « Pétain », « Roosevelt », « hitlérien », « gazogène », « totalitaire », « Gaulle » (de)... ce qui est plausible ; mais aussi, « Québec », « Montréal », « canadien »... ? L'avant-guerre se caractériserait par « tumeur », « inoculation », « syphilis », « syphilitique », « uretère »... ?

28. L'interface du GBNV permet d'accéder à la recherche avancée de Google Books, ce qui donne un aperçu de ce qu'est un *livre* selon ce corpus. On trouvera de sinistres titres d'époque comme : *Hitler, caporal stratège* ; mais aussi : *Statutes of the Province of Québec – Volume 8 – Page 490* ; *De la vitamine C chez Valeriana officinalis et de l'influence de la vitamine B1 sur son métabolisme* ; *Revue Internationale des Industries Agricoles : International Review of Agricultural Industries, Issues 5-7*... On en arrive à la conclusion que cette collection est constituée de tout ce qui a pu se numériser facilement à l'époque, ou qui était déjà en ligne, pour faire masse ; il en résulte que ce fatras d'opportunités n'est représentatif que de sa propre histoire, mais pas d'une réalité extérieure.

Univers indéfini et mondes clos

29. L'expertise industrielle de Google s'est construite sur Internet, un univers ouvert en expansion, dont on ne peut pas arrêter le contour[7]. Ses index sont conçus pour répondre à des questions imprévisibles. Si quelqu'un s'intéresse à « 4chql'am » ou « mou1r », il aura des résultats, qui peuvent même le satisfaire. Notons qu'il n'a aucune assurance qu'il n'existe pas un document en

7. Cf. (Koyré 1988).

ligne qui répondrait mieux à sa requête, mais faute de moteur plus puissant pour comparer, il croit Google.

30. Lorsque l'on doit prendre en compte la totalité de l'index pour par exemple calculer la fréquence d'un mot, alors la liste est fermée, et les erreurs prennent du poids. Cette culture du moteur de recherche devient nuisible lorsqu'il s'agit de constituer le corpus. Il ne suffit pas de faire grossir un nombre pour qu'il soit plus représentatif. Un échantillon bien pondéré de 2 000 personnes représente mieux l'opinion que 15 000 partisans partageant les mêmes convictions. La croissance doit être conduite par un principe qui garantit une cohérence.

31. Un corpus doit être un *tout*, avec une frontière définie, qui permet de mesurer le dedans et le dehors. L'œuvre d'un seul auteur, par exemple, est un miroir du monde complètement faussé, mais on en sent intimement la courbure. Le corpus littéraire du *TLF* est une histoire dont nous connaissons beaucoup de noms, avec les manies et les lubies, mais on a une vie de lectures pour en corriger les biais. Google Books est comme un miroir de télescope composite (Bachelard 1993), avec du cristal et des cailloux, il ne montre plus le monde, on ne voit que le miroir. L'humain reste l'instrument le plus sensible à lui-même, la difficulté est juste de l'étalonner.

Conclusion

32. Les conclusions historiques que l'on peut tirer de Google Books Ngram Viewer sont décevantes, faute d'un minimum de rigueur professionnelle de leur part. Cette analyse ne s'est pas effrayée de la désinvolture bibliographique, ou du bruit des caractères, mais il faut en conclure qu'il est fortement déconseillé de se fier à ces données lexicales. Elles ne sont pas seulement bruitées avec des biais homogènes, les déséquilibres peuvent être massifs sur certaines années et emmener vers de fausses conjectures.

33. Mais au fond, on peut sortir assez rassuré de cette expérience. Google défriche un chemin qui est beaucoup plus facile à suivre ensuite. Il n'est pas difficile de faire mieux avec un peu de soin et d'astuce. Les procédures numériques pour corpus bruités développées ici ne resteront pas lettre morte, elles peuvent être transposées à des séries historiques mieux définies, par exemple la presse numérisée par la BNF, qui aura le même bruit OCR mais des métadonnées autrement plus précises. Il y a bien plus de résultats à en attendre pour l'histoire, avec pourtant des masses bien moindres.

Pour consulter les données mobilisées dans le chapitre, voir **https://hnso-corpus.nakala.fr/**

Archelec, les archives électorales françaises de la Ve République, du papier au numérique : reflet fidèle ou distorsion ?

Odile Gaultier-Voituriez

Introduction

1. Les archives électorales françaises de la Ve République du Centre de recherches politiques de Sciences Po (CEVIPOF) ont commencé à être numérisées en partenariat avec la bibliothèque de Sciences Po à partir de 2013 dans le cadre du projet *Archelec*. Constitué dans un but de recherche en interne par les politologues et les sociologues du CEVIPOF, le fonds d'archives papier contient des professions de foi, bulletins de vote, tracts, affiches mais aussi des résultats électoraux, de la presse et des travaux de chercheurs effectués à partir de ces documents.

2. Nous réfléchirons donc à ce sujet à partir de plusieurs questions, soulevées au fur et à mesure de la mise en œuvre du projet, bien au-delà de ce qui était prévu au début : Quelles typologies faut-il numériser ? Quels choix doivent être effectués ? Quel risque juridique peut-on prendre ? Quels sont les droits qui s'appliquent à ces différents types de documents – droit d'auteur, droit à l'image entre autres ? Quelles typologies peuvent être numérisées et mises à disposition gratuitement en ligne sans risque ou avec un risque faible ? Quelles sont les modalités de réutilisation à envisager ? Nous commencerons néanmoins par poser les bases de départ de la constitution du fonds d'archives et de son contexte, puis du projet de numérisation, de sa mise en œuvre et des étapes, et de toutes les questions qui se sont posées pour réfléchir enfin à ce qu'est le corpus mis en ligne par rapport au fonds d'origine et à ses nouveaux usages.

Le fonds d'archives de départ

Le CEVIPOF

3. Le Centre d'étude de la vie politique française contemporaine (CEVIPOF) est créé en 1960 par Jean Touchard, secrétaire général de la Fondation nationale des sciences politiques (FNSP), pour participer à sa mission de recherche dans le domaine des élections (géographie et sociologie électorales), des institutions et des partis et de l'analyse de la vie politique, de la pensée et de l'histoire des idées politiques menée par André Siegfried puis par François Goguel et Jean Touchard. Le CEVIPOF devient une unité mixte de recherches (UMR) du Centre national de la recherche scientifique (CNRS) dès 1968 et prend le

nom de Centre de recherches politiques de Sciences Po en 2003.

4. Dès ses débuts, le CEVIPOF rassemble de nombreuses sources d'abord destinées à la recherche du laboratoire, surtout autour des élections. Le centre de documentation conserve actuellement la production scientifique des chercheurs, les ressources électorales (imprimés, résultats et archives), les tracts politiques et sociaux, un corpus papier et numérique de plus de 25 000 sondages d'opinion sur la vie politique et sociale en France produits par les instituts de sondages comme BVA, CSA, l'IFOP, Louis Harris, la Sofres, etc. depuis 1950, les enquêtes qualitatives menées par les sociologues[1] et les archives administratives du centre.

5. Le fonds d'archives électorales est un fonds original, produit par et pour la recherche, reflet de l'activité du CEVIPOF. Il a été partiellement numérisé, avec ses avantages, notamment pour un accès élargi, et ses limites juridiques et scientifiques, car le fonds numérisé n'est plus l'exact miroir du fonds papier. Nous y reviendrons.

La constitution du fonds par et pour la recherche

6. Les données électorales ont été collectées et utilisées par de nombreux chercheurs du CEVIPOF en science politique dans le cadre des grandes enquêtes à l'occasion des élections présidentielle, législatives, mais aussi régionales ou européennes[2]. Le CEVIPOF a obtenu les documents officiels des scrutins par les préfectures grâce à une circulaire envoyée à la demande du président de la Fondation nationale des sciences politiques (pour les élections législatives de 1962 par exemple) ou auprès du bureau des élections du ministère de l'Intérieur. Ensuite, la collecte a été menée et l'est toujours par le centre de documentation qui conserve le fonds. Les tracts éphémères – qui ne font pas l'objet de publications officielles et ne relèvent pas du dépôt légal* – demandent un réel travail de collecte volontaire par le biais de dons personnels spontanés ou provoqués. L'exhaustivité est donc un mythe et le fonds est complémentaire de sources rassemblées par d'autres acteurs.

7. Le fonds continue à s'accroître régulièrement. Un volume important de documents variés a pu être regroupé : tracts, professions de foi officielles et bulletins de vote, mais aussi affiches, discours, dossiers de candidat, périodiques et résultats relayés dans la presse nationale et locale. Enfin, des objets (bandeaux, T-shirts, sacs, pin's ou ballons de baudruche distribués lors des meetings) et des notes de travail produites par les chercheurs complètent ces archives.

1. Entre autres sur les jeunes, les femmes, le catholicisme, l'engagement politique, le rôle des parents d'élèves, le Parti socialiste, la CGT ou la CFDT.

2. Alain Lancelot, Jean-Luc Parodi et Jean Ranger dans les années 1950 à 1960, puis Roland Cayrol, Daniel Derivry, Guy Michelat, et Colette Ysmal dans les années 1960 et 1970, et Élisabeth Dupoirier, Gérard Grunberg, Jérôme Jaffré, Pascal Perrineau, François Platone et Jean Ranger dans les années 1970-2000.

8. Classées dans l'ordre chronologique des élections, ces sources électorales sont inventoriées et mises à la disposition du public. Elles sont rassemblées en 274 cartons, soit 34 mètres linéaires, et couvrent plus de quatre-vingts ans de scrutins, de l'échelon communal au niveau européen, des élections législatives de 1936 jusqu'aux dernières élections, actuellement les européennes de 2019[3]. Les documents les plus anciens concernent les élections municipales de 1936. L'essentiel du fonds porte sur la période postérieure à la Seconde Guerre mondiale, depuis les élections à l'Assemblée nationale constituante d'octobre 1945, c'est-à-dire les IVe et Ve République. À partir des législatives de 1956, les documents de campagne et les professions de foi sont également conservés. Chaque élection est traitée en trois étapes : avant, pendant et après le scrutin, soit la campagne, les résultats et l'analyse politique. Les documents de campagne pour les élections municipales et cantonales sont, par rapport aux élections législatives et présidentielles, relativement peu nombreux, excepté pour les municipales de 1977. Cette année-là a été menée une importante récolte de documents de campagne, essentiellement pour la région parisienne. Après ce scrutin, on observe une nette diminution de ce type de documents et de résultats officiels d'élections locales. À partir de 1983, le fonds d'archives est surtout constitué des professions de foi et des résultats des élections présidentielles, législatives et régionales.

3. Gaultier-Voituriez, Odile, éd. 2019. *Inventaire des archives électorales : 1936-2019*. 7e éd. Paris, France : CEVIPOF. https://www.sciencespo.fr/cevipof/sites/sciencespo.fr.cevipof/files/ArchElecINV_CEVIPOFCEVIPOF_2019-07-01.pdf.

L'utilisation du fonds papier

9. Les archives électorales ont d'abord été utilisées par les politologues et les sociologues du laboratoire, puis l'usage s'est étendu à un public plus large, à Sciences Po et en dehors de l'institution. Le nombre de recherches a augmenté et les sujets se sont diversifiés : femmes en politique, communication politique, place du Front national et politiques publiques, par exemple. Le fonds ayant acquis avec le temps une profondeur chronologique, les historiens de la période contemporaine ont aussi commencé à le mettre à profit. Les étudiants et les enseignants-chercheurs préparent des mémoires de master, thèses, habilitations, projets de recherche[4], rapports, communications, articles académiques et ouvrages. De nouveaux usages ont vu le jour ces dernières années : iconographie, édition[5], émissions de télévision, production de films et expositions[6]. Des candidats aux élections et des particuliers passionnés par la politique viennent également consulter les documents.

4. Les projets de recherche peuvent être très larges : en 2015, les programmes de tous les candidats aux présidentielles de 1981 à 2012 ou, en 2013, les législatives partielles de 1986 à 1997. Un projet plus circonscrit a concerné en 2014 le Front national et les femmes.

5. Notamment pour la reproduction de tracts dans des manuels scolaires.

6. Par exemple au Centre mondial pour la paix de Verdun, en 2018-2019, les 60 ans de la Ve République et ses présidents. Des affiches originales ont été prêtées et de nombreux fac-similés de tracts ont été reproduits.

Le projet de numérisation

Choisir et mettre en œuvre la numérisation

10. L'évolution du Web a conduit le CEVIPOF à réfléchir à la numérisation de ce fonds pour le mettre à disposition de manière plus large sur Internet afin de démultiplier et de diversifier le public au-delà des chercheurs : journalistes, particuliers, érudits, collectionneurs, généalogistes et même le grand public. Le Web permet la recherche en ligne du texte intégral, une analyse de contenu et des études statistiques avec des logiciels spécialisés.

11. Les documents éphémères électoraux sont encore rares parmi les archives disponibles en ligne[7]. Les éphémères n'ont pas d'obligation de dépôt légal, même si le service des recueils de la Bibliothèque nationale de France en reçoit un certain nombre en don. En dehors des Archives nationales pour les fonds du ministère de l'Intérieur et de son bureau des élections, les autres fonds ou collections de matériel électoral sont souvent circonscrits à un département ou à une région – dans les archives départementales surtout – ou à un parti. C'est le cas de la Fondation Jean-Jaurès, qui a mis à disposition du public de très nombreux documents du Parti socialiste sur la base de données Archives socialistes[8].

12. Pour mettre en œuvre cette numérisation, le CEVIPOF a noué en 2013 un partenariat avec la bibliothèque (Direction des ressources et de l'information scientifique[9]) de Sciences Po[10] qui possédait une expertise en matière de numérisation de ses dossiers de presse[11] et souhaitait élargir le champ de la numérisation. Le CEVIPOF et la bibliothèque de Sciences Po ont alors répondu à un premier appel à projets sur lequel nous reviendrons.

Les questions juridiques soulevées

13. L'écueil juridique apparaît très vite, dès le début du projet. De nombreuses questions se posent : Quelles typologies retenir ? Quel risque juridique prendre ? Les typologies retenues sont tout d'abord celles de documents qui n'ont pas encore été numérisés et que d'autres institutions ne numériseront pas, mais aussi celles qu'il est possible légalement de mettre à disposition. Les droits qui s'appliquent pour la consultation sur place des documents et pour leur mise en ligne ne sont pas les mêmes. La mise à disposition de sources numérisées sur le Web

7. Cf. http://www.numerique.culture.fr/pub-fr/index.html
8. Cf. https://archives-socialistes.fr
9. La bibliothèque de Sciences Po est devenue la Direction des ressources et de l'information scientifique (DRIS) qui se centre sur le contenu, notamment des données de la recherche, et sur une offre de service aux chercheurs.
10. Je tiens ici à remercier vivement Sylvaine Detchemendy, co-pilote du projet *Archelec* pour la bibliothèque, Donatienne Magnier, responsable du département Valorisation Numérisation, Sophie Forcadell, co-coordinatrice du projet *Archelec 4*, et l'ensemble des personnes impliquées dans le projet à la bibliothèque et au CEVIPOF.
11. Accessible depuis le campus numérique de Sciences Po : http://dossierspresse.sciences-po.fr/consult/.

implique différentes législations : droits d'auteur, droit à l'image, droit à l'oubli et données personnelles[12].

14. En 2012, afin de préparer les nouveaux projets de numérisation, la bibliothèque de Sciences Po fait appel au cabinet Alain Bensoussan qui évalue le risque existant, selon une échelle graduée de faible à fort. Ce risque n'est pas nul. Il est donc impossible de numériser l'ensemble du fonds pour un accès gratuit en ligne. L'analyse de risque[13] conduit à la démarche suivante, en trois étapes pour chaque type de documents : un droit est-il applicable et lequel[14] ? Qui en est titulaire ? Et qui a intérêt à en empêcher la publication ?

15. Trois typologies de documents sont écartées : les ouvrages, les périodiques et la presse. Ils nécessitent une description différente de celle des autres pièces du fonds, ils seront peut-être numérisés dans le cadre de Gallica, la bibliothèque numérique de la Bibliothèque nationale de France et, surtout, ils relèvent incontestablement du droit d'auteur. Le titulaire des droits est souvent connu et il les exploite lui-même. Le risque est donc maximal.

16. Selon l'analyse du cabinet Alain Bensoussan[15], les documents de partis peuvent être regroupés en trois catégories de documents : internes, d'information et électoraux. Parmi les documents internes se trouvent les rapports et les procès-verbaux de réunion (écrits scientifiques). Les documents d'information, relevant du droit d'auteur, regroupent les tracts (œuvres complexes), les affiches (œuvres graphiques) et les discours transcrits (œuvres dérivées). Les documents électoraux enfin, qui relèvent aussi du droit d'auteur, rassemblent les programmes (œuvres graphiques et écrits) et les listes des candidats (œuvres dérivées et œuvres complexes). La titularité des droits des documents de partis, œuvres collectives, appartient généralement au parti, sauf pour les documents externes s'ils ont été produits par une agence et que les droits n'ont pas été cédés au parti[16]. Les professions de foi, diffusées officiellement et publiquement en grand nombre dans les boîtes aux lettres et mises à disposition dans les bureaux de vote, restent néanmoins des œuvres collectives dont le titulaire des droits est le parti. Le degré de risque est ensuite évalué[17]. Le risque zéro consiste à obtenir l'autorisation du parti s'il existe toujours. Dans un certain nombre de cas, il est difficile, voire impossible, d'identifier un titulaire des droits. Le risque de revendication est donc faible et la personne qui revendique devra apporter la preuve qu'elle en est l'auteur ou l'ayant droit. En outre, l'exploi-

12. Le règlement général sur la protection des données entré en vigueur le 25 mai 2018 est venu renforcer cet aspect.

13. D'après : « Outils d'analyse, fonds documentaire, IEP ». 2012. Document interne. Cabinet Alain Bensoussan. 2-4.

14. Si c'est une œuvre de l'esprit, il s'agit du droit d'auteur. Si une personne est reconnaissable, c'est le droit à l'image qui s'applique. S'il s'agit d'une personne publique, dans l'exercice de sa vie publique et sans atteinte à la vie privée, c'est le droit à l'information qui s'exerce.

15. D'après : « Rapport d'audit, fonds documentaire, IEP ». 2012. Document interne. Cabinet Alain Bensoussan. 26-27.

16. *Ibid*. 27-28.

17. *Ibid*. 29-30.

tation a lieu dans un but non commercial et les documents sont anciens.

17. Après cette analyse, et en fonction de la réalité des documents repérés dans le fonds, les typologies retenues pour la numérisation se précisent : pas de périodiques, ni de brochures, ni de coupures de presse, qui relèvent du dépôt légal. Dans un premier temps seront numérisés les cartons sériels composés d'éphémères électoraux à diffusion publique officielle, à risque faible : professions de foi, bulletins de vote et affiches. Dans un second temps et après la mise en ligne de la première série, il est prévu de numériser les cartons *varia* contenant les autres documents de parti, à risque faible. Il demeure le problème de l'iconographie pour lesquelles le photographe est titulaire des droits, notamment sur les affiches. Une liste des photographes identifiés est préparée. Ils sont peu nombreux et une recherche active sera effectuée, qui permettra d'instruire une demande d'autorisation.

18. Les modalités de réutilisation sont aussi étudiées en fonction de ces contraintes juridiques. Les documents sont donc mis en ligne sous la licence Creative Commons* Attribution – Pas d'utilisation commerciale – Pas de modifications 4.0 international (CC BY-NC-ND 4.0). Elle permet une réutilisation non commerciale à l'identique en citant la source.

Les projets successifs et leur périmètre

19. Depuis 2013, cinq projets *Archelec* ont vu le jour. Le premier, baptisé *Archelec 1*, a été financé grâce à un appel à projet du segment 5 – numérisation – de la Bibliothèque scientifique numérique (BSN) à hauteur de 39 000 euros. Ce financement a permis la numérisation par un prestataire externe de 35 000 professions de foi et bulletins de vote des élections législatives de 1958 à 1993. Le choix a porté sur ce type d'élections qui couvrent l'ensemble du territoire et peut donc intéresser un public élargi ainsi que des chercheurs travaillant sur des séries longues. Le projet devant être réalisé en dix-huit mois, la mise en ligne a été effectuée en recueils et non à la pièce, par élection, département, circonscription et tour, sur Internet Archive[18]. Ce site a été retenu en raison de sa robustesse, de sa maintenance et de son moissonnage par les moteurs de recherche, malgré des possibilités limitées de métadonnées*.

20. Le projet *Archelec 2* a été consacré à l'élection-reine de la Ve République, la présidentielle. Il couvre les années 1965 à 2012, soit neuf élections, et 1 200 documents beaucoup plus variés, des professions de foi et bulletins de vote officiels aux tracts officieux, autocollants, correspondance et objets. La numérisation a été préparée sans financement extérieur sur la station de numérisation nouvellement acquise par la bibliothèque et la mise en ligne a eu lieu à la veille de la présidentielle de 2017. Le lot *Archelec 3*

18. Cf. https://archive.org/details/archiveselectoralesducevipof

concerne un nombre beaucoup plus réduit de documents, pour les législatives de 1986 à la proportionnelle. Ces élections ayant eu lieu en même temps que les régionales, il fallait pouvoir dissocier les deux scrutins lors de la préparation.

21. *Archelec 4* est un projet différent des précédents, puisqu'il n'inclut pas de numérisation. Il a été gagné à la suite de la réponse à un appel de l'infrastructure Collex-Persée dans l'axe des services à la recherche, pour un montant de 29 000 euros, en 2018. Actuellement en cours, il permet d'y associer étroitement des chercheurs utilisateurs des données pour leur fournir des outils adaptés. Le projet prévoit donc le découpage à la pièce des recueils de documents numérisés dans le cadre d'*Archelec 1*, des métadonnées plus nombreuses pour chaque unité documentaire, la création d'une base de logos de partis politiques extraits des professions de foi, la conception d'un tutoriel et l'organisation d'une journée d'études au printemps 2020. En revanche, une tension est apparue entre les besoins des chercheurs qui souhaitent accéder à des jeux de données et les nécessités de mettre à disposition une base de données documentaire permettant l'interrogation. Les deux sont donc prévus parallèlement.

22. *Archelec 5* enfin a permis la numérisation des 700 documents officiels et officieux produits lors des élections européennes de 1979 à 2014. Ils seront très bientôt mis à disposition. Ces cinq projets donneront donc accès courant 2020 à la pièce aux archives des élections présidentielles (1965-2012), législatives (1958-1993) et européennes (1979-2014). Mais toutes les élections ne sont pas couvertes chronologiquement sur l'ensemble de la Ve République, les différents types de scrutins ne sont pas représentés et toutes les typologies documentaires non plus. Le corpus mis en ligne est-il donc le reflet réel du fonds papier ?

Le corpus mis en ligne : reflet fidèle ou distorsion par rapport au fonds originel ?

Deux fonds différents

23. La nécessité de réfléchir aux questions juridiques en fonction des typologies de documents conduit à des choix qui ne relèvent pas de logiques scientifiques en matière de numérisation des archives électorales du CEVIPOF. Le public se trouve donc en face de deux fonds différents : l'ensemble papier, cohérent, relativement complet – même s'il n'est pas exhaustif –, diversifié, reflet d'une campagne électorale comme d'une activité de recherche, consultable sur place, d'une part, et le corpus virtuel numérisé, limité – selon le type d'élection, les bornes chronologiques et la typologie documentaire – qui n'est pas un miroir du fonds physique. Le principe archivistique du respect de l'intégrité du fonds est-il alors encore à l'œuvre ? Probablement pas car le fonds n'est pas disponible dans son entièreté.

24. Se posent également des questions scientifiques : si le chercheur n'a pas accès sous forme numérique à toutes les sources disponibles en papier, peut-il avoir une perception juste de l'ensemble du contenu du fonds d'archives électorales ? Il est alors indispensable de préciser la méthodologie retenue, les contraintes juridiques et les choix effectués en fonction de ces obligations. Il devient d'autant plus nécessaire de signaler d'autres fonds complémentaires pour pouvoir toujours croiser les sources[19]. Une communication précise et juste est essentielle pour valoriser le corpus numérisé mais aussi pour en indiquer les limites, à travers tous les moyens possibles (mails, listes de diffusion, écrans numériques, vidéos, réseaux sociaux, etc.). L'effort d'*Archelec 4* porte notamment sur cette dimension pédagogique, en prévoyant de fournir un tutoriel et en organisant une journée d'études le 1er juillet 2020 qui associe des chercheurs impliqués dans l'utilisation des documents numérisés.

25. Des effets déformants apparaissent également à l'usage. La recherche par mots-clés (département, circonscription ou nom, par exemple) aboutit à des résultats qui figurent tous sur le même plan. La structure arborescente n'apparaît plus, de même que le contexte.

19. Les documents sont à mettre en perspective avec les archives évoquées plus haut, mais aussi avec les fonds d'hommes politiques et de partis conservés aux Archives d'histoire contemporaine du Centre d'histoire de Sciences Po, avec les richesses de la bibliothèque de Sciences Po, les fonds de l'Assemblée nationale et du Sénat et ceux de La Contemporaine.

De nouveaux usages

26. Au-delà d'utilisations déjà existantes mais élargies et renouvelées comme celles de journalistes pour la rédaction d'articles locaux lors d'élections, notamment pour analyser si les promesses ont été tenues, la reproduction plus aisée de fac-similés dans le cadre d'expositions, comme à Verdun en 2018-2019, il est notable que de nouveaux usages sont apparus grâce à la mise à disposition en ligne, même si beaucoup d'utilisateurs ne sont pas connus, à la différence du lecteur venu consulter le fonds en salle de lecture. En octobre 2019, plus de 453 000 vues ont eu lieu depuis les débuts de la mise en ligne en octobre 2015, en provenance de pays variés, au-delà bien sûr de ce qu'il est possible de consulter sur place.

27. La mise en ligne de grandes séries de documents a permis leur appropriation et leur traitement quantitatif par une équipe de chercheurs de Berkeley et de Sciences Po (Caroline Le Pennec et Paul Vertier) pour une recherche longitudinale en économie portant sur « *Downsian Convergence on Non-policy Issues: Evidence from Campaign Manifestos at French Legislative Elections* ». Ces chercheurs sont associés à *Archelec 4* avec d'autres académiques de Sciences Po, notamment Martial Foucault et Nicolas Sauger, respectivement directeur du CEVIPOF et directeur du Centre de données socio-politiques de Sciences Po, pour préciser leurs besoins en termes de métadonnées à intégrer à la base. D'autres projets de recherche sont en cours pour mettre à profit *Archelec*. L'un consiste à associer ces données aux résultats électoraux. L'autre

est mené par Lou Safra, chercheuse au CEVIPOF en psychologie politique, et consiste à analyser les photographies des candidats pour évaluer le potentiel de vote en leur faveur.

Conclusion

28. La numérisation d'un fonds papier original d'éphémères, constitué par et pour la recherche, permet de toucher un nouveau public et d'exploiter différemment les sources. Si elle offre de formidables opportunités, elle implique aussi de nombreuses contraintes, juridiques et scientifiques. Le fonds numérisé n'est pas le reflet fidèle du fonds d'origine. Une réelle pédagogie est aussi nécessaire pour faire comprendre quelles sont les ressources disponibles en ligne par rapport à celles qui ne sont pas numérisées, voire qui ne le seront jamais, et quels sont les choix qui ont présidé à la mise en ligne. Mais la numérisation a aussi des retombées positives pour la connaissance et le rayonnement du fonds et elle a déjà permis plusieurs dons intéressants de tracts et des professions de foi, de volumes très variés. L'offre étant liée à la demande, le public est maintenant en attente de la mise à disposition d'autres corpus numérisés, notamment ceux des régionales ou des législatives plus récentes. Le financement d'*Archelec 4* a également permis un élargissement des services proposés aux chercheurs au-delà de la numérisation.

Le traitement numérique des sources : la construction des corpus et des instruments de recherche comme enjeu pour la mise à disposition des données

Céline Alazard, Jean Vigreux et Serge Wolikow

1. Avec la généralisation des pratiques numériques, toute la production scientifique est transformée. Les sources à partir desquelles les chercheurs travaillent, notamment les historiens, forment un domaine précocement concerné et durablement marqué par de nouvelles évolutions. Dans le processus de transition au numérique qui affecte l'ensemble du travail scientifique, la constitution des sources reste préalable et donc impactée par les nouvelles pratiques numériques d'abord dispersées avant d'être systématiques.

2. L'extension de l'usage par les historiens de la photographie numérique pour collecter les archives, le partage des données entre chercheurs et archivistes, entre laboratoires et dépôts d'archives dans le cadre de projets menés en commun sur des sources rares et jusqu'alors inaccessibles ont fait émerger la nécessité de produire des guides de bonnes pratiques et des « boîtes à outils ». Dans cette perspective, la Maison des sciences de l'Homme (MSH) de Dijon[1] a travaillé à la constitution de deux corpus à partir des archives du communisme et celles de la vigne et du vin, puis à la structuration et à la normalisation d'instruments de recherche* électroniques.

3. Il s'agit de programmes inscrits dans deux thématiques, « Critique et mouvements sociaux » et « Vigne et vin » de la MSH, labellisées comme collections d'excellence en 2017 par le GIS Collex-Persée*. L'objectif de ces travaux était en premier lieu d'élaborer les outils nécessaires au développement des recherches dans ces deux domaines. Cela impliquait en particulier la possibilité de lancer des requêtes croisées sur une documentation textuelle hétérogène ne permettant pas l'océrisation*. Au départ, il semblait que le travail d'indexation allait constituer le principal sinon le seul défi ; or, il est très vite apparu que la vérification des données, le nommage des fichiers, la structuration des données comme ensuite la mise en ligne des documents et leur lien avec les métadonnées* avaient été sous-estimés. C'est l'ensemble de ce processus pluriannuel, partant des sources pour aboutir à la mise en ligne des documents en passant par l'établissement des corpus et la création des instruments de recherche, que nous proposons d'expliciter dans cette communication, avec pour exemple les corpus « communisme » et « vigne et vin ».

1. Cf. http://msh-dijon.u-bourgogne.fr/

La constitution des corpus : archives et documentation

4. Avant de présenter les premiers chantiers mis en place à la MSH de Dijon, il est important de rappeler les problèmes posés par la définition même du terme de corpus. Si la définition du Centre national de ressources textuelles et lexicales (CNRTL) entend « recueil réunissant ou se proposant de réunir, en vue de leur étude scientifique, la totalité des documents disponibles d'un genre donné, par exemple épigraphiques, littéraires, etc.[2] », cette définition a évolué avec les initiatives institutionnelles et les préoccupations des chercheurs, notamment les premiers appels à projets ANR *Corpus et outils de la recherche en SHS* en 2006 et 2007. Le corpus apparaît alors comme un ensemble de documents caractérisé par une délimitation ou sélection nécessaire à l'analyse selon le processus méthodologique imposé par la discipline[3].

5. C'est dans cette perspective que les premiers travaux sur des corpus ont été entrepris à la MSH, relativement aux programmes de recherche alors engagés. La MSH de Dijon depuis 2004, et le consortium Archives des mondes contemporains (ArcMC[4]) depuis 2012, soutiennent et développent une démarche collaborative et collective pour répondre aux problèmes liés à la nature des archives et de la documentation de la période contemporaine (XIXe et XXIe siècles).

6. Ces dernières sont caractérisées tout à la fois par leur masse, l'hétérogénéité des supports (papier, microsillon, photographie, disquettes informatiques, fichiers du Web, etc.), leur fragilité, la diversité de leur statut juridique (public ou privé), et la dispersion des lieux où elles sont conservées. Par ailleurs, la situation de ces archives est très diverse : les unes ont déjà été indexées et numérisées, quand d'autres doivent encore être constituées en corpus. La numérisation et la mise en ligne constituent alors une réponse pertinente aux problèmes de conservation et d'accessibilité pour les chercheurs, problèmes qui résultent de la nature et de la situation même de ces archives. Dès lors il s'agit de prendre en considération toutes les étapes de la chaîne de production de la documentation numérique, depuis la numérisation des documents papier jusqu'au traitement des documents numériques natifs (texte, image, audiovisuels, etc.), afin d'aboutir à la constitution de corpus numériques inédits, et permettre leur mise à disposition et leur publication.

2. Cf. https://www.cnrtl.fr/definition/corpus
3. Ben Henda, Mokhtar. 2018. « L'ingénierie des corpus ». Cours en ligne présenté à Master Humanités numériques. https://cel.archives-ouvertes.fr/cel-01716602.
4. Cf. https://arcmc.hypotheses.org/a-propos-2

Les premiers chantiers : prise en charge et traitement des archives du communisme

7. En cohérence avec les recherches scientifiques menées par les chercheurs de l'université de Bourgogne sur l'histoire du communisme, un des premiers chantiers entrepris fut celui des archives de ce mouvement, et notamment du Parti communiste français (PCF), avec des réflexions et travaux alimentés lors des programmes *Incomka* (projet du Conseil de l'Europe) et *Paprik@2F* (Portail archives politiques recherches indexation Komintern et fonds français – ANR Corpus 2013-2016[5]).

8. Pour rappel, après l'effondrement de l'Union soviétique, les archives du communisme deviennent la propriété de l'État russe et sont ouvertes aux chercheurs – avec des inventaires approximatifs comme seuls outils – et selon des pratiques parfois commerciales filtrant la consultation. En 1992, le Conseil de l'Europe, dans le cadre de la politique générale d'aide aux nouvelles démocraties, mais aussi dans un souci de sauvegarde des fonds d'archives, se saisit de la question des archives. C'est ainsi que le 8 mai 1992, lors d'une séance du Conseil de l'Europe à Strasbourg, est constituée une commission mixte d'historiens et d'archivistes qui lance conjointement un plan d'action au niveau européen.

9. Le projet est alors piloté par un comité international, l'Incomka pour International Committee for Computarization of Comintern Archives. Ce comité, dont fait partie Serge Wolikow, chercheur à l'université de Bourgogne, est composé de représentants du Conseil de l'Europe, des Archives de la Fédération de Russie, du Conseil international des archives, et de huit organisations partenaires : la direction des archives de France, les Archives de la République fédérale d'Allemagne, les Archives fédérales suisses, les Archives nationales de Suède, les Archives nationales d'Italie, la Bibliothèque du Congrès de Washington (États-Unis d'Amérique) et l'Open Archives Society de Budapest. Le Centre Georges Chevrier (UMR CNRS uB 5605) de l'université de Bourgogne, en partenariat avec les Archives de France, a quant à lui participé au processus en apportant son expertise.

10. Le projet *Incomka* a permis la mise en œuvre et la réalisation de deux projets distincts, tout en les connectant :

- la numérisation de 5 % des fonds d'archives (soit un million de pages) dont celles du Komintern (l'Internationale communiste), permettant ainsi d'envisager l'histoire des sections nationales de la IIIe Internationale, avec pour la France celle de la Section française de l'Internationale communiste (SFIC), puis du Parti communiste français (PCF). Les pièces les plus significatives dans l'ensemble des fonds de l'Internationale communiste ont été sélectionnées pour bénéficier de la numérisation
- la numérisation des inventaires des fonds de l'Internationale communiste conservés à Moscou et leur transformation en une base de données indexée par thèmes et donnant accès aux documents numérisés

5. Cf. https://anrpaprika.hypotheses.org/

11. Dans le prolongement de ces travaux, le programme de recherche ANR *Paprik@2F*, qui s'est déroulé de 2013 à 2016, sous la direction de Jean Vigreux, a eu pour objectif la création d'un outil de recherche dans les archives politiques du PCF, hébergé sur le portail PANDOR (Portail archives numériques et données de la recherche[6]) de la MSH de Dijon. La principale originalité de l'outil mis en place est de rendre accessibles des informations jusqu'alors dispersées en offrant une seule porte d'entrée vers des informations conservées dans des lieux différents. Le portail ainsi constitué assure à la communauté scientifique la pérennité de l'accès aux données. *Paprik@2F* s'est aussi appuyé sur les ressources des Archives nationales, en particulier sur les archives de surveillance ou de répression et offre ainsi un panorama le plus exhaustif possible des ressources disponibles. Dans le cadre de ce programme national, les Archives nationales ont pu faire l'inventaire complet, mais aussi la numérisation et la restauration de documents qui étaient jusqu'alors inaccessibles car sous scellés (ceci est le cas par exemple des fonds de la section spéciale de la cour d'appel de Paris, série Z/4).

12. La question de la construction du corpus à partir de bases et de types de documents hétérogènes est un enjeu majeur du programme. La base de données Incomka contient le fonds dit « 517-1 », regroupant les archives de la SFIC, conservées aujourd'hui aux Archives d'État russes d'histoire socio-politiques (RGASPI). Ce fonds a fait l'objet de plusieurs actions de rapatriement dès les années 1970 au gré des relations entre le PCF et le Parti communiste de l'Union soviétique et c'est ainsi qu'une copie partielle (865 dossiers sur les 2 055 existants) de documents originaux du fonds 517 a été transmise progressivement par les autorités soviétiques à l'Institut Maurice-Thorez entre 1972 et 1976, puis à l'Institut de recherches marxistes entre 1983 et 1986. Cette copie comprend environ 100 000 clichés pour la période 1921-1939, dont en particulier les procès-verbaux et décisions des organes de direction du PCF (Comité directeur, Comité central, Bureau politique, Secrétariat), soit environ 22 000 pages, mais également les documents concernant les relations entre la direction nationale et les structures régionales du PCF (notes, rapports, correspondance, propagande, etc.), l'activité de secteurs et commission de travail (le travail en direction des femmes, la main-d'œuvre étrangère, les paysans, l'activité anticoloniale, les écoles et la formation des militants, l'agitprop, les journaux de cellule), soit environ 56 000 pages.

13. Initialement sous forme microfilmée, ce fonds a fait l'objet d'un ample programme de numérisation et d'indexation réalisé par la MSH en partenariat avec, dans un premier temps, la Bibliothèque marxiste de Paris (BMP) et dans un deuxième temps les Archives départementales de la Seine–Saint-Denis.

14. Concernant les fonds des organismes centraux du Komintern, ont été menées de minutieuses opérations d'identification et de sélection des documents et notices qui renvoyaient au fonds français dans la base de données

6. Cf. https://pandor.u-bourgogne.fr/

Incomka à l'activité des communistes français et aux liens entre le PCF ainsi que les autres Sections de l'Internationale communiste, ceci dans le but de les extraire et de les transformer en instruments de recherche encodés en XML*-EAD*. À ces documents ont été ajoutés de nouveaux fonds grâce à des missions spécifiques de rapatriement organisées au RGASPI à Moscou.

15. Ce processus a été répliqué pour le fond 517-1, qui a été indexé dans une base « 4D » réalisée à la fin des années 1990 puis reprise et retravaillée par la MSH à l'occasion du partenariat autour du programme *Paprik@2F*. Dans la perspective d'une mise à disposition facilitée et généralisée des documents, la poursuite des opérations d'extraction et de conversion des notices de la base « 4D » en XML-EAD est actuellement en cours dans le cadre du programme de recherche *ABRICO*[7] (*Archives brochures et informations communistes*) financé dans le cadre d'une réponse à l'appel à projets « Numérisation » de Collex-Persée en 2018.

Les corpus vigne et vin : une démarche documentaire pour une recherche pluridisciplinaire

16. Les recherches menées sur la vigne et le vin ont également débouché sur la constitution de corpus et prolongé la réflexion méthodologique entreprise à Dijon. Pour ce faire, la MSH s'appuie sur le groupe thématique de recherche « Vigne et vin » qu'elle structure. Il s'agit d'un lieu d'études pluridisciplinaires qui a entre autres objectifs la création d'outils de recherche s'appuyant sur les plateformes technologiques de la MSH, permettant ainsi la prise en charge et le traitement d'archives publiques et privées, de la documentation des milieux professionnels, et la constitution d'une bibliographie internationale.

17. C'est ainsi que des programmes de recherche successifs ont permis la constitution de ressources numériques mises à la disposition des chercheurs, mais également des professionnels de la filière vitivinicole et du grand public, sur le portail de publication PANDOR. La première action a eu lieu en 2008 avec la numérisation et le traitement archivistique d'une partie des archives de l'INAO (Institut national des appellations d'origine[8]), organisme créé en 1935 regroupant les services du ministère de l'Agriculture et les représentants des milieux professionnels de la viticulture et du vin. Cette action s'est inscrite dans la perspective d'une mise en place de normes communes de production, adaptées cependant à la situation des différents vignobles.

18. Le sous-fonds « Archives des comités nationaux », composé des documents de travail et des comptes-rendus des séances des différents comités nationaux depuis leur création, a été entièrement numérisé. La période couverte s'étend de 1935, date de création du Comité national des appellations d'origine pour les vins et eaux-de-vie

7. Cf. http://www.collex.eu/wp-content/uploads/2018/11/Fiche_presentation_ABRICO.pdf

8. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00005

(CNAO), à la décennie 1980. Ces archives sont des sources essentielles pour l'histoire technique, sociale et économique du vin, des eaux-de-vie, des cidres, des poirés et de l'ensemble des produits agroalimentaires français, mais elles sont également décisives pour l'histoire spécifique de l'Appellation d'origine contrôlée et de l'organisme garant de ce statut, à savoir l'INAO.

19. Quelques années plus tard, en 2014, en réponse à un appel à projets de BSN 5 (appel à projets de la Bibliothèque scientifique numérique dédié à la numérisation), la MSH a été lauréate avec son projet *Numérisation, traitement documentaire et mise en ligne des bulletins de l'Organisation internationale de la vigne et du vin (OIV)*[9]. Ce programme a donc consisté en la numérisation, l'indexation et la mise en ligne de la collection complète des bulletins de l'OIV de 1928 à 1999. En tant qu'organisme technique, juridique et scientifique international, l'OIV travaille à l'harmonisation et à l'élaboration des pratiques et normes mondiales, à l'amélioration des conditions d'élaboration et de commercialisation des produits vitivinicoles. La MSH a ainsi élaboré un outil interdisciplinaire de recherche destiné à la communauté scientifique, au grand public et aux professionnels de la vigne et du vin. Cet instrument patrimonial unique permet d'appréhender les évolutions juridiques, techniques, scientifiques et commerciales touchant à la vigne et au vin dans les pays membres de l'OIV ainsi que dans les interrelations entre pays producteurs ou consommateurs de vins dans le monde au XXI[e] siècle.

20. Plus récemment, la MSH a été lauréate d'un deuxième projet dans le cadre de l'appel à projets « Numérisation » de Collex-Persée avec le programme *Convex* (Collection numérique vitivinicole d'excellence[10]) actuellement en cours de réalisation au sein de la plateforme ADN (Archives – documentation – numérisation), permettant la poursuite des travaux d'ores et déjà entrepris sur la thématique. La construction de ces différents corpus et de leurs instruments de recherche associés prolonge la démarche documentaire amorcée à Dijon. Le caractère diversifié et original des nouveaux corpus à traiter a suscité le développement d'approches holistiques et multiscalaires d'un grand intérêt heuristique. Ces corpus rendent possible une meilleure connaissance de l'histoire des sciences et techniques. L'exploitation de ces données permet de mieux appréhender, au niveau national comme international, la construction et l'ancrage des normes de production et de commercialisation des vins, offrant des outils d'analyse rares aux juristes, historiens, géographes ou économistes, ainsi qu'aux professionnels du vin.

21. C'est bien dans cette perspective que sont actuellement numérisés, décrits et indexés le *Bulletin international de la répression des Fraudes* (1910-1916) et les *Annales des fal-*

9. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00019

10. Cf. http://www.collex.eu/wp-content/uploads/2018/11/Fiche_presentation_CONVEX.pdf

sifications et des fraudes (1917-1938[11]). Ces deux titres font référence à la revue technique et informative publiée par la Société des experts chimistes de France. La part de la collection concernée par le projet couvre la période 1908 à 1938, le choix chronologique de la collection à numériser ayant été dicté par plusieurs facteurs. Alors que 1908 constitue l'année du lancement du *Bulletin de la répression des fraudes*, les années 1935-1938 s'imposent comme un moment charnière de mutation des réglementations viticoles, avec le décret-loi du 30 juillet 1935 sur les AOC ainsi que la loi Chouffet du 13 janvier 1938, et de bouleversements dans le contrôle et l'expertise de ces normes, avec la mise en place entre 1937 et 1938 de la brigade spéciale chargée de la surveillance des vins et eaux-de-vie à AOC sur le territoire français.

22. Cette revue voit le jour dans une période cruciale de transformation de la vitiviniculture mondiale, période d'intense mise en réglementation des productions, de leur commercialisation et moment d'affirmation des sciences chimiques, de la pharmacologie, de la médecine et de l'œnologie dans l'expertise des denrées alimentaires. Ce fonds vient donc compléter les sources à caractères réglementaires, techniques et scientifiques présents dans les documentations de l'INAO et de l'OIV d'ores et déjà accessibles via le portail PANDOR de la MSH de Dijon, offrant ainsi un champ élargi pour les recherches liées à ces thématiques.

11. En 1917, le *Bulletin de la répression des fraudes* est absorbé par les *Annales des falsifications et des fraudes*.

La construction des instruments de recherche et les enjeux de l'indexation

Rendre cohérents et interrogeables de grands ensembles hétérogènes

23. L'exploitation de ces corpus et archives numériques permettant d'accéder à des pièces spécifiques à l'intérieur de grands volumes documentaires, de même que le traitement quantitatif de données hétérogènes implique le développement d'outils spécifiques de recherche.

24. Dans le cadre du programme *Paprik@2F* par exemple, la rencontre entre chercheurs et archivistes a tout d'abord permis d'inventorier les fonds existants, puis de mettre en regard deux ensembles hétérogènes : les archives nationales et les archives du Komintern. Ce projet s'est appuyé sur les compétences de la plateforme technologique ADN pour Archives – documentation – numérisation[12] de la MSH de Dijon pour publier des instruments de recherche en XML-EAD ou adapter les documents textes existants et exposer les métadonnées dans un entrepôt OAI* en Dublin Core*.

25. Cette plateforme est labellisée par le Réseau national des maisons des sciences de l'Homme[13] (RNMSH) depuis 2012 (aujourd'hui l'une des plateformes du réseau Scripto) et

12. Cf. http://msh-dijon.u-bourgogne.fr/offre-technique/plateforme-adn.html
13. Cf. https://www.msh-reseau.fr/

reconnue par l'université de Bourgogne depuis 2013. Elle fait partie des 33 plateformes technologiques labellisées par l'établissement[14], tous domaines de recherche confondus. ADN propose une chaîne de traitement complet des données, qu'elles soient sur support traditionnel (textes papier, images fixes et mobiles, sons) ou nativement numériques. Cette chaîne comprend toutes les étapes permettant de transformer ces données en ressources intelligentes : indexation, préparation pour la fouille de données, publication, etc.

26. Le stockage et la pérennité de l'archivage des données numériques sont assurés par la MSH de Dijon en étroite collaboration avec la Direction du numérique de l'université de Bourgogne, suivant les recommandations de la TGIR Huma-Num[15]. Cette démarche s'inscrit ainsi dans le cadre des Humanités numériques en rendant l'ensemble des ressources numériques interrogeables via l'interface de publication de la MSH.

27. Au cours du programme *Paprik@2F*, les ressources destinées à alimenter le portail PANDOR ont fait l'objet d'un travail préalable de recensement des formats d'origine et un cahier des charges précis concernant le format de destination choisi a été établi au moment même de la constitution du corpus. Cette expérience a confirmé l'importance d'un traitement documentaire et archivistique préalable à la numérisation.

14. Cf. https://www.u-bourgogne.fr/recherche-scientifique/plate-formes-technologiques-du-grand-campus

15. Cf. https://www.huma-num.fr/

28. Au final, le travail archivistique a consisté dans la mise en ligne, sur PANDOR, de 5 instruments de recherche en XML-EAD (notices et images) concernant les fonds du monde communiste français de 1917-1947 (fonds qui seront présentés dans la partie suivante).

29. Autre exemple, le traitement numérique des formes brèves de l'imprimé a mis en évidence la nécessité d'entreprendre en amont des réflexions méthodologiques et de faire des choix technologiques en vue de proposer un outil adapté permettant de rendre interrogeable et intelligible un ensemble documentaire conséquent. La collecte numérique à la base de ce programme a impliqué la mise au point de bonnes pratiques fondées sur l'expérience de chercheurs comme de plusieurs laboratoires et de MSH qui, associés, ont travaillé de manière collaborative dans le consortium Archives des mondes contemporains (ArcMC), labellisé par la TGIR Huma-Num de 2012 à 2016 et aujourd'hui structuré en réseau.

30. Car la question des formes brèves de l'imprimé constitue un autre chantier pour la réflexion menée sur « archives et documentation ». Ces imprimés ont en effet en commun d'avoir été composés souvent dans l'urgence en fonction de l'actualité : il s'agit de documents très divers tels que des brochures, des tracts, des affiches, des caricatures, etc. Ces documents hétérogènes, constitués en corpus, permettent des recherches combinant histoire du politique, histoire des idées et histoire de l'édition en prenant en considération également la diffusion des textes.

31. C'est ainsi que la MSH et le consortium ArcMC ont associé chercheurs et ingénieurs pour réfléchir conjointement au traitement et à la valorisation de ces gisements documentaires « en sommeil » ou en déshérence. À ce jour, 3 000 des 20 000 brochures de l'ancienne Bibliothèque marxiste de Paris sont d'ores et déjà en ligne sur le portail PANDOR de la MSH de Dijon[16] et 1 000 supplémentaires seront publiées en 2020, également intégrées au programme *ABRICO* précédemment mentionné. Mais quels sont les critères de « sélection » dans un ensemble documentaire si volumineux ? Plutôt que de procéder à un échantillonnage qui porterait atteinte à l'intégrité des fonds, la sélection s'est appuyée sur la cohérence de contenus.

32. L'ambition a été de proposer un corpus important pour l'après Seconde guerre mondiale, non seulement pour mesurer l'implantation du PCF alors premier parti politique de France mais également pour appréhender au mieux ses logiques d'activités dans les contextes de la guerre froide, de la décolonisation et des problèmes sociaux du fordisme. L'objectif est également de prendre en compte la brochure comme forme spécifique de l'imprimé de propagande. Les sélections ainsi constituées permettent actuellement de procéder à la recherche portant sur une période doublement décennale (1945-1968). En résumé, si le traitement numérique et l'élaboration des instruments de recherche, tout en s'inspirant de principes communs, restent différenciés en fonction de l'origine et de la nature de la documentation, c'est au niveau de l'interface de publication que les ressources sont intégrées et globalement interrogeables.

16. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00009

Publication et diffusion : perspectives de partage

33. Pour rendre possible la mise à disposition des ressources constituées et traitées à la communauté académique et au grand public, la MSH s'est dotée du portail PANDOR, lancé en 2014, qui est un puissant outil d'interrogation et de valorisation des ressources numériques gérées par la plateforme ADN. Il permet la publication et un accès unifié aux corpus numériques créés par la plateforme ADN dans le cadre de programmes de recherche portés ou soutenus par la MSH.

34. Il s'agit d'instruments de recherche archivistiques, catalogues de bibliothèque normalisés en XML-EAD et documents numérisés. Il propose les fonctionnalités suivantes :

- la navigation à l'intérieur du cadre de classement – en regard des axes thématiques portés par la MSH
- la navigation à l'intérieur du plan de classement au niveau de l'instrument de recherche
- la navigation par index
- la recherche en texte intégral dans les notices
- la recherche en texte intégral dans les documents numérisés (module XML ALTO*)

- des formulaires de recherche simple et avancée
- une interopérabilité de l'outil avec les grands portails nationaux via le protocole OAI-PMH (PANDOR est moissonné par Isidore* et prochainement par France Archives).

35. PANDOR permet ainsi aux chercheurs, aux collectivités et institutions, aux entreprises, mais aussi au grand public de localiser et de consulter un ensemble de données (en fonction des conditions d'accès propres à chaque corpus), le plus souvent inédites, issues de programmes de recherche pluridisciplinaires. Il couvre les champs thématiques des sciences humaines et sociales représentés à la MSH et intègre tous les types de données (textes, images fixes et mobiles, son, multimédia), qu'elles soient natives ou le fruit d'une numérisation. En répondant aux standards internationaux en matière de traitement de données, PANDOR permet, grâce à une fine description des contenus, de repérer des documents difficiles d'accès. Témoin et acteur de la recherche développée à la MSH de Dijon, il inclut aussi des archives et des productions de chercheurs. De fait, cet outil rend possibles des recherches croisées sur des fonds de typologies hétérogènes.

36. Concernant plus spécifiquement la thématique labellisée par Collex-Persée « Critique et mouvements sociaux », PANDOR met à disposition un ensemble conséquent de ressources : outre les fonds évoqués précédemment et le programme *Paprik@2F* (brochures de la Bibliothèque marxiste de Paris[17], fonds de la Section française de l'Internationale communiste[18] [1922-1939], fonds de la direction du PCF[19] [1922-1939], fonds de l'Internationale communiste[20] [1917-1947], fonds des Archives nationales[21]), la collection regroupe et propose des revues et périodiques tels que les *Cahiers d'histoire*[22] (1966-2001), *Correspondance internationale*[23] (1921-1939), *Économie et Politique*[24] (1954-1999), *Nouvelle Critique*[25] (1948-1980) ou encore *Société française*[26] (1981-1999). Cette collection met en relief les difficultés rencontrées pour la mise en place de traitements documentaires communs sur des sources hétérogènes. C'est le choix d'une structuration basée sur le XML-EAD, initialement réservé aux archives, qui a permis une recherche globale mais fine sur un ensemble regroupant brochures, revues et archives.

37. Pour la collection « Vigne et vin », la dématérialisation, le traitement documentaire et la publication des corpus présentés plus haut (INAO, bulletins de l'OIV et Convex) offrent en premier lieu, un accès large à des res-

17. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00009
18. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00034
19. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00036
20. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00033
21. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRAN_IR_050130 et https://pandor.u-bourgogne.fr/ead.html?id=FRAN_IR_054916
22. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00008
23. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00032
24. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00026
25. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00048
26. Cf. https://pandor.u-bourgogne.fr/ead.html?id=FRMSH021_00010

sources difficiles d'accès, et pour certaines, rares, voire quasi indisponibles. Il s'agit par là d'un travail de préservation d'un patrimoine écrit, mais l'objectif est avant tout scientifique puisque ces différents programmes traitent des ensembles documentaires qui contribuent à une meilleure connaissance de l'histoire de la vigne et du vin sur des échelles variées. Ces corpus offrent de nombreuses entrées de recherche allant au-delà du vin, et questionnent, par exemple, la place des savoirs scientifiques « objectifs » dans la production des prescriptions, des décisions et des expertises. Cette documentation s'impose également comme un outil exceptionnel d'appréhension de la construction et de l'évolution des territoires viticoles (et plus largement agricoles), mais aussi de leurs réglementations, de leurs marchés et, plus généralement des normes et représentations qui les traversent, au niveau local, national et international.

38. Si ces corpus sont destinés en premier lieu aux chercheurs étudiant la vigne et le vin, la large palette des sujets concernés peut également capter l'attention d'autres spécialistes travaillant, par exemple, sur l'histoire des sciences, sur celle du droit, de l'alimentation, ou encore le statut de l'expertise et de la prescription. Cette documentation intéresse également l'ensemble des acteurs institutionnels ou économiques de la vigne et du vin, tout autant qu'un large public souhaitant consulter ces ressources.

39. L'objectif affiché à moyen terme est de participer à la constitution d'un centre de ressources internationales rassemblant pour la première fois un ensemble exhaustif de documentations historiques sur la vigne et le vin et qui a pour vocation de devenir une référence internationale qui pourra devenir un point d'appui incontournable pour la diffusion des données mondiales sur cette thématique.

40. Autre outil de publication, le site ArcMC[27], créé par le consortium du même nom et hébergé par la TGIR Huma-Num, signale en tout 70 fonds ou bases de données en ligne appartenant aux partenaires et aux équipes associées qui le composent et sont regroupés en pôles thématiques. Ainsi, les ressources concernant les thématiques abordées aujourd'hui sont visibles dans le pôle « Mouvements sociaux et organisations ouvrières » du site web ArcMC qui, outre les fonds publiés sur PANDOR, signale certains ensembles documentaires de La Contemporaine, de la Fondation Jean Jaurès, du Centre d'histoire sociale du XX^e siècle et de Ciné-Archives.

41. Le pôle ArcMC « Les mondes ruraux et du vin » signale quant à lui les fonds disponibles sur PANDOR grâce à des liens hypertextes. Force est de constater qu'une même base de données ou un même instrument de recherche archivistique ou catalogue peut être visible sur plusieurs interfaces, soit par moissonnage (exemple Isidore) soit par signalement et des renvois à l'aide de liens, accroissant par là même la valorisation des contenus et leur partage.

27. Cf. http://arcmc-corpus.huma-num.fr/

42. Les méthodologies et développements technologiques ici présentés et ceux envisagés doivent encore se poursuivre et s'étoffer (fouille de donnée, balisage TEI*) sur des corpus tels que les brochures de la Bibliothèque marxiste de Paris par exemple. Car avec près de 5 000 brochures publiées prochainement et couvrant une période de 70 ans et un large spectre idéologique et politique, ce corpus se prête à des études comparées et méthodiques sur les évolutions du langage politique mais aussi sur la diversité des lexiques politique dans un même moment historique (temps forts de la crise économique des années trente, du fascisme, de la révolution russe, etc.). Quant aux bulletins de l'organisation internationale de la vigne et du vin des années 1930 à 2000, ils permettent une étude comparée des différents savoirs mobilisés et formalisés : dans le domaine juridique, économique, géographique mais aussi de la chimie et de la biologie, et ainsi de travailler sur l'application des savoirs scientifiques à un domaine spécifique tel que celui du vin. Ces exemples impliquent des expériences qui depuis une décennie procèdent d'une démarche co-construite par les chercheurs et les ingénieurs professionnels des archives ou de la documentation. De nouveaux programmes de recherche, engagés pour 2020-2022 dans le cadre du GIS Collex-Persée, explicitent encore davantage ces démarches autour d'objets numériques avec des corpus spécifiques dont la construction procède d'une démarche scientifique et pas uniquement de la rétroconversion de corpus papier préexistants. C'est le cas en particulier du projet de bibliothèque numérique multilingue et internationale du vin, de la constitution d'une collection numérique des congrès du Parti communiste français sur près de cent ans ou encore celle des cahiers du Centre d'études et de recherches marxistes (CERM) des années 1960 et 1970.

43. À ces perspectives scientifiques et technologiques s'ajoute celle de la constitution de réseaux autour de ces corpus numériques, qui appelle le développement et le partage de préconisations et d'outils communs autour de programmes de recherche fondés sur ces ressources. Les réseaux thématiques en humanités numériques du type Réseau Archives des mondes contemporains sont ainsi propices à l'émergence de collaborations et partenariats entre institutions, nécessairement de statuts différents (universités, laboratoires de recherche, services d'archives publiques, bibliothèques, etc.) travaillant de concert à l'élaboration de nouvelles méthodologies. Il s'agit d'instituer des modes de partage des savoirs et des méthodes, comportant différents volets : l'acquisition et la diffusion de nouveaux outils, la formation des jeunes chercheurs, le partage des données et la confrontation des savoirs. Pour mener à bien ces opérations, il semble opportun et souhaitable que des réseaux de ce type puissent prendre appui, pour pérenniser leur structuration, sur les MSH et leur réseau national.

Les corpus textuels numériques (re)spécifiés

Damon Mayaffre

Introduction

1. Les années 2000 ont vu le triomphe du corpus en linguistique et, par-delà, dans les SHS. Non pas que l'objet corpus n'ait existé de longue date auparavant, non pas que les linguistes l'aient ignoré jusqu'alors mais au sens où la linguistique *sans* ou *hors corpus* apparaît à ce moment-là comme une spéculation intellectuelle marginale pratiquée seulement par une minorité. Sémanticiens, phonologues, lexicologues, dialectologues, etc. se revendiquent tous du corpus ; même la syntaxe générative semble concernée comme par exemple dans le numéro « La syntaxe de corpus » de la revue éponyme *Corpus* au début du siècle.

2. Et la fièvre *corpus* qui a saisi l'hexagone scientifique, sinon le monde, au début des années 2000 n'est jamais retombée, attestant qu'il s'agissait plus qu'un effet de mode. On lira ainsi à titre d'exemple un premier bilan documenté dans (Laks 2008), on consultera les actes du colloque thématique du Cercle belge de linguistique (Mellet et Longrée 2009) ; on mentionnera le 23e colloque international du Cercle linguistique du Centre et de l'Ouest (université de Poitiers – 5 et 6 juin 2009) intitulé « L'exemple et le corpus : quel statut ? ». À l'échelle internationale le *peer-reviewed journal* intitulé *Corpora*, créé en 2006, fait paraître son 15e volume en 2020. À l'échelle nationale, la revue *Corpus* créée en 2001 fait paraître ses 20e et 21e numéros cette même année. À Lorient, Poitiers, Grenoble, les Journées annuelles de linguistique de corpus (JLC) ont tenu leur 10e session en 2019, et les Journées internationales biannuelles d'analyse de données textuelles (JADT) se sont imposées dans le concert national et européen avec un millésime à Nice en 2016, à Rome en 2018 et à Toulouse en 2020. Et ces indices sur l'engouement *corpus* ne sont pas exhaustifs, puisque nous pourrions ajouter, autres exemples, la parution récente de *La Mesure et le Grain. Sémantique de corpus* de François Rastier (2011) ou de *Explorer un corpus textuel : Méthodes, pratiques, outils* de Frédéric Landragin et Céline Poudat (2017), prolongement certain, 20 ans après, de l'ouvrage de Benoît Habert, Adeline Nazarenko, André Salem, *Les Linguistiques de corpus* (1997).

3. Cependant, loin de tirer un bénéfice direct du triomphe du corpus, la linguistique de corpus stricto sensu au sens par exemple de (Aijmer et Altenberg 2004 ; Biber, Conrad et Reppen 1998 ; Habert, Nazarenko et Salem 1997 ; Kennedy 1998 ; Partington 1998 ; Partington, Morley et Haarman 2004 ; Sinclair 1991 ; Tognini-Bonelli 2001 ; Rastier 2001, 2005, 2011 ; Williams 2005 ;

Landragin et Poudat 2017)[1] s'en trouve ébranlée comme si, par une ruse de l'histoire scientifique, la banalisation de son objet avait brouillé voire dissout son identité propre. Pire : revendiqués désormais par tous, les corpus semblent ne plus appartenir scientifiquement à personne jusqu'à devenir inopérants. A minima l'objet *corpus* demande aujourd'hui à être (re)spécifié.

4. C'est donc dans ce cadre paradoxal – triomphe des corpus ; dilution de la linguistique de corpus – que cette contribution essaye de (re)questionner l'objet *corpus textuel numérique*, dans sa dimension textuelle comme dans sa dimension numérique. Les 5 portraits proposés (1. « Le corpus comme matrice » ; 2. « Le corpus comme contexte » ; 3. « Corpus réflexifs et herméneutique endogène » ; 4. « Le corpus comme texte » ; 5. « Corpus et numérique ») tendent vers une exigence méthodologique forte qui sera partout suggérée sans pouvoir être hélas aboutie nulle part dans cette contribution. Dès lors qu'il est théoriquement établi et empiriquement constitué, tel que nous allons essayer de le discuter, comment traiter scientifiquement un corpus textuel numérique ? Quelles méthodes et quels logiciels utiliser ? Le *deep learning** aujourd'hui peut-il compléter les parcours de la logométrie traditionnelle ? etc. Ces aspects méthodologiques introduits dans cet article sont traités ailleurs, et nous ne pouvons renvoyer ici le lecteur qu'à la bibliographie[2].

Le corpus comme matrice

5. Travailler sur corpus, c'est vouloir travailler sur des données attestées ; sans quoi l'introspection suffirait. Comme les seules données langagières attestées – c'est-à-dire les seules performances linguistiques abouties qu'un locuteur produit lorsqu'il s'exprime – sont les discours[3] puis, en tant que formes empiriques et stabilisées du discours, les textes, nous ne considérons, dans cette contribution, que les corpus textuels[4].

6. Et peut-être est-ce la nature nécessairement textuelle des corpus traités qui modifie en substance les choses et notre réflexion ?

7. Avec François Rastier, nous pensons qu'un texte n'a pas de signification, qu'une grammaire formelle du texte est vaine, que les unités textuelles (quand bien même reposent-elles sur les formes matérielles graphiques repérables) sont mouvantes, plurielles, complexes, par-

1. Il va de soi que ces auteurs divergent entre eux sur certains points. Nous les rassemblons ici en tant que linguistes ayant pris à bras le corps l'objet *corpus* et l'idée d'une linguistique de corpus. Quelques grands ancêtres pourraient être ajoutés comme Firth ou Palmer.

2. Pour commencer, peut-être, trois ouvrages récents et un article : (Lebart, Pincemin et Poudat 2019 ; Mayaffre et Vanni 2021 ; Née 2017 ; Mayaffre, Pincemin et Poudat 2019).

3. « Discours » au sens large donc, c'est-à-dire aussi bien des monologues que des dialogues ou des séquences interactionnelles longues ou courtes.

4. Répétons : un locuteur ne dit pas des mots isolés, il ne fabrique pas des phrases grammaticales : il produit des discours, aussi courts soient-ils, qui prennent la forme empirique de textes analysables.

fois discontinues et opèrent à différents niveaux de granularité linguistique (lexique, grammaire, syntaxe, graphie, pragmatique). Un texte n'a pas de signification mais un sens (ou plutôt des sens) qu'il ne s'agit pas de re-trouver mais de co-construire dans des parcours de lecture contrôlés. En linguistique du texte, il s'agit donc moins d'établir ou de restituer, que d'interpréter[5].

8. Partant, un corpus de textes n'est pas une *base de données* à interroger : le sens ne se laisse pas enfermer dans une *base* ; le sens n'est jamais *donné*. Les corpus textuels apparaissent ainsi très différents de certains corpus-ressources lexicographiques ou phonologiques qui consignent dans de vastes tableurs leurs données. Par la nature de ses composants – les textes –, le corpus textuel n'est donc pas une banque[6] ou une base de *data*. C'est un lieu, lui-même construit, où s'échafaude le sens, où se scénarise l'interprétation. En d'autres termes, nous pensons que le corpus est *moins le réceptacle du sens que sa matrice* ; moins un observatoire d'une langue qui serait déjà-là, qu'un observé vivant, mouvant, dynamique qui par sa constitution même et par son organisation, produit un sens toujours à inventer.

9. Cette première affirmation semble avoir des conséquences majeures. Nous en relèverons ici succinctement deux seulement.

10. D'un point de vue méthodologique d'abord, la méthode pour traiter ce corpus-matrice – la logométrie ou le *deep learning* notamment – prendra une valeur heuristique plus que probatoire : interroger plutôt que prouver, interpréter autant qu'établir. Notre travail s'inscrit à l'intérieur d'une linguistique de corpus à vocation herméneutique – l'herméneutique matérielle numérique[7] – et non dans le traitement automatique de la langue (TAL*). Si la logométrie est, loin de l'impressionnisme, une méthode formalisante s'appuyant fermement sur le matériel du texte, elle formalise moins des données que des parcours interprétatifs : la nuance est fondamentale.

11. D'un point de vue épistémologique ensuite, le corpus-matrice redéfinit notre posture face au texte, et le sens de la

5. La dimension herméneutique de la linguistique (textuelle) est, on le comprend dès à présent, au cœur de cette contribution. Elle est au centre de l'œuvre de Rastier et particulièrement de (Rastier 2001). Nous sommes marqué par le programme qui ouvre le chapitre IV « Herméneutique matérielle » : « Il reste à unir, au sein d'une sémantique des textes, les acquis de la philologie et de la linguistique comparée, pour restituer aux sciences du langage leur statut de disciplines herméneutiques » (p. 99). Nous retrouvons le même type de programme sous la plume de (Adam 2008, 30) lorsque l'auteur parle « du tournant herméneutique et plus largement de l'ouverture de la linguistique à l'interprétation ». Ajoutons que notre dette scientifique a été contractée, en amont de ces linguistes du texte, envers Jacques Guilhaumou (2006). En France, c'est lui qui a engagé précocement l'école française d'analyse du discours dans ce *tournant herméneutique* décisif.

6. Ainsi, contrairement à la terminologie anglo-saxonne, nous distinguons clairement *banque* de textes (ressource matérielle collective dans laquelle le chercheur pourra puiser ses textes) et *corpus* de textes (objet construit de manière *ad hoc* par lequel le chercheur problématise sa recherche). Le British National Corpus, le Brown Corpus ou le Lancaster-Oslo-Bergen (LOB) Corpus relèvent pour nous des banques de textes au même titre que Frantext, la BMF ou le LASLA.

7. Voir le titre de notre thèse HDR dont cet article est largement issu : Mayaffre, Damon. 2010. « Vers une herméneutique matérielle numérique. Corpus textuels, Logométrie et Langage politique ». HDR, Nice, France : Université Nice Sophia Antipolis. https://tel.archives-ouvertes.fr/tel-00655380.

démarche. À moins d'aspirer à une démarche paradoxale et circulaire, qui consisterait à supposer (à « hypothéser ») un sens déjà-là qu'il suffirait de rechercher et d'établir, tout en prétendant que le corpus produit un sens qu'il nous resterait à inventer et à co-construire, notre mode heuristique demande à être renversé. A minima, nous parlerons, avec (Tognini-Bonelli 2001) et avec toute la littérature anglo-saxonne, d'études *corpus-driven* (*versus* d'études *corpus-based*) : un corpus qui n'est pas une ressource que l'on soumet à l'interrogatoire mais un objet qui nous dirige dans le questionnement. Plus hardiment, nous parlerons d'un retournement de la méthode hypothético-déductive qui domine en SHS : là où l'on interrogeait *top-down* le corpus, l'on se propose en effet de se laisser interroger *bottom-up* par lui. Posons ici le principe naïvement en renvoyant à (Mayaffre 2010, chap. 2) pour le détail.

Le corpus comme contexte

12. Matrice du sens, le corpus l'est car, dans son entier, il informe les textes qui le composent. Principe d'architextualité ou *détermination du local* (dans ce cas, le texte) *par le global* (dans ce cas, le corpus) dirait (Rastier 2001, 92) ; condition d'une comparaison différentielle non hiérarchisée des textes diraient (Adam et Heidmann 2005, 102 et ss.).

13. Soulignons simplement que cette affirmation théorique et récente, que la plupart des auteurs de linguistique textuelle ou de linguistique de corpus revendiquent désormais, épouse le principe même, technique et originel, de l'analyse de données textuelles, de la lexicométrie ou de la logométrie dès les années 1960 : l'idée d'un corpus-norme ou d'un corpus-référence, l'idée d'une statistique endogène ; comme si cette méthode était idéale pour ces linguistiques, comme si la méthode avait devancé, en pareil cas, la théorie. À la statistique endogène de (Guiraud 1954), (Muller 1977) ou (Brunet 2011) a en effet répondu une « stylistique endogène » (Viprey 1997), ou une « lexicologie textuelle » endogène au texte (Valette 2008) : finalement, la linguistique de corpus n'est rien d'autre qu'une *linguistique endogène*, et son outillage par la statistique (endogène par essence) apparaît naturel aussi bien chez (Biber 1988, 1995 ; Biber, Conrad et Reppen 1998), (Habert, Nazarenko et Salem 1997) que (Malrieu et Rastier 2001).

14. Matrice du sens, le corpus entretient, dès lors, un dialogue direct avec la notion de *co(n)texte* puisque l'on admet que la co(n)textualisation est la condition de la maïeutique du sens. Le sens naît en/du co(n)texte, avons-nous plusieurs fois écrit (Mayaffre 2007b, 2014) en soulignant la vanité d'une linguistique ou simplement de pratiques méthodologiques décontextualisantes. Le sens naît en/du corpus pourrait-on renchérir ici liant corpus et contexte dans une relation étroite, quasi synonymique.

15. Le corpus peut être en effet conçu comme une forme privilégiée du co(n)texte. Plus précisément, nous définis-

sons le corpus comme *la forme maximale du co(n)texte*. De la lettre au mot, du mot à la phrase, de la phrase au paragraphe ou à la partie, de la partie au texte, du texte au corpus : le phénomène de co(n)textualisation (ou, autrement dit, l'extension de l'objet du linguiste) semble devoir aller jusque là et s'arrêter là.

16. Précisons bien cependant pour ne pas paraître naïf : en définissant le corpus comme forme maximale du co(n) texte, nous entendons forme maximale *formalisable* du co(n)texte, car le co(n)texte (le co-texte proche ou l'intertexte plus lointain) est insondable et à proprement parler *insaisissable*. Plus loin encore, le contexte lorsqu'il s'étend au-delà du co-texte ou de l'intertexte pour toucher à la situation socio-historique générale et aux conditions de production des discours est une chose qui échappe pour partie aux études strictement linguistiques ; le *hors corpus* nous mène vers un horizon scientifique insondable.

17. Pour un texte donné, donc, une infinité de corpus pourront être construits formant autant de co(n)textes maximaux formels, au sein desquels seront écrits autant de scénarios interprétatifs. Le corpus donne un corps – un corps linguistique – au contexte, et hors de la réification qu'il propose, le co-texte ou le contexte, l'intertextualité ou l'interdiscursivité sont des magmas sans forme, sans aucun doute passionnants mais platoniques ou insaisissables. Dit plus simplement, le corpus représente pour nous la forme empirique ou matérielle du contexte ; celle qui accepte de se soumettre à l'observation linguistique. C'est le cadre matériel, immédiat, formalisé, d'une interprétation contrôlée.

Corpus réflexifs et herméneutique endogène

18. Pour répondre à cette double définition – le corpus comme matrice du sens ; le corpus comme objectivation du contexte linguistique nécessaire à l'interprétation –, les corpus textuels doivent être bien formés (Landragin et Poudat 2017).

19. Outre les critères désormais bien connus d'équilibre, de représentativité ou d'exhaustivité, d'homogénéité (notamment générique), de contrastivité et de clôture, etc., nous avons proposé non seulement que les corpus soient gros pour offrir un cadre suffisant[8] mais structurés de manière adéquate.

20. La proposition qui a été la mieux reprise par la communauté scientifique est la nécessaire *réflexivité du corpus*. Sans revenir dans le détail sur une idée développée par le

8. À vrai dire, la taille des corpus est une question insoluble, inutilement polémique. Nous savons que le traitement statistique est d'autant moins contestable que les populations sont importantes, et la puissance des machines repousse chaque jour les limites de la veille. Aussi « *more data, better data* » pourrait être une devise pertinente s'il n'y avait la nécessité d'embrasser le texte qualitativement : trop gros, les corpus deviennent illisibles et ininterprétables. Puisqu'il faut donner un ordre de grandeur, posons, pour ce qui nous concerne, que nous traitons plutôt des corpus de 2 000 000 de mots que de 200 000 (trop petits pour mesurer des régularités) ou de 20 000 000 (trop gros pour être lisibles).

menu dans (Mayaffre 2002) et (Mayaffre 2007b) et déclinée dans (Mayaffre *et al.* 2020 ; Mayaffre 2020), posons dans la continuité de ce qui précède que l'enjeu du *corpus réflexif* est de constituer un ensemble sémantique auto-suffisant qui internaliserait les ressources co-textuelles nécessaires à l'interprétation de chaque texte. En miroir, les textes du corpus doivent s'éclairer mutuellement ; se *réfléchir* les uns les autres ; chacun d'entre eux constituant le co-texte immédiat de tous, et l'ensemble constituant l'intertexte de chacun.

21. Comme l'importance de la notion a déjà été soulignée précédemment, avançons ici l'idée que le *corpus réflexif* est la condition d'une herméneutique *endogène*. C'est au sein du corpus que les parcours interprétatifs sont proposés ; le corpus réflexif dans son ensemble formalisant l'intertextualité – une intertextualité parmi d'autres possibles – des textes constituants, soumis à l'analyse.

22. Internaliser les ressources interprétatives dans des corpus réflexifs, pour une herméneutique endogène : l'ambition paraîtra démesurée mais rappelons le modeste point de départ – un malaise épistémologique – de la réflexion. Une discrimination non acceptable sépare souvent les textes-objets-du-corpus et les textes-ressources-interprétatives-hors-corpus. Les premiers font l'objet d'un mode de sélection, d'une attention philologique et, par définition, d'un traitement linguistique minutieux. Les seconds sont convoqués à discrétion, cités à la hussarde et, hors du corpus, échappent au traitement proprement dit. Cette discrimination saute à l'œil car, sources ou ressources, il s'agit bien dans les deux cas de *textes*.

23. Notons que ce co-texte ou cet intertexte que les corpus réflexifs entendent manufacturer peuvent apparaître à l'usage objectif : il s'agit de traiter ensemble deux textes contemporains jugés comme apparentés historiquement – le premier constituant le co-texte immédiat objectif du second, le second le co-texte immédiat objectif du premier –, comme par exemple, dans nos travaux, les discours de Jospin chef de gouvernement et les discours de Chirac président de la République, durant la période de cohabitation (*i.e.* explicitement les deux locuteurs se répondent). Mais la réflexivité mise en scène peut être plus subjective et arbitraire, et mettre volontairement en dialogue deux textes « étrangers » l'un à l'autre, mais dont on suppose, par hypothèse, que la mise en rapport – le face-à-face *réflexif* – au sein du corpus produira des effets de sens et suscitera des parcours de lecture critiques et fertiles. Précisons par-là, avec force, combien le corpus est un objet construit sur la base d'hypothèses de travail, et combien cette construction détermine l'analyse. Avec des méthodes *corpus-driven* ou émergentistes comme la logométrie ou le *deep learning*, les hypothèses de travail ne doivent pas présider au traitement linguistique du corpus (laissons remonter librement et sans *a priori* du corpus des informations linguistiques pertinentes) : il est suffisant et impérieux que les hypothèses de travail président au recueil – *moment philologique* (Adam et Heidmann 2005, 83) – et à l'organisation réflexive – *projection herméneutique* – des textes constituant du corpus.

Le corpus comme texte

24. De la lettre au corpus, en passant par le mot, la phrase ou le paragraphe, et en s'arrêtant sur le texte, la linguistique de corpus procède à/de l'extension de l'objet de la linguistique vers des réalités ou des globalités toujours plus vastes et toujours plus complexes ; les corpus représentent pour nous le *terminus ad quem* d'une linguistique contextualisante.

25. Par facilité sans doute, à chaque palier de complexité franchi, l'analyste a eu tendance à se retourner vers le palier inférieur pour en faire remonter des schémas d'analyse qui lui étaient familiers. Ainsi, hier, par exemple, lorsqu'on est passé du phrastique au transphrastique, s'est-on imaginé établir – en vain – une *grammaire du texte* comme il en existait une précédemment de la phrase.

26. Ainsi, aujourd'hui, en passant du texte au corpus peut-on envisager d'expliquer – avec fruit ? – la « corporalité » comme on explique la textualité ; et précisons que la tâche s'annonce passionnante mais d'autant plus compliquée que la notion de textualité est elle-même à peine stabilisée en linguistique textuelle et objet encore de riches discussions.

27. Peut-on considérer un corpus comme un texte ? Peut-on considérer un corpus textuel comme un macro-texte qu'il s'agirait alors de traiter avec des outils théoriques en partie balisés par Hjelmslev ou Bakhtine, Hasan, Halliday, Adam ou Rastier ?

28. Nos travaux se gardent bien d'apporter une réponse définitive à une interrogation qui engage sinon l'avenir de la linguistique de corpus en tout cas son intérêt actuel. Mais plusieurs indices, comme autant de pistes de réflexion, peuvent être pressentis. Deux méritent d'être rappelés ; nous verrons qu'ils sont nuancés ; et que dans ces nuances résident des programmes de recherche.

Cohérence-cohésion du texte – cohérence-cohésion du corpus

29. Si l'on peut supposer que le corpus, à l'image du texte, est un ensemble cohérent et cohésif, il l'est nécessairement de manière différente. En s'aventurant dans le *distinguo* heideggerien sans doute peut-on prétendre que la cohésion-cohérence d'un texte lui est ontologique ; la cohésion-cohérence du corpus lui est ontique. Le texte est cohérent par nature, par essence, par définition[9] ; le corpus, lui, l'est par existence – en tant qu'*étant* –, par construction, par hypothèse.

30. Il ne s'agit pas ici de naturaliser (ontologiser) l'objet texte qui est lui-même un objet construit, artefactuel, mais de rappeler que le texte existe – sous une forme ou une autre – dans la société sans l'analyse scientifique, là où le corpus existe uniquement en laboratoire par le fait du seul

9. Les « propos incohérents » qu'essayent de tenir certains auteurs n'en peuvent mais c'est précisément cette incohérence étudiée du propos qui fait la cohérence du texte.

chercheur, et seulement le temps de la recherche[10]. Le texte est un construit, mais un construit social ou culturel « de première main » par le fait du couple auteur-lecteur. La construction du corpus textuel est, elle, de « seconde main » par le seul fait de l'analyste.

31. La textualité – ce qui fait qu'un texte est un texte – est définitoire du texte et peut être perçue par tout lecteur, *sans quoi il n'admettrait pas qu'il s'agit là d'un texte*. La corporalité, elle, est une pétition de principe ou un parti pris, un espoir, un postulat, le fruit d'un travail singulier ou d'une projection particulière évidente pour le seul chercheur. Exprimée en langage mathématique, la cohésion-cohérence d'un texte est axiomatique ; la cohésion-cohérence du corpus est hypothétique. Bref, un texte qui ne serait pas cohérent-cohésif ne serait plus un texte. Un corpus qui n'est pas cohérent-cohésif est seulement un corpus manqué, c'est-à-dire manquant de pertinence et d'efficacité heuristique ; mais cela reste un corpus[11].

32. La différence est donc importante. Pourtant, elle n'est pas définitive. Quoique d'une autre nature, la cohérence-cohésion du corpus est l'enjeu de la linguistique de corpus exactement comme la cohérence-cohésion du texte est celui de la linguistique textuelle : c'est cette tension commune vers une textualité/corporalité, conçue avec (Charolles 1995, 10) comme « *principe général* gouvernant l'interprétation[12] » du texte/corpus, qui rapproche les deux disciplines. Simplement, si du point de vue de la cohérence-cohésion du texte, de (Halliday et Hasan 1976) à (Calas 2006) ou (Adam 2008), l'essentiel est déjà réalisé, du point de vue du corpus, l'essentiel reste à faire, même si les travaux de (Viprey 1997, 2006, 2005a) sur la micro/macro-distribution des unités dans le corpus et la *texture* des corpus balisent une partie du terrain. Et à ce stade, pressentons seulement, d'un point de vue méthodologique, que le parallèle texte/corpus et textualité/corporalité demande quelques ajustements : si le texte et la textualité peuvent encore être considérés comme des objets *micro* réclamant l'approche qualitative, le corpus et la corporalité, en tant qu'objets *macro*, semblent exiger une approche quantitative.

Sérialité du corpus – linéarité du texte

33. Si texte et corpus (textuel) présentent certaines similarités au point que l'on peut envisager entre eux un simple rapport d'échelle, une différence profonde de structure semble les distinguer : le corpus est fondamentalement

10. Rappelons les échecs répétés d'archives de corpus ou de banques de corpus. Tous les chercheurs ont rêvé de sauvegarder et patrimonialiser leurs corpus (et non seulement les textes) mais force est de constater que ces tentatives sont le plus souvent vaines. Les corpus semblent destinés à disparaître après l'analyse et la validation de l'étude par des jurys. Certes, des *benchmark corpus* existent pour comparer des méthodes et hiérarchiser des logiciels mais il ne s'agit pas de corpus SHS en vue d'une analyse.

11. Le lecteur aura remarqué le parti pris de mentionner ensemble, globalement, la *cohérence* et la *cohésion*. Dans le détail, et de manière hiérarchique, il serait facile de montrer que la *cohésion* du corpus pose plus de problèmes encore que sa *cohérence*. Cf. *infra* la question de la sérialité des corpus (*versus* la continuité des textes).

12. Surligné par Michel Charolles. Voir aussi son article moins abouti de 1983 : Charolles, Michel. 1983. « Coherence as a Principle in the Interpretation of Discourse ». *Text 3* (1). https://doi.org/10.1515/text.1.1983.3.1.71.

un objet *sériel* (Mayaffre 2002), le texte est d'abord un objet *linéaire*.

34. Le corpus est une *collection* de textes réunis sur la base d'hypothèses de travail. Au-delà du stade critique d'une collection de textes qui en compterait un seul, les corpus peuvent donc être considérés comme des séries. (Et faut-il souligner encore ici que les séries, en linguistique comme ailleurs, se prêtent bien au traitement statistique ?)

35. Certes, certaines de nos séries, particulièrement en histoire, sont ordonnées linéairement. Nous pensons aux *séries textuelles chronologiques* dont André Salem a décrit les caractéristiques (Habert, Nazarenko et Salem 1997, 207 et ss.) et qui, précisément, par leur *progression* chronologique, et la permanence de leurs locuteurs individuels ou collectifs, peuvent à juste droit être traitées comme des textes : il s'agit-là d'un champ de recherche à part entière de la linguistique de corpus dont (Viprey 2004) ou (Metwally 2017), sur les numéros du *Monde diplomatique*, échelonnés sur plusieurs décennies, ont décrit le fonctionnement.

36. Mais hors des séries textuelles chronologiques, la plupart des corpus n'ont pas de structure linéaire évidente ; de manière significative leurs parties (les textes qui les composent ou des regroupements de textes que l'on aura constitués en parties) peuvent être indifféremment ordonnées sans que le traitement en soit changé. Ainsi dans le corpus de la campagne électorale de 2007 ou de 2017 que nous avons eu l'occasion de traiter, les textes des candidats Laguiller, Buffet, Royal, Bayrou, Sarkozy et Le Pen ou Mélenchon, Hamon, Macron, Fillon, Le Pen peuvent contraster et se singulariser indépendamment de leur ordre de saisie (Mayaffre *et al.* 2017 ; Mayaffre 2020)[13].

37. Si le corpus est avant tout sériel donc, et non nécessairement organisé linéairement, le texte lui est toujours linéaire ; c'est un objet linéaire. À l'exception de quelques productions surréalistes ou de quelques jeux d'auteurs marginaux en littérature[14], un texte a toujours un commencement, un prolongement et une fin ; il peut être défini comme une *suite* (Maingueneau 1996, 81 ; Détrie, Siblot et Verine 2001, 349) ; et l'élément fondamental de sa lecture est sa progression, pour nous de gauche à droite, de haut en bas. Contrairement aux parties du corpus, les parties du texte (ses phrases, ses chapitres, ses séquences…) ne peuvent être inversées sans remettre en cause l'édifice. Certes, depuis l'abandon du rouleau pour le codex ou le *polyptychon*, rien n'interdit au lecteur de briser par sa lecture cette progression implacable et de papillonner aléatoirement d'une page à l'autre, d'arrière en avant, ou de chapitre en chapitre. Certes encore, aujourd'hui, le numérique permet des lectures hypertextuelles dont la caractéristique est justement de s'affran-

13. Cette idée pourrait être nuancée (mais non contredite). On aura en effet noté qu'un *ordre* politique ici s'est imposé à nous. Et lors de l'analyse, certaines distributions semblent renvoyer à cet ordre ou cette *progression* du corpus. Ainsi par exemple constatera-t-on une progression de l'emploi de « patrie » (Mayaffre 2008c, 63 fig. 1) à mesure que le corpus se *déroule* (au sens politique *et* typographique) vers la droite.

14. Précisément il s'agit là de jeux, dont la règle sous-jacente est bien la linéarité attendue… et transgressée.

chir du linéaire : il s'agit d'une révolution majeure qui est au cœur même des propositions méthodologiques du traitement informatique et statistique des textes et sur lesquelles nous reviendrons. Mais il n'en reste pas moins vrai que la linéarité apparaît irréductible à la textualité et constitue le socle de sa définition[15].

38. Suite continue *versus* série discontinue, linéarité du texte *versus* sérialité du corpus : touchons-nous donc cette fois-ci à une différence définitive ? Non pourtant.

39. Objet linéaire, *d'abord*, le texte est aussi traversé de sérialité et de réticularité : c'est l'apport essentiel des travaux de (Viprey 1997, 2006, 2005a) que nous avons essayé de reprendre d'un point de vue théorique et pratique dans l'ensemble de nos écrits (pour une approche globale : Mayaffre 2010). Objet sériel, *en premier*, le corpus est – ne serait-ce que parce qu'il est composé de textes linéaires – traversé par la linéarité et la séquentialité : c'est l'apport essentiel par exemple des travaux de (Mellet et Longrée 2009) ou (Longrée et Mellet 2013).

40. Autrement dit se dessine un double mouvement scientifique qui venant de deux pôles opposés converge en un programme de recherche commun : la linguistique de corpus, partant de la série, vise aujourd'hui à réintroduire la linéarité dans ses traitements. La linguistique textuelle, partant du linéaire, admet la sérialité comme élément complémentaire de son objet. À la suite de Jean-Michel Adam qui, venant du texte, a présenté l'unité de ce programme aux JADT 2006 (Adam 2006) avant d'en esquisser des pistes dans (Adam 2008, 179-182), nous avons essayé, venant du corpus, d'en souligner quelques enjeux (Mayaffre 2007a) ou (Mayaffre 2014).

41. Dit en des termes admis depuis longtemps en linguistique, il s'agit de rappeler que toute écriture/lecture (celle d'un texte ou celle d'un corpus ; a fortiori celle d'un *corpus textuel*) articule une compétence syntagmatique et une compétence paradigmatique. Longtemps essentiellement paradigmatique, l'approche statistique des données textuelles doit prendre en compte désormais aussi la dimension syntagmatique, séquentielle et plus généralement encore co(n)textuelle de son objet. Dans ce cadre, nos principaux travaux insistent classiquement sur le traitement des co-occurrences (Mayaffre 2008c, 2008b, 2008a ; Mayaffre et Viprey 2012 ; Mayaffre 2014) car la co-occurrence articule, dans son essence même, un processus sélectif et un processus combinatoire[16]. Plus précisément, le calcul de la co-occurrence a pu y être conceptualisé comme le premier mouvement d'une statistique lexicale contextualisante, faisant passer nos pratiques, écrivions-nous, d'une lexicographie passive (le relevé d'occurrences) à une lexicologie active sémantiquement.

15. Citons ici la concession définitive du linguiste qui a remis le mieux en cause cette linéarité pour introduire dans le traitement la réticularité : « Nul ne saurait mettre en doute qu'un texte se manifeste dans l'ordre du temps et/ou de l'espace orientés, se caractérise par un début, un milieu, une fin, ordonnés et non interchangeables, et ce à quelque échelle que ce soit de l'organisation macro-séquentielle à la fine succession des périodes » (Viprey 2006, 74).

16. Concrètement, il est possible de montrer que le calcul des co-occurrences mobilise/produit une liste paradigmatique (le mot pôle et les mots co-présents) et une fenêtre syntagmatique ou co(n)textuelle dans laquelle ces mots se combinent.

Dans la perspective herméneutique qui est la nôtre, nous avons en effet pu définir la co-occurrence comme *la forme minimale du contexte* (forme minimale calculable) nécessaire à l'interprétation. En une phrase simple mais définitive : constater que le mot a et le mot b sont co-occurrents, c'est contextualiser minimalement l'un par l'autre[17]. Ainsi pouvons-nous prétendre tenir les deux extrémités du *contexte* c'est-à-dire de la ressource interprétative : à un pôle, le corpus réflexif (forme maximale formalisable du contexte), à l'autre pôle la co-occurrence (forme minimale calculable du contexte).

Corpus et numérique

42. À vrai dire, assimiler le corpus à un macro-texte à traiter est moins, aujourd'hui, une option théorique ou une vue de l'esprit qu'une possibilité pratique offerte par le numérique. De manière générale, l'ensemble de notre réflexion sur les corpus retracée dans cette contribution se trouve déterminée par le phénomène numérique. Particulièrement, la notion de *corpus réflexif* ne serait qu'une vaine spéculation si le numérique ne rendait pas possible la structuration architextuelle du corpus et une navigation/exploitation hypertextuelle généralisée : ici c'est bien le numérique et lui seul qui organise le dialogue souhaité entre textes, et objective ainsi l'intertextualité.

17. Après d'innombrables travaux, la pertinence de l'approche n'a plus à être démontrée. En un seul exemple grossier : calculer que la cooccurrence de « classe » est « ouvrière » dans le texte A et « tableau » dans le texte B permet de désambiguïser sémantiquement (et idéologiquement) le mot « classe » et les textes A et B.

43. Après d'autres, nous avons donc écrit que le passage du papier au numérique ne représente pas un simple changement technique de support du vecteur principal de la culture humaine (le texte) mais une révolution culturelle et épistémologique (anthropologique aussi sans doute), sans guère de précédent dans l'histoire ; sans doute supérieur à la révolution Gutenberg de la Renaissance (Mayaffre 2007b). De MacLuhan à la médiologie de Debray, et aujourd'hui au développement des humanités numériques, tout indique que le média numérique informera fondamentalement la forme, le fond, la signification ou sens des textes à venir.

44. Pour le simple propos qui nous intéresse ici – le statut des corpus textuels –, cette révolution peut être résumée par une formule paradoxale qui retourne l'état de l'art traditionnel : le numérique *dématérialise le texte* et *matérialise le corpus*.

45. Là où l'on avait tendance à naturaliser le texte en l'assimilant, dans sa fixité matérielle, à son support physique traditionnel (la page et le livre), la philologie numérique rend évidente son artefactualité. Pluralité des formats et des codages, choix multiples et individuels dans l'affichage, multiplication des niveaux d'étiquetage, d'annotations, d'enrichissement, circulation, sans limite et jusqu'à l'ubiquité, par fichier joint, parcours de lecture variés, etc. : tout se combine pour souligner aujourd'hui la volatilité ou la relativité du texte, sa dimension artefactuelle, conventionnelle ou culturelle (*i.e.* non naturelle). En l'arrachant du scriptorium et de la bibliothèque,

en le « défixant » de l'ouvrage papier qui jusqu'ici le supportait, le numérique a définitivement dénaturalisé et dématérialisé le texte, retrouvant ainsi une pratique ancienne de l'Antiquité jusqu'au Moyen Âge[18]. Dans les mots de Dominique Legallois :

46. Bien sûr, il faut mentionner l'hypertextualité liée à la mutation numérique du texte, qui oblige à une reconsidération de l'unité textuelle et du texte lui-même : en tant qu'ensemble de possibilités de parcours, le texte devient alors *une unité virtuelle* (souligné par nous. Legallois 2006, 7)

47. Inversement, là où l'on considérait le corpus seulement comme une idée ou une virtualité, le numérique le matérialise, l'incarne, le réifie en le rendant, quelle que soit sa longueur, palpable et manipulable, exploitable et ré-exploitable, archivable et échangeable[19]. Si le terme de corpus avait tendu à disparaître dans les années 1980, avant de s'imposer aujourd'hui, c'est qu'il était sans grande pertinence et sans contrainte ; flasque jusqu'ici, il est devenu désormais un concept dur. Le corpus était en effet un idéal (« potentiellement tous les textes susceptibles de m'intéresser ») : c'est aujourd'hui un matériau (« réellement, seuls les textes que j'ai *saisis* en machine et que je peux matériellement soumettre au traitement »). Hier encore horizon, il est devenu aujourd'hui, à la faveur du numérique, un continent, dont la clôture constitue une limite mais sur lequel il est désormais possible de circuler. En ce sens, rappelons que le développement de la linguistique de corpus (c'est-à-dire le dépassement du texte seul par le corpus, considéré alors comme le macro-objet de la linguistique) est un produit de la révolution numérique. Du Brown corpus au British national corpus en passant par le *Trésor de la langue française*, des corpus lemmatisés du LASLA aux corpus XML* de la Base de français médiéval en passant par le Nouveau corpus d'Amsterdam – sans rien dire des corpus particuliers des chercheurs –, tous les linguistes de corpus travaillent aujourd'hui sur corpus numériques, et tous réfléchissent à des méthodes numériques pour traiter leur objet numérique.

48. Récemment, le traitement des corpus numériques par l'intelligence artificielle et le *deep learning*, nous a permis de montrer comment des notions aussi fuyantes que l'interdiscours ou l'intertexte pouvaient être, pour la première fois, formalisées grâce au numérique. Après apprentissage (*machine learning**), la machine classe et décrit le parler de Macron, de de Gaulle, de Pompidou ou de Hollande, et retrouve automatiquement les observables linguistiques qui traversent le corpus présidentiel sous la V^e^ République : l'intertexte commun que les présidents partagent (Mayaffre *et al.* 2020 ; Mayaffre et Vanni 2021).

18. Cf. la réflexion engagée sur la philologie numérique par divers auteurs déjà cités : (Rastier 2001, 73-97 chap. 3 « Philologie numérique » ; Viprey 2005b).

19. Nous avons souligné *supra* la difficulté d'archiver sur le long terme les corpus d'étude, mais le temps de la recherche le corpus numérique peut être stocké, mobilisé et re-mobilisé, échangé, confronté à d'autres, etc.

Conclusion

49. Dans un article récent, dans lequel l'auteur fait référence au sein d'une riche bibliographie internationale aux travaux de François Rastier ou de la revue *Corpus* et, plus modestement, aux nôtres (Laks 2008) rappelle l'opposition fondamentale entre une linguistique du *datum* et une linguistique de l'*exemplum*. Magistralement, Bernard Laks démontre que les corpus existent depuis toujours en philologie et en linguistique, et qu'il ne saurait y avoir de modélisation du langage sans prise en compte des usages c'est-à-dire de données attestées recueillies dans de vastes compendiums.

50. Pour pertinente qu'elle soit dans l'histoire des sciences du langage, l'opposition retracée perd de son efficacité aujourd'hui, en opposant tout le monde à personne et en plaidant une cause entendue désormais par tous ; depuis la fin du XX^e^ siècle, il n'y a plus nécessité de dénoncer la grammaire générative et ses *exempla* controuvés, comme le temps des luttes politiques contre le rideau de fer apparaît révolu. Aujourd'hui, ce n'est plus le corpus qui sépare en SHS puisqu'il est admis unanimement. C'est son statut heuristique ou épistémologique qui fait toujours clivage.

51. Pour Laks, un corpus rassemblerait des données qu'il conviendrait ensuite d'extraire, de décrire, de modéliser. Mais la notion de « données », elle-même, nous semble dangereuse. Les données sont seulement *ce que l'on se donne* ironisent (Malrieu et Rastier 2001, 554 note) ; elles ne peuvent que difficilement prétendre être le contenu objectif de la langue ou un échantillon représentatif de l'infini du langage. Dangereuse par son illusion objectivante, la notion de « données » semble surtout inopérante pour désigner un *texte* dans son épaisseur et sa complexité, dans son expression que la philologie toujours discute, dans son organisation et sa productivité sémantiques qui restent à découvrir.

52. Un corpus textuel – puisque tels sont nos corpus – ne donne jamais accès à aucun contenu objectif : il problématise seulement la lecture et organise l'interprétation. Il ne restitue point un sens positif : il le produit. Pas plus que nous pouvons envisager de travailler sans corpus, nous ne pouvons concevoir le corpus seulement comme une chambre froide de dissection du sens ou comme un cercueil de données.

53. Le corpus est un ensemble dynamique et contrastif pour une sémantique différentielle. C'est un tout vivant, clos sur lui-même mais réflexif pour une herméneutique endogène. C'est une composition matérielle car saisie, contraignante par sa clôture, exigeante par sa réflexivité, mais que le numérique aujourd'hui fertilise et anime.

54. Il n'y a de sens que d'interprétation : les corpus textuels numériques bien formés sont pour nous le lieu effectif de cette interprétation, et la condition nécessaire de son contrôle.

La méthode diplomatique face à l'information numérique

Marie-Anne Chabin

Introduction

1. La diplomatique est une discipline universitaire ancienne et pourtant largement méconnue, injustement méconnue. La définition classique, celle que l'on trouve par exemple dans l'*Enclyclopedia Universalis* sous la plume de Robert-Henri Bautier, dit que :

2. > La diplomatique est la science qui étudie la tradition, la forme et la genèse des actes écrits en vue de faire leur critique, de juger de leur sincérité, de déterminer la qualité de leur texte, d'apprécier leur valeur exacte en les replaçant dans la filière dont ils sont issus, de dégager de la gangue des formules tous les éléments susceptibles d'être exploités par l'historien, de les dater s'ils ne le sont pas et enfin de les éditer. Science autonome, elle est aussi et avant tout une des sciences auxiliaires de l'histoire.

3. Cependant, déjà en 1961, Georges Tessier, qui venait de prendre sa retraite après trente ans de professorat de la diplomatique à l'École des chartes, écrivait déjà : « La notion de science auxiliaire appliquée à la diplomatique nous paraît aussi contestable et aussi périmée que la limitation de la méthode aux seuls documents médiévaux » (Tessier 1961, 670).

4. Il est intéressant de noter que la définition de *Wikipédia*, référence principale pour le grand public et source importante pour les chercheurs, reste traditionnelle malgré quelques mots plus modernes : « une science auxiliaire de l'histoire qui étudie la structure, la classification, la valeur, la tradition et l'authenticité des documents officiels[1] ». La définition de *diplomatics* dans le *Wikipédia* anglais est, quant à elle, plus ouverte : « *a scholarly discipline centred on the critical analysis of documents: especially, historical documents*[2] ». Une autre définition, dans un cours en ligne de l'École nationale des chartes, insiste sur la finalité juridique de l'objet étudié, sans précision de période : la diplomatique est l'étude de « tout écrit utilisé ou utilisable comme titre, fondamentalement pour prouver un droit » (Guyotjeannin 2011).

5. On voit là les tendances, voire les tentations, de la discipline. Face à la société numérique, à l'œuvre depuis un demi-siècle déjà, la diplomatique se trouve à une croisée de chemin : s'adapter et intégrer la matière numérique qui est devenue majoritaire dans la production des écrits,

1. Cf. https://fr.wikipedia.org/w/index.php?title=Diplomatique&oldid=181561319 (consulté le 18 avril 2021).
2. Cf. https://en.wikipedia.org/w/index.php?title=Diplomatics&oldid=999517788 (consulté le 18 avril 2021).

soit se cantonner aux corpus anciens, toujours riches d'enseignement mais figés. C'est la perspective de cette évolution vers une « diplomatique numérique » qui est analysée ici.

De l'acte écrit à la trace numérique

6. L'objet de la diplomatique est d'abord le diplôme, qui donne son nom à la discipline. Le terme diplôme désigne à l'origine un document officiel, émanant d'un souverain et établissant un droit ou un privilège. L'étymologie, avec le préfixe « di- », rappelle que ces actes étaient pliés en deux.

7. L'acte de naissance de la discipline est la publication en 1681 par Jean Mabillon, sous le titre *De re diplomatica* (un ouvrage de plus de 1000 pages), de la méthode qu'il a construite et appliquée pour sortir d'une controverse opposant deux congrégations religieuses, Jésuites et Bénédictins, au sujet de l'authenticité des actes des rois mérovingiens conservés à l'abbaye de Saint-Denis. Des jésuites avaient entrepris d'écrire une somme sur la vie des saints mais le fait que l'un deux ait dénoncé comme faux une partie des actes royaux (autrement dit une partie des diplômes), faisait porter un fort discrédit sur l'ordre bénédictin dont les possessions territoriales devaient beaucoup aux donations des rois de France.

8. Mabillon, lui-même bénédictin mais reconnu et apprécié pour son intégrité intellectuelle, se vit confier la tâche de procéder à une étude rigoureuse des documents afin de faire apparaître la vérité. Sa méthodologie constitue les bases de la critique diplomatique. Elle est originale car personne avant Mabillon n'était allé aussi loin dans la description des concepts et du mode opératoire qui permettent de discerner le vrai du faux. Sur le plan de la méthode, Mabillon a scientifiquement constitué un corpus d'actes avant de procéder à une observation cartésienne de chaque partie des documents afin de se livrer à une critique comparée, élément par élément.

9. Ce faisant, Mabillon a élaboré ce qu'on pourrait nommer une grille-type d'analyse des documents dans le but de faire ressortir leur véracité ou leur fausseté. Tout acte, et plus largement tout écrit, comporte d'une part des caractères externes (ce que l'on voit : matière, écriture, mise en page) et des caractères internes (l'expression écrite). Pour éviter d'utiliser le vocabulaire technique de la diplomatique mabillonienne, on peut dire que cette expression écrite se divise en trois zones d'information : identification des protagonistes, texte proprement dit avec la description du contexte et la décision, et une zone de validation et de datation. De l'analyse, suivie d'une critique comparée, naît la démonstration d'authenticité.

10. Au cours des trois siècles suivants, les travaux de diplomatique ont été approfondis par plusieurs écoles de chercheurs, en France mais aussi en Italie et en Allemagne. L'étude diplomatique a été étendue aux documents de la période moderne (XVI^{e}-XVIII^{e} siècles) et à l'époque contemporaine. On distingue alors dans les études la

diplomatique générale et la diplomatique spéciale, appliquée à un type de document particulier (par exemple, les actes pontificaux ou les documents comptables).

11. La masse documentaire est en constante augmentation depuis le Moyen Âge et la forme des documents s'est diversifiée, notamment avec la production croissante de brouillons et de copies aux côtés des originaux. Ceci a donné lieu à la constitution de dossiers dès le XVII^e siècle, même si le terme « dossier » n'apparaît dans l'administration qu'au XIX^e siècle. Or, en observant l'évolution de la production documentaire des deux derniers siècles, on voit bien que le dossier n'est, dans de nombreux cas, que la ramification technique et informationnelle de la pièce unique qui faisait le même office naguère. « On constate ainsi que le discours du dossier, au-delà de la mosaïque documentaire de sa présentation physique, répond à la composition du discours diplomatique classique » (Chabin 2000).

12. Depuis quelques décennies, la recherche en diplomatique connaît un regain d'intérêt. Une des raisons en est la publication, échelonnée entre 1989 et 1992, d'un article fondateur intitulé *Diplomatics : New Uses for an Old Science*. À une époque où la bibliographie sur la diplomatique en langue anglaise est quasi inexistante, cet article de Luciana Duranti, professeur en sciences de l'information à l'université de Colombie britannique à Vancouver, porte à la connaissance des chercheurs anglophones la quintessence de la discipline diplomatique telle que l'ont formulée Mabillon et ses successeurs. Dans cet article, l'auteur évoque assez peu les supports numériques tout juste naissants mais l'idée commence à émerger ici et là que les défis de gestion dans le temps des bases de données ont peut-être à s'inspirer des méthodes de la diplomatique, le numérique étant, dans sa fragilité, possiblement plus proche du manuscrit que de l'imprimé.

13. Cette problématique est au cœur du colloque franco-américain organisé en 1992 et 1993 par la bibliothèque historique Bentley de l'université d'Ann-Arbor, Michigan et l'École nationale des chartes à Paris. « Dans un environnement où l'information est distribuée par des ordinateurs et qui est caractérisée par la libre circulation des textes, les questions d'authenticité, d'intégrité et d'édition des documents vont devenir d'une importance vitale » (Blouin 1996). C'est aussi l'occasion de rappeler que « la finalité et la forme d'un document sont des éléments essentiels en diplomatique. Ce sont aussi des moyens de caractériser des documents qui sont devenus moins importants dans le contexte des archives modernes, en raison de l'importance accordée au dossier. La finalité d'un document, pris isolément, est en général soumise au contexte du dossier (ou de la série) dans lequel il est conservé » (Blouin 1996).

14. Au tournant du siècle, l'image et l'audiovisuel s'invitent dans le débat et deviennent à leur tour objets de la diplomatique. « L'approche diplomatique est aujourd'hui une nécessité absolue, ne serait-ce que pour des raisons quantitatives : les documents audiovisuels (images fixes, images animées et enregistrements sonores) représentent

déjà pour le XXe siècle une masse d'archives équivalente à celle des archives sur papier. Le chercheur se trouve devant les documents audiovisuels dans une situation comparable à celle des historiens romantiques devant les documents écrits : il doit opérer une révolution mentale » (Delmas 2003).

15. L'idée est celle d'un examen critique des caractéristiques externes et des caractéristiques internes des images, mais plus encore une posture, un état d'esprit pour l'analyse et la comparaison méthodique :

16. > Ce que cette (éventuelle) discipline [la diplomatique des images numériques] emprunte à la diplomatique, ce ne sont pas des méthodes ou des techniques ni même des concepts, mais une culture, un esprit d'analyse et de critique orienté vers un ensemble de finalités similaires : dégager l'authenticité et la valeur des actes ou des images, retracer leur filière, comprendre et connaître leur utilisation. Il existe je crois un intérêt à « tenir ensemble », à regrouper sous une même conception analytique des méthodes et des techniques disparates mais qui toutes tendent à obtenir le maximum d'informations sur l'élaboration, les transformations et utilisations des images numériques dans le but de qualifier leur authenticité et leur valeur. (Peccatte 2009)

17. Le champ de la diplomatique s'ouvre encore un peu plus avec le développement de la messagerie électronique à partir du milieu des années 1990 et plus encore avec la diffusion du smartphone (2007) qui changent les comportements et les modes de production de l'écrit : tout le monde produit, tout le monde laisse des traces, tout le monde cherche de l'information. Dans l'environnement numérique, les gens continuent d'avoir des relations d'ordre administratif, contractuel, social, etc. Ces relations sont tracées par des outils numériques avec, de plus en plus, géolocalisation et horodatage. Lorsqu'il y a une contestation ou une action à poursuivre, la feuille de papier rectangulaire n'est plus là mais l'ensemble de données reliées entre elles constitue l'équivalent d'un document d'archives, c'est-à-dire un contenu enregistré sur un support pour attester de l'engagement d'une personne envers une autre. « En résumé, l'objet de la diplomatique numérique est un agrégat accompagné de sa trace » (Chabin 2013).

18. Ainsi, la diplomatique, longtemps cantonnée au monde des archivistes et des historiens, s'ouvre à d'autres disciplines :

19. > En environnement numérique, la collecte et la mise en mémoire des données ne repose plus uniquement sur les compétences professionnelles des archivistes. De nouveaux acteurs s'invitent dans les flux d'information : via moteurs de recherche, réseaux sociaux, applications diverses, tout un chacun peut en effet consulter mais aussi alimenter à chaque instant des stocks de plus en plus importants. Jeux, achats, échanges ou créations constituent autant de terrains où se dépose la mémoire numérique, par captation de nos traces d'usage. Ce sont ces traces dont il faut faire mémoire (Merzeau 2012).

Travaux de diplomatique numérique

20. « Peut-on parler de diplomatique numérique ? » (Chabin 2011). Au cours de la table ronde « Archives numériques et numérisées : de nouvelles disciplines en perspectives ? » qui clôturait la journée d'études « Sciences humaines et patrimoine numérique » le 25 novembre 2010 à l'INA, quatre des cinq intervenants ont spontanément évoqué la diplomatique parmi ces nouvelles disciplines ou plutôt disciplines à renouveler pour mieux connaître, gérer et exploiter ces nouvelles archives. Par ailleurs, on peut relever le titre du colloque international de diplomatique qui s'est tenu à Paris en 2013 : « *Digital Diplomatics. What is diplomatics in the digital environment?* ». Même si l'expression « diplomatique numérique » est également utilisée pour parler de l'utilisation des technologies numériques au service de la diplomatique médiévale, on peut répondre positivement à la question : oui, on peut résolument parler de diplomatique numérique.

21. Si un doute surgit quant à l'authenticité ou la fiabilité d'un objet numérique (document, fichier, mail, post, image…), la diplomatique invite à étudier non seulement la vraisemblance du discours mais aussi la forme, le contexte, la traçabilité de l'information car la falsification peut toucher aussi bien la validation de l'écrit que son contenu. Les outils d'analyse de la validité d'un objet numérique ne peuvent pas être moins sophistiqués que les outils de production car les faussaires et ceux qui les traquent ont toujours entretenu une saine émulation sur le plan des moyens. La part technologique dans l'expertise de la qualité ou de la véracité d'un original numérique est donc essentielle. Mais la science ne suffit pas. L'exercice requiert aussi de la méthode. Comment appréhender la critique d'une information ? Comment, où et par qui a-t-elle été produite ou diffusée ? Dans quelle intention ? De quand date-t-elle ? A-t-elle été modifiée, déformée depuis sa création ?

22. Cette méthodologie est au cœur des travaux du groupe de recherche InterPARES (International Research on Permanent Authentic Records in Electronic Systems) fondé en 1999 par Luciana Duranti. InterPARES a développé quatre grands projets successifs. Le projet regroupe aujourd'hui, sur tous les continents, plus de cinquante universités et organisations, nationales et multinationales, publiques et privées. La mission de la première phase était de développer des stratégies et des normes pour la conservation authentique des documents électroniques. « Pour définir les critères d'authenticité, l'équipe reprit d'abord à son compte le postulat de la diplomatique classique selon lequel, par-delà toutes leurs différences de nature, de provenance, de date, tous les documents sont assez semblables entre eux pour que l'on puisse concevoir une forme documentaire typique idéale, qui contiendrait tous les éléments constitutifs des documents » (Duranti 2003).

23. Le but d'*InterPARES 2* était de garantir que la part de mémoire de la société produite sous forme numérique dans des systèmes interactifs, dynamiques et participa-

tifs peut être créée dans une forme exacte et fiable, et conservée en une forme authentique, tant sur le court que sur le long terme, pour l'usage de ceux qui l'ont créée et plus largement de la société, en dépit de l'obsolescence technologique et de la fragilité des supports. La recherche s'est concentrée sur les documents créés dans des systèmes dynamiques, expérientiels et interactifs au cours d'activités artistiques, scientifiques et gouvernementales. Parmi le panel d'études de cas d'*InterPARES 2*, on peut citer : les travaux du théâtre Arbo Cyber (arts du spectacle, arts visuels et arts médiatiques) ; *Obsessed Again...*, une œuvre pour basson et électronique interactive écrite en 1992 par un compositeur canadien, Keith Hamel ; un atlas cyber-cartographique de l'Antarctique ; l'informatisation du livre foncier d'Alsace-Moselle ; le système de classement électronique de la Cour suprême de Singapour.

24. *InterPARES 3*, qui a débuté en septembre 2007 et s'est achevé en avril 2012, avait pour objectif la mise en œuvre des conclusions des deux premières phases du projet dans des organisations ou des unités d'archives dotées de ressources limitées. Le quatrième projet, nommé *InterPARES Trust* ou *ITrust* (2012-2019) insiste sur les questions de confiance et de fiabilité des dossiers et des données dans les environnements en ligne. Son objectif est de générer les cadres théoriques et méthodologiques permettant de développer des politiques, procédures, réglementations, normes et législations locales, nationales et internationales, afin de garantir la confiance du public sur la base de preuves de bonne gouvernance, d'une économie numérique forte et d'une mémoire numérique persistante.

25. Bien évidemment, la diplomatique n'est pas la seule discipline sollicitée par *ITrust*. Les chercheurs sont également des experts en archivistique, gestion des documents, droit, technologies de l'information, communication et médias, journalisme, commerce électronique, informatique de santé, cybersécurité, gouvernance et assurance de l'information, criminalistique numérique, ingénierie informatique et politique de l'information.

26. La diplomatique numérique en tant que telle poursuit cependant ses objectifs spécifiques, notamment celui de décrire et qualifier les objets d'information pertinents pour une analyse d'authenticité. C'est notamment l'enjeu de la complétude de l'objet d'information que l'on veut analyser. Peut-on valablement se prononcer sur la confiance à accorder à un objet d'information sans prendre en compte le fait de savoir si cet objet est entier, complet, intègre, ou bien s'il s'agit d'un morceau de document, d'un extrait de fichier, d'une information tronquée qui n'est pas identifiée comme telle, voire d'un montage composite qui déforme le message de l'une ou l'autre composante à l'insu des différents auteurs concernés ?

27. Pour la critique de l'information, dans sa dimension de preuve ou de connaissance, pour l'appréciation du degré de confiance que l'utilisateur peut accorder à l'écrit et à l'image, mais aussi pour attirer l'attention de l'utilisateur sur la complétude d'un objet d'information, pour éviter

une troncature ou une abstraction du contexte, la diplomatique est pertinente, dans l'environnement numérique comme dans l'environnement papier.

28. La matière pour une diplomatique numérique est immense. Pour développer la recherche, on peut suggérer au moins deux domaines d'investigation :

1. Un travail de modernisation des concepts et du vocabulaire de la diplomatique appliquée aux objets numériques, dans le respect de la discipline, en faisant ressortir la valeur ajoutée d'attribuer un contour spatial et temporel à un groupe de données avant de décrire son contexte et de critiquer sa portée ; l'objectif est de forger des outils pour qualifier plus concrètement les éventuels éléments de désinformation
2. La constitution d'une typologie d'écrits numériques représentatifs des nouvelles formes de l'information engageante susceptible d'être « convoquée » dans un différend social, une investigation judiciaire, une recherche de vérité ; ce sont donc tous les cas de forgerie et d'usurpation d'identité, de falsification, de troncature et de décontextualisation, de déformation du message, etc. mais aussi toutes les typologies d'écrits existantes, officielles ou socialement établies, justement pour pouvoir évaluer l'écart entre la norme et les pratiques. Cette typologie est indissociable de la création de corpus illustratifs dont l'intérêt est tout autant de disposer d'un référentiel de recherche que de proposer un matériel pédagogique

29. Dans ce contexte, la diplomatique n'est plus une étude rétrospective attachée à une branche de l'histoire. Elle s'affirme comme une démarche prospective tournée vers un plus grand nombre de disciplines, dont le droit et la sociologie.

Quelques exemples de recherche en diplomatique numérique

30. Pour illustrer les promesses de la diplomatique numérique, cette dernière partie de l'article présente trois exemples d'application de la critique diplomatique à des objets d'information nativement numériques ou diffusés via Internet et les réseaux sociaux.

31. Le premier cas est celui d'une forgerie, c'est-à-dire d'un document fabriqué de toutes pièces en usurpant l'identité d'un tiers. Il s'agit d'une série de faux courriers, qualifiés d'arnaques au RGPD (règlement général pour la protection des données personnelles), diffusés sous forme de fax ou de fichiers PDF en pièce jointe de mail, adressés à des centaines d'entreprises françaises à partir de juin 2018, date d'entrée en vigueur du règlement. Le texte de ces courriers insiste précisément sur le risque de forte sanction en cas de manquement avéré au RGPD. Ces courriers provenaient de plusieurs pseudo-sociétés de conseil voire d'escrocs qui, misant sur la peur du gendarme de la part des organisations qui auraient négligé de se mettre en conformité avec les exigences du règle-

ment européen, espéraient soutirer de l'argent à quelques clients naïfs.

32. L'exercice a consisté à analyser un de ces courriers en découpant le document selon la grille de Mabillon exposée rapidement ci-dessus. Les indices de fausseté sont circonscrits méthodiquement (figure 1) : le logo (un montage sans lien avec l'Union européenne) ; la formulation aberrante du verbe d'action au centre de l'écrit, sous forme d'un impératif (« Régularisez-vous ») ; les pseudo-références ou des absences de références ; le code-barre fantaisiste ; les incohérences de police d'écriture ; les erreurs de terminologie, de syntaxe et d'orthographe (Chabin 2018b).

33. C'est un cas de diplomatique traditionnelle portant sur un document A4 ; l'intérêt principal de l'analyse est de mettre en évidence que les arguments diplomatiques démontrant la fausseté du courrier sont différents et complémentaires des critères avancés par les juristes qui se sont livrés au même exercice en s'appuyant sur leurs propres discipline et expertise pour analyser le document et ses formulations.

34. Le second cas concerne la notion d'image à la une dans les médias d'information et sur les réseaux sociaux. Son absence étant dommageable à la visibilité d'un post, l'image à la une est devenue une quasi-obligation mais le choix est parfois difficile, entre images-prétexte, droits à l'image et difficultés d'illustrer certains sujets abstraits. Le décalage flagrant entre l'illustration

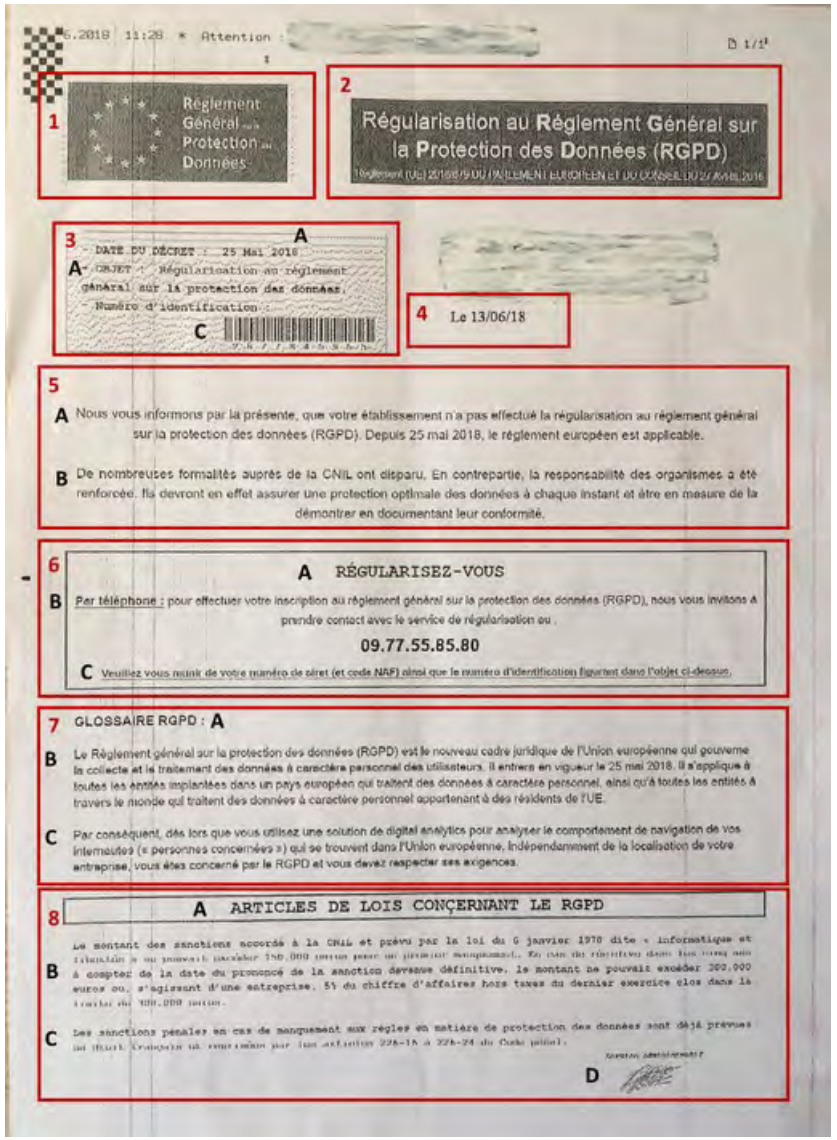


Figure 1. Analyse diplomatique d'un faux courrier diffusé par fax.
Crédit : Marie-Anne Chabin.

de certains articles partagés sur les réseaux sociaux et le titre du sujet est aujourd'hui courant, l'image étant incontournable, quelle qu'elle soit. Voici deux exemples : un article de 2018 sur la manie du président Trump de déchirer les documents officiels de la Maison-Blanche, illustré par une image montrant les morceaux d'une page de livre en espagnol soigneusement déchirée ; le second est l'annonce d'une décision de la Cour de justice de l'Union européenne de novembre 2018 concernant un fromage à tartiner néerlandais, illustrée par une cagette de crottins de Chavignol... En l'espèce, la désinformation est sans grande conséquence. Mais le même procédé pourrait se révéler désastreux dans un autre contexte (politique, santé publique...). D'où l'expression de désinformation subliminale pour qualifier cette mauvaise pratique (Chabin 2018a).

35. Cette analyse d'image à la une a également mis en évidence que le fait de partager sur un réseau social un article de presse en ligne accentue la désinformation dans la mesure où l'ordre « titre-image » du média est inversé sur le réseau social et devient « image-titre ». En effet, sur son mur, l'internaute voit d'abord l'image (sans sa légende initiale qui passe la trappe) puis, en dessous, le titre de l'article, disposition qui ne peut que l'inciter davantage à associer ce qu'il voit à ce qu'il lit.

36. L'objectif de la démarche n'est pas tant de juger le procédé que de le qualifier, de donner un nom explicite aux différentes parties de l'objet numérique que l'internaute a sous les yeux afin qu'il puisse savoir dans quelle mesure il peut se fier à ce qu'il voit lorsque cette information a été produite en plusieurs temps et provient de plusieurs auteurs. De ce point de vue, il serait intéressant de pouvoir distinguer visuellement l'image choisie par l'émetteur de l'information rapportée (un ex-fonctionnaire américain dans le premier cas, la Cour de justice dans le second) d'une image prétexte sans lien organique avec le contenu de l'article.

37. Le troisième cas de recherche est plus complexe car il ne s'agit plus de la critique diplomatique classique de l'authenticité d'un objet d'information mais de l'analyse de la composition et de la complétude d'un immense jeu de données formant un tout archivistique, en l'occurrence les données issues du Grand débat national lancé par le président de la République en janvier 2019 dans le contexte du mouvement social des Gilets jaunes.

38. L'étude porte sur le matériau du Grand débat national et sur les synthèses officielles qui en ont été tirées. Ce matériau regroupe les cahiers citoyens, des courriers, les comptes rendus de plusieurs catégories de réunions et les données de la plateforme où les citoyens étaient invités à s'exprimer, notamment au travers de questionnaires en ligne sur quatre thèmes : « Transition écologique », « Fiscalité et dépenses publiques », « Démocratie et citoyenneté » et « Organisation de l'État et des services publics ».

39. Les synthèses officielles, produites par un prestataire en communication (OpinionWay) et ses sous-traitants, ont

été diffusées mi-avril 2019. Le discours officiel était que cette quantité de données ne pouvait être traitée que par l'intelligence artificielle. La question de l'efficacité des algorithmes n'est pas en cause ici (même si le débat entre approche sémantique et approche cognitive est pertinent). La problématique soulevée est la complétude du corpus soumis à l'algorithme en termes de qualité des données. La majorité des documents ont été traités (les retardataires n'ont pas pu être pris en compte faute de temps au regard de l'impératif politique de mi-avril), mais toutes les informations enregistrées et présentes dans les divers types de contributions n'ont pas été exploitées. Le corpus analysé est quasi exclusivement textuel, à l'exclusion d'autres informations.

40. Or, l'analyse d'un simple échantillon de quelques milliers de contributions met en évidence la richesse d'informations véhiculées par le contexte, la forme, le style, etc. En effet, qu'il s'agisse des cahiers ou des comptes rendus de réunion, des questionnaires papier ou des questionnaires en ligne, la préparation des données soumises aux algorithmes a consisté à extraire les phrases, dénommées « réponses », « verbatim » ou « idées » pour faire un corpus de textes homogène. En dehors de quelques statistiques globales sur les dates et les lieux (sans corrélations avec les contenus), les informations que représentent l'agencement des contributions, les titres donnés aux contributions par les contributeurs eux-mêmes, les dates et les lieux, la mise en page, l'écriture et le style, les répétitions, les silences, etc. n'ont pas été exploités ou très peu (Chabin 2020).

41. Il est important de remarquer que la démarche diplomatique, dans ce cas du matériau du Grand débat, n'a pas pour objet de déterminer si les contributions sont authentiques ou pas ; le mode de collecte (mairie, plateforme en ligne) fait que la question n'est pas là. Initialement, les documents recueillis par la mission du Grand débat sont bien ce pourquoi ils se donnent ; en revanche, on constate un écart non négligeable entre la masse d'information créée et les données exploitées. C'est un peu comme si le destinataire du message (le gouvernement) n'en avait reçu que la moitié, l'autre moitié s'étant perdue en chemin… La réduction de l'ensemble des contributions, dans leur diversité d'expression, à une grande base d'énoncés textuels appauvrit l'exploitation de l'information. Il s'agit aussi de mesurer les écarts entre les éléments de forme de l'écrit papier et les éléments de forme de l'écrit numérique, et d'évaluer les conséquences de ces écarts. Il est à parier que la prise en compte du contexte formel de l'expression des idées aurait enrichi notablement le corpus des données analysées et, partant, les résultats de la consultation démocratique.

42. Cette opération du Grand débat, vraisemblablement représentative de nouvelles pratiques de traitement de l'information pour les décennies à venir, interpelle la diplomatique numérique, avec d'autres disciplines, sur les enjeux de préparation des données et de prise en compte de l'expression des contributeurs dans son intégralité.

Conclusion

43. Cet article pourrait s'intituler « Manifeste pour une diplomatique numérique » dans la mesure où il expose plus d'arguments et d'illustrations du possible que les résultats consolidés d'un projet de recherche académique. La multiplication des cas de fausseté, de décontextualisation, d'appauvrissement de l'expression par sa réduction au seul contenu exprimé en mots, doivent inciter les chercheurs en sciences de l'information à réagir et à s'investir davantage dans l'arène numérique.

44. Il faut espérer, d'une part, que les enjeux liés aux mauvais traitements infligés à l'information convaincront plus de diplomatistes à se tourner vers le numérique, et d'autre part, que des chercheurs d'autres disciplines s'intéresseront davantage à la forme de l'écrit (quel que soit le nom qu'ils lui donnent). La question du vrai et du faux ne peut progresser que dans une approche pluridisciplinaire et internationale.

Le goût de l'archive à l'ère numérique : gestes et récits historiens, du document au corpus

Caroline Muller et Frédéric Clavert

Introduction

1. En novembre 2016, Caroline Muller, alors professeur agrégée en histoire contemporaine à l'université de Reims Champagne-Ardennes et co-autrice de cet article, s'interroge sur Twitter : « peut-être [faudrait-il] écrire l'équivalent du "*Goût de l'archive*" numérique, tant les corpus se multiplient et les gestes se créent » (figure 1). Frédéric Clavert, alors maître-assistant en histoire contemporaine à l'université de Lausanne et co-auteur de cet article, fait partie de ceux qui ont répondu et crée un pad, c'est-à-dire un document d'écriture collaborative en ligne : une vingtaine d'historiens et d'historiennes y collaborent, imaginant comment l'on pourrait écrire un livre sur le goût de l'archive à l'ère numérique.

Figure 1. Tweet de Caroline Muller daté du 27/11/2016

Crédit : Caroline Muller et Frédéric Clavert

2. Le tweet fait référence à un ouvrage qui a fait date dans l'historiographie non seulement française, mais également anglo-saxonne : traduit en anglais, portugais, allemand, *Le Goût de l'archive* d'Arlette Farge est publié en 1989. Ce « petit » livre, comme l'autrice nous l'a décrit dans un échange de mails, décrit la relation d'Arlette Farge à l'archive dans ses dimensions pratiques – y compris le choix d'une place dans la salle de lecture –, émotionnelles – et ce jusque dans la relation avec les personnes, des femmes, mentionnées dans les archives d'Arlette Farge –, et matérielles – la manipulation de l'objet, sa couleur, son aspect, qui selon Arlette Farge influencent la manière d'appréhender l'archive donc de l'interpréter. Ce livre est aujourd'hui, souvent, considéré comme une description du parangon du travail historien, image quasi indépassable du chercheur ou de la chercheuse en centre d'archives, au corps défendant de l'autrice elle-même.

3. L'interrogation qui marque le tweet de Caroline Muller est au cœur du projet qui en a découlé : la relation de l'historien ou de l'historienne à ses archives, à son corpus, a évolué à notre ère numérique. Si nous sommes partis d'une relecture d'Arlette Farge, notre objet porte désormais sur les routines numériques « discrètes » des historiens face à l'archive, y compris au moment de la constitution du corpus et notre problématique est ainsi de nous pencher sur les conséquences possibles pour l'écriture de l'histoire de cette introduction discrète, tant logicielle que matérielle, de l'informatique, dans un sens large, dans nos pratiques. Nous allons dans ce chapitre revenir en premier lieu sur l'expérience de l'écriture de notre livre en ligne, *Le Goût de l'archive à l'ère numérique*, puis examiner la diversité des pratiques numériques qu'il reflète avant d'esquisser quelques pistes de réflexion.

Écrire à l'ère numérique : le goût de l'archive à l'ère numérique, un livre vivant

4. *Le Goût de l'archive à l'ère numérique* (Clavert et Muller 2017) est un livre « vivant », qui s'écrit en ligne, est modifié au fur et à mesure de son écriture. L'expérience n'est pas en soi nouvelle : Ian Milligan, Scott Weingart et Shawn Graham l'ont déjà expérimentée (2015)[1], ainsi que Jack Dougherty et Kristen Nawrotzki pour leur *Writing History in the Digital Age* (2013)[2]. La logique de ces deux livres est toutefois un peu différente : si la version en ligne du *Macroscope* est présentée comme un brouillon final, le *Writing History* dispose d'une version en ligne qui fait office de brouillon, d'une version en ligne finale et, comme le *Macroscope*, d'une version imprimée.

5. Toutefois, ces deux exemples, de vivants sont devenus figés par l'édition d'un livre papier ou par l'adjonction d'un livre en ligne désormais immuable. Reprenant les mêmes logiciels, *Le Goût de l'archive à l'ère numérique* est, du moins pour le moment, un livre sans fin dont l'écriture est toujours en cours. Nous utilisons, comme les deux exemples cités, le logiciel de création de sites web Wordpress[3] augmenté par un greffon, CommentPress[4]. Ce dernier autorise auteurs et autrices à structurer leur site web comme un livre – avec table des matières notamment –, à ouvrir aux commentaires non seulement chaque chapitre, mais également chaque paragraphe (figure 2). L'écriture devient alors collective, publique, dynamique. Le système que nous utilisons a ses limites : si l'historique de l'écriture des articles est bien conservé, les lecteurs et lectrices n'y ont pas directement accès, ne pouvant ainsi facilement voir l'évolution de l'article.

1. Version en ligne : http://www.themacroscope.org/?page_id=584.

2. Version en ligne : https://writinghistory.trincoll.edu/.

3. Cf. https://fr.wordpress.org/.

4. Cf. https://wordpress.org/plugins/commentpress-core/ – CommentPress est développé par The Institute for the Future of the Book (http://futureofthebook.org/) qui se présente comme un *think-and-do tank* dont le groupe de New York est affilié aux bibliothèques de la New York University.

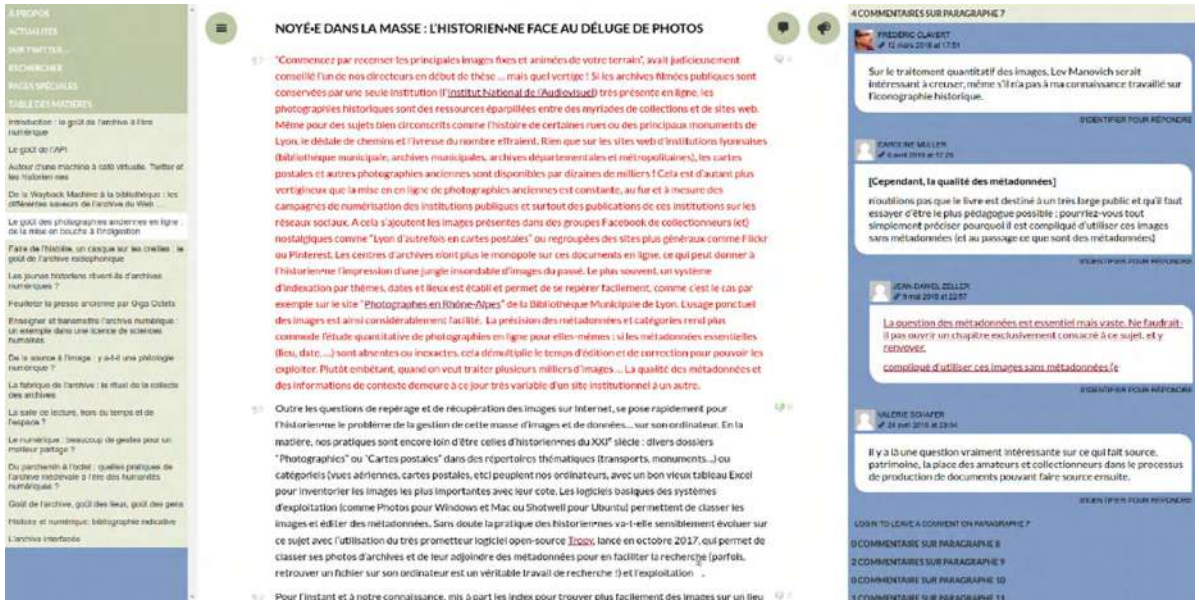


Figure 2. Capture d'écran du site *Le Goût de l'archive à l'ère numérique* (30 octobre 2019)

Crédit : Caroline Muller et Frédéric Clavert

6. Toutefois, la possibilité de commenter au niveau du paragraphe a permis un enrichissement de nombreux chapitres, instaurant une discussion pour l'essentiel entre la co-directrice et le co-directeur de l'ouvrage d'une part et les auteurs d'autre part, parfois entre auteurs, parfois encore entre auteurs et lecteurs[5]. À ce jour, le livre compte 319 commentaires. D'une certaine manière, l'usage du commentaire non seulement par les auteurs et autrices mais également par des lecteurs et lectrices rend visibles les pratiques d'annotation des livres (Moulin 2011), théoriquement privées, en les transformant en pratique collaborative, en donnant la possibilité à l'auteur ou l'autrice de les intégrer (ou de ne pas les intégrer) dans leur propre écriture[6].

7. *Le Goût de l'archive à l'ère numérique* est une publication collective, comprenant 15 chapitres, en plus de l'introduction et de la bibliographie indicative, écrits par 15 auteurs de générations et d'historiographies différentes, pour la plupart historiens ou historiennes, mais également philologues et archivistes. Une coopération avec l'Association des archivistes français (AAF) et notamment avec Céline Guyon, Julien Benedetti et Dominique Naud a permis l'écriture de trois chapitres par des archivistes, portant sur la collecte des archives numériques, sur la salle de lecture et sur les gestes de l'archivage. Si le cœur de ce projet est ainsi un livre vivant, nous souhaitons en faire des « captures » papier : ainsi est paru, sur la base de chapitres publiés sur gout-numerique.net, un numéro de *La Gazette des archives* sur le goût de l'archive à l'ère numérique (Clavert et Muller 2019).

5. Nous développons certains exemples dans la partie suivante.

6. Il serait intéressant de réinsérer d'ailleurs cette pratique du commentaire dans un livre vivant au sein d'une histoire plus large de l'annotation, telle qu'envisage de le faire Claudine Moulin dans *Between the Lines and in the Margins. A Cultural History of Annotation* (à paraître) en collaboration avec l'Institut d'études avancées.

8. La publication de ce numéro de *La Gazette des archives* a d'ailleurs engendré de nombreux questionnements sur l'écriture des chapitres : comment retranscrire dans une version papier la richesse de l'écriture fluide, « liquide » d'un livre en ligne ? Comment faire comprendre au lecteur de la version « papier » la richesse du jeu des commentaires de la version en ligne, qui font partie intégrante de l'expérience de lecture ? Faut-il reporter sur la version en ligne les modifications effectuées, à la demande des relecteurs de la revue, sur les chapitres ? Nous avons apporté des réponses *ad hoc* et largement discutables, relatant l'expérience de l'écriture en ligne dans l'introduction, considérant que les deux versions (papier et en ligne) avaient une vie différente. L'expérience de l'écriture et de la lecture en ligne ou sur « papier » sont ainsi distinctes[7].

9. Cette même coopération avec l'AAF a également abouti à l'organisation d'une journée d'études en novembre 2018, dont le cœur était une discussion entre Sean Takats, alors professeur associé en histoire moderne à l'université George Mason et responsable au Roy Rosenzweig Center for History and New Media des projets *Zotero*[8] et *Tropy*[9], et Arlette Farge modérée par Emmanuel Laurentin sur France Culture[10]. La journée d'études, issue du projet d'écriture du livre en ligne, l'enrichit en retour par la publication, sur le blog du livre, des enregistrements de la journée.

10. Les chapitres du *Goût de l'archive à l'ère numérique* tournent pour beaucoup autour de la collecte des archives – du point de vue historien comme archiviste – sur la constitution des corpus, sur le gigantisme de ces corpus numérisés ou nés numériques, sur les relations à l'archive et à la salle de lecture, sur l'enseignement et les générations en cours de formation d'historiens. Les quinze chapitres, qui confinent d'une certaine manière à une forme d'auto-ethnographie, nous permettent ainsi de dégager quelques conclusions intermédiaires sur les pratiques informatiques et numériques exposées par nos auteurs.

Diversité des pratiques numériques discrètes

11. Notre premier constat est la très grande diversité des pratiques informatiques et numériques des historiens, historiennes, philologues et archivistes qui ont participé à l'écriture du livre, auteurs ou lecteurs. Et cette diversité montre l'ampleur d'une zone grise entre les chercheurs et chercheuses rejetant tout ou partie des pratiques numériques et celles et ceux se réclamant explicitement des humanités numériques.

7. Ces questions se posent pour tout livre ayant une version en ligne, par nature facilement modifiable, et une version papier. Un cas extrême fut celui du *History Manifesto* (Guldi et Armitage 2014), qui a fait l'objet de nombreux débats, parfois très tendus, dont les auteurs ont été accusés de modifier la version en ligne sans indication claire dans un premier temps (Cohen 2015).

8. Zotero est un logiciel libre de gestion des données bibliographiques. https://www.zotero.org.

9. Tropy est un logiciel libre de gestion des photographies prises en centre d'archives. https://tropy.org/.

10. La discussion est disponible en ligne : http://www.gout-numerique.net/640.

12. Des contrastes forts peuvent apparaître, par exemple entre ce constat de Stéphanie Pirez-Huart : « Pour autant, chaque chercheur rencontré, qu'il soit débutant ou plus confirmé, insiste sur le besoin régulier de retour au document. Car l'écran ralentit l'appropriation des informations et le contact avec l'archive est un pan indispensable du métier du chercheur, et c'est bien cette archive qui reste la base de son travail. » (Pirez-Huart 2018) et le point de vue de Frédéric Clavert sur son usage des bases de données (son corpus) : « L'interrogation de la base de données est un vrai dialogue entre la base et l'historien.ne dans la mesure où la requête alors écrite est issue d'une véritable question de recherche » (Clavert 2017).

13. Ces contrastes sont liés aux spécialisations des deux auteurs – respectivement médiéviste et contemporanéiste –, à la nature des archives et des corpus constitués par les auteurs, de la charte médiévale éventuellement numérisée au tweet né numérique, mais également à la manière d'envisager l'analyse et l'interprétation du corpus et à la manière de le lire : la lecture proche, impliquant critique interne et critique externe, d'un manuscrit repose sur une méthodologie différente d'une lecture « distante » (*distant reading**) (Moretti 2005) qui revient à demander à l'ordinateur de lire pour l'historien (voir *infra*).

14. Le plus grand regret régulièrement exprimé est celui de la médiation supplémentaire entre l'historien et son corpus qu'implique la numérisation d'archives « nées papier » : « Ce qui m'amène à ma dernière idée : je pense que ce qui continue à me gêner dans le document numérique, c'est qu'il est médiatisé. Il crée une distance supplémentaire (je suis historienne, ne croyez pas que j'ignore qu'il existe une distance infranchissable entre moi et mes sources…), là où je voudrais la réduire au maximum. Toucher, manipuler, feuilleter… » (Beaulande 2019).

15. Les débats parfois menés en commentaire ont également révélé des écarts importants entre traditions sous-disciplinaires. Ainsi, cet échange (figure 3) entre Caroline Muller et Julie Giovacchini autour du chapitre de cette dernière (Giovacchini 2018) :

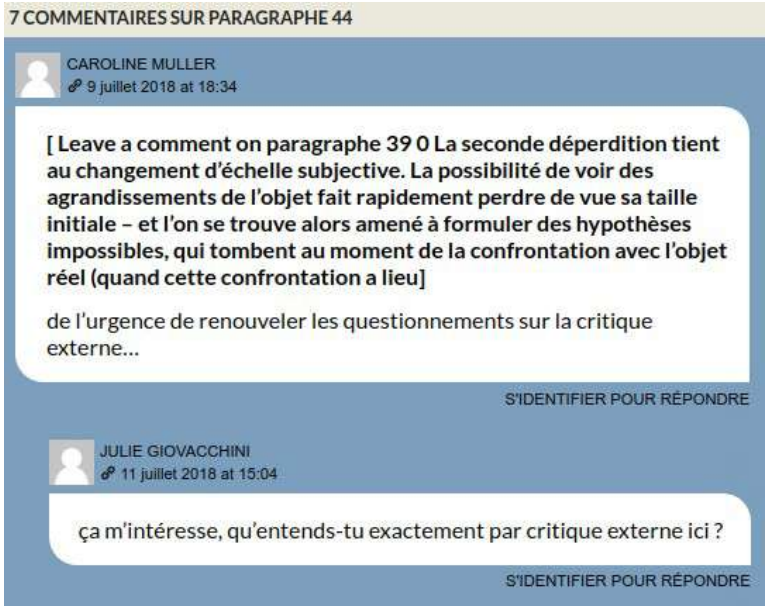


Figure 3. Échange de commentaires autour de l'article de Julie Giovacchini
Capture d'écran. Crédit : Caroline Muller et Frédéric Clavert

16. Ces différences interdisciplinaires sont préexistantes, bien entendu, mais les pratiques informatiques et numériques nous encouragent, comme l'avait déjà noté Peter Haber (Haber 2011), à une relecture et un renforcement de nos méthodes, ce qui peut aussi avoir pour conséquence d'amplifier les écarts disciplinaires. Ainsi se repose une éternelle question, celle de la constitution d'un vocabulaire commun, mais cette fois portant sur nos méthodes et notre lien à l'archive à l'ère numérique.

17. Toutefois, quelques éléments communs peuvent être dégagés. En premier lieu, l'ère numérique est marquée par l'introduction de nouveaux objets dans la salle de lecture des centres d'archives : l'ordinateur, bien sûr, mais également et peut-être surtout l'appareil photographique numérique (figure 4), mentionné par Sébastien Poublanc (Poublanc 2018) du côté historien, par Julien Benedetti du côté des archivistes (Benedetti 2018) et parfois, pour son appareil photo, du téléphone portable[11]. Ces objets sont également porteurs de logiciels – bases de données, logiciels de traitement des données bibliographiques, logiciels de gestion des photographies, etc. – y compris, d'ailleurs, des éléments d'intelligence artificielle.

18. L'appareil photographique numérique (APN) n'est pas uniquement un nouveau mode de lecture des sources, mais aussi une manière complètement différente d'appréhender l'archive et de travailler, qui réduit la présence en centre d'archives sans toutefois nécessairement réduire la présence de l'archive, transformée par l'APN en fichier stockable sur un ordinateur pour une lecture ultérieure. L'APN induit au contraire de devenir « archivore » – pourquoi ne pas photographier ce carton d'archives dont l'utilité n'est pas certaine ? Cette boulimie peut-elle s'avérer piégeuse, engendrer une noyade de la chercheuse dans son corpus ?

Figure 4. L'appareil photographique, outil de l'historienne
© Johanna Daniel.

19. L'introduction de nouveaux objets n'est pas le seul élément notable ressortant des chapitres du *Goût de l'archive à l'ère numérique*. L'importance croissante des salles virtuelles d'archives – Gallica, Europeana, l'IGN, l'Inathèque, etc. – est frappante. Cette salle virtuelle d'ar-

11. Voir le chapitre « Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique » de Johanna Daniel dans cet ouvrage.

chives est souvent liée à des centres d'archives, bibliothèques et autres institutions patrimoniales comme Gallica ou l'Inathèque et nécessite parfois d'être à la fois dans une salle de lecture physique et devant une salle de lecture virtuelle (Loriou 2018).

20. La salle de lecture virtuelle peut aussi mettre en scène des archives nées numériques. Les archives du Web – consultables parfois de chez soi[12], parfois en bibliothèque (le dépôt légal du Web, à la BNF ou à l'Inathèque) – engendrent toute une série de questionnements méthodologiques, y compris sur les pratiques historiennes numériques et, en premier lieu, la question de l'indigestion, de la masse et donc à nouveau de leurs modes de lecture[13].

21. Ces derniers sont, avec les archives numérisées – y compris par l'historien lui-même par l'usage de son appareil photo –, désormais plus divers que jamais. La lecture proche, lecture historienne classique, des sources, est chamboulée par la lecture distante ou à distance, qui est une manière de demander à l'ordinateur de lire, mais non d'interpréter, à la place de l'historien, quand la masse de données – glissement terminologique significatif – est trop importante pour une lecture humaine. Mais en filigrane de chaque chapitre apparaît également la lecture machinique, c'est-à-dire une lecture proche appareillée par l'ordinateur : en d'autres mots, la lecture des fichiers textes aidée, par exemple, par la recherche par mots-clés[14].

22. Enfin, dans de très rares cas, la salle de lecture virtuelle mute en interface de programmation (API*) – ces éléments logiciels permettant à deux programmes, Twitter et un programme installé sur un serveur par exemple, de communiquer et d'échanger, notamment, des données. L'historien n'est néanmoins pas un programmeur, mais un lecteur de sources différent (Clavert 2017).

23. En dehors de la salle de lecture, les nouveaux modes de communication ressortent également comme un élément commun aux auteurs du *Goût de l'archive à l'ère numérique*. Les réseaux sociaux numériques (Muller 2017), les carnets de recherche – le mot sérieux pour dire « blog » – sont présents à différents niveaux de ce livre, y compris, d'ailleurs, dans sa genèse même.

Des pistes de réflexion

24. Le livre vivant *Le Goût de l'archive à l'ère numérique*, les pratiques informatiques et numériques discrètes qu'il donne à voir, offrent des pistes de travail pour le futur de ce projet. Nous en détaillerons quelques-unes ici. La première piste est de réinterroger ce qu'est l'« archive ». Ainsi Céline Guyon, archiviste, s'interroge-t-elle sur le

12. Cf. https://archive.org/.

13. Voir (Schafer 2018), à compléter avec (Schafer et Thierry 2015).

14. Sur les trois modes de lecture, voir (Hayles 2016), notamment p. 156, cité par (Schwerzmann 2018).

répertoire émotionnel – l'un des points les plus importants relevés par Arlette Farge – de l'archive numérique :

25. > L'environnement de la collecte a profondément changé. La collecte est comme mise à distance par la technologie : nous avons besoin de l'informatique pour pénétrer au cœur de l'archive, jusqu'à la donnée, le corps de l'archiviste ne suffit plus pour repérer, classer, sélectionner. Le goût de l'archive ne joue plus avec le même répertoire émotionnel. C'est un répertoire beaucoup plus « intellectuel » et moins sensible qui s'exprime par exemple par la satisfaction d'avoir compris puis résolu une anomalie, qui empêchait le déroulement sans anicroche du processus de versement. (Guyon 2018)

26. L'émotion suscitée par l'archive et éventuellement sa collecte a des conséquences, du moins si l'on suit Arlette Farge. Le passage à un répertoire plus « intellectuel » peut-il influencer l'interprétation du contenu du document ? Peut-il influencer sa (non-)collecte ?

27. *Le Goût de l'archive* d'Arlette Farge est également parfois utilisé comme un modèle de ce que devrait être l'historien, de son travail en archives – l'image de l'historien et de son travail. Cette vision de soi, définition de soi par le biais de cet ouvrage ne correspond plus – si jamais il a correspondu un jour à la majorité de la profession – au travail effectif de l'historien. Les études qui ont été effectuées au moment du lancement du projet *Tropy* ont ainsi montré le phénomène massif de la prise de vue en centres d'archives et le décalage existant entre les récits de soi et les récits en archives et la pratique effective du métier d'historien face aux sources primaires (Mullen 2016). Le chapitre de Sébastien Poublanc, qui relate une expérience pédagogique, montre que les futurs historiens se glissent de manière différenciée dans le moule historien (Poublanc 2018) : certains acceptent pleinement l'image de l'historien en centres d'archives dressée par Arlette Farge, d'autres en doutent nettement plus. Dans ce chapitre, tout n'est alors que décalage face à l'archive : entre le doctorant (qui emmène un appareil photo dans les archives) et son directeur de thèse (qui ne l'emmène pas), entre des enseignants et leurs étudiants quant à la « salle de lecture » et le contact avec l'archive, ou encore entre le goût de l'archive d'Arlette Farge et l'expérience d'aujourd'hui – surtout quand on peut réutiliser les matériaux numérisés par d'autres historiens ou historiennes. De ces décalages, certaines questions émergent, vitales quant à notre rapport aux sources et qui n'ont pas encore toutes des réponses : les nouveaux outils aboutissent-ils à l'abandon de certaines sources ? Faut-il s'interroger sur une nouvelle vision de l'historien face à son corpus ? Quel outillage commun aux historiens, au-delà de l'ordinateur et de l'appareil photographique numérique ?

28. La question de l'outillage commun, d'ailleurs, n'épargne pas les pratiques liées au recours à la salle de lecture « virtuelle ». Quelles nouvelles possibilités ouvre-t-elle pour l'écriture de l'histoire ? La presse numérisée du XIX^e^ siècle est un cas emblématique :

29. Le recours à Gallica est désormais un des réflexes primaires de tout.e.s historien.ne.s. Dans les bibliothèques numériques, la presse ancienne a occupé une place pionnière au sein des programmes de numérisation. La presse ancienne a fait partie des premiers chantiers de numérisation de la BNF. L'accès en ligne à n'importe quel numéro des principaux quotidiens du XIXe et XXe siècles rend l'utilisation des archives de la presse ancienne non seulement plus simple mais plus systématique dans nombre de travaux d'historien.ne.s. (Gaillard 2018)

30. Non seulement sa numérisation a poussé à une plus large et plus complète utilisation de la presse des XIXe et XXe siècles, mais elle en devient une nouvelle source au sens où il est désormais possible de lui poser de nouvelles questions de recherche :

31. Il faudra probablement attendre la mise en place de véritables outils statistiques qui allient les possibilités de Gallica et de la textométrie pour avoir des résultats satisfaisants. C'est d'ailleurs à ce travail que s'attelle Pierre-Carl Langlais dans le cadre de l'ANR *Numapresse*. L'idée de ce travail collectif est d'analyser la presse ancienne à grande échelle avec des méthodes automatisées de fouilles de données pour faire apparaître des dynamiques invisibles à l'œil nu comme la vitalité des contenus médiatiques par exemple. (Gaillard 2018)[15]

15. Claire-Lise Gaillard y fait référence aux travaux de Pierre-Carl Langlais (2017) et au projet *Numapresse* (http://www.numapresse.org/).

32. Comme d'autres archives, nativement numérique (les archives du Web, les corpus fondés sur les réseaux sociaux numériques et bientôt les archives publiques et privées de manière générale) ou non (Gallica), la pratique risque de nous faire tomber dans un ordre illusoire (Milligan 2013), qui pousse à confondre masse et exhaustivité.

Conclusion

33. Nouvelles pratiques, encore peu ou pas explicitées, nouvelles relations à l'archive mais sans nouvelle définition de la profession par elle-même, une méthodologie qui risque d'être désormais « en miettes » (Dosse 2010). Nous avons mentionné cet ordre illusoire qui est l'un des risques liés aux nouvelles pratiques, discrètes, des historiens à l'heure numérique.

34. Parmi les développements futurs du projet, nous aimerions analyser et comprendre le poids des nouvelles cultures du document décrites dans ce chapitre dans les choix méthodologiques d'analyse et d'interprétation de nos sources. Certaines sources, certaines périodes, certaines historiographies aussi, peut-être, se prêtent mieux que d'autres à l'usage de ces nouvelles pratiques, d'où l'importance de les expliciter.

35. Cette explicitation passera fort probablement par un dialogue entre les métiers – entre historiens et archivistes, conservateurs des bibliothèques ou des musées, acteurs classiques de l'écriture de l'histoire et de la médiation à

l'archive, mais pas uniquement. De nouveaux venus sont désormais à prendre en compte, à commencer par les chercheurs et chercheuses en sciences informatiques.

36. Dans les premières pages de son célèbre *Douze leçons sur l'histoire*, Antoine Prost rappelle les débats entre histoire et sciences sociales, notamment avec la sociologie naissante. Les débats opposant François Simiand et Charles Seignobos, par exemple, dans un contexte du début du XX^e^ siècle qui voit le triomphe des sciences expérimentales se sont poursuivis tout au long du siècle. Et d'une certaine manière, comme le rappellent les débats autour du *History Manifesto* (Guldi et Armitage 2014)[16], peut-être sommes-nous, à l'ère numérique, toujours à l'état de crise en histoire. Cette fois, néanmoins, ce sont bien les sciences informatiques, et non plus la sociologie ou les sciences expérimentales, qui nous poussent à revenir sur nos propres pratiques et à les repenser.

16. Une traduction partielle a été publiée, avec des articles très critiques d'autres historiens, dans (« La longue durée en débat » 2015).

Enrichir les corpus, structurer les données

Enrichir un corpus de sources numérisé en histoire de l'éducation. Le cas du *Bulletin administratif de l'instruction publique*

Solenn Huitric

Introduction

1. Le développement de rencontres ayant pour thème les humanités numériques[1] permet à tout le moins de mettre en évidence le nombre de projets de recherche recourant à des technologies numériques entendues dans un sens large, et de poser la question des effets de ces opérations sur nos pratiques de recherche. Ces questionnements sont d'autant plus salutaires que des opérations comme la numérisation de documents sont de plus en plus fréquemment menées, pour répondre aux enjeux auxquels sont confrontées les institutions patrimoniales et aux besoins de projets de recherche. En histoire, nous saluons presque unanimement les vertus de la numérisation et, essentiellement, la possibilité d'accéder à distance à des documents autrement conservés dans des dépôts d'archives plus ou moins accessibles selon les cas. Or, cette dématérialisation des documents permet davantage qu'un simple accès facilité et la numérisation de corpus d'archives constitue également l'occasion d'une réflexion sur les modalités d'enrichissement du corpus que cette opération permet. Les archives sont le matériau premier de l'historien (Farge 1997) et, de la même façon que nous intégrons dans nos questions de recherche les logiques de conservation des documents, il nous faut prendre en compte les opérations qui guident la mise en ligne de corpus.

2. C'est dans cette perspective que le présent article prend appui sur un retour d'expérience autour du projet de Bibliothèque historique de l'éducation (BHE) et, plus précisément, sur la numérisation du *Bulletin administratif de l'instruction publique* (*BAIP*)[2]. L'ensemble du projet *BHE* vise à proposer une version enrichie d'une numérisation de corpus en lien avec l'histoire de l'éducation. Les travaux sur l'histoire des politiques et des institutions éducatives développés ces dernières années se sont pour beaucoup appuyés sur le dépouillement manuel de sources sérielles, administratives et imprimées. Cependant, nous nous sommes rendu compte que ces documents sont souvent difficilement accessibles et que les collections complètes sont peu nombreuses. En outre,

1. L'expression est ici reprise dans une acception large, il n'entre pas dans le propos de cet article de travailler sur une définition possible. Voir (Berra 2012).

2. Le projet *BHE* est né en 2014 et se concentre sur la numérisation de corpus documentaires cohérents en histoire de l'éducation. La recherche présentée ici correspond à la première vague du projet (2014-2017). La BHE entrera en 2020 dans la troisième vague du projet.

le support imprimé rend plus difficile tout traitement sériel d'une certaine ampleur. C'est pour tenter de dépasser ces cadres que leur numérisation a été envisagée. Elle a associé différentes équipes en fonction de leur spécialisation : laboratoires de recherche en histoire (LARHRA) mais également en analyse de corpus textuels (ICAR) et l'unité mixte de service (UMS) Persée. Pour cette première phase de travail, le projet a bénéficié d'un financement permettant que deux personnes s'y consacrent à temps plein, l'une pour la documentation dans la chaîne de traitement de Persée, la deuxième pour la préparation scientifique de l'intégration des corpus sur le portail. La réflexion autour des enjeux de numérisation s'est principalement concentrée autour d'un corpus précis, le *BAIP*, publié tous les mois entre 1850 et 1932 et contenant tous les actes pris par le ministère de l'Instruction publique sur cette période.

3. Les différentes étapes du projet, les interrogations que nous avons dû résoudre et les problèmes survenus ne me semblent pas propres à ce corpus précis. Le premier objectif de ce texte est de poser la question de la position que les historiens peuvent adopter dans cette configuration, étant donné le rapport de notre discipline aux documents d'archives. Un deuxième objectif se concentre davantage sur le changement de rapport à la source que permet et que suppose la numérisation, sur les possibilités que ce changement ouvre ainsi que sur les défis qu'il pose. Afin de montrer que ces enjeux ne se posent pas aux historiens une fois la numérisation achevée mais à chacune des étapes du processus, trois temps sont présentés, qui reprennent les grandes étapes du déroulement du projet *BHE*.

Les historiens et les archives en ligne : usages et hésitations

4. Le premier point d'appui du projet *BHE* n'est pas nouveau : nous utilisons tous des archives disponibles en ligne. En fonction de nos méthodologies de recherche et du type de documents dont nous avons besoin, cet usage varie mais la numérisation nous permet au moins de lire à distance. Au début du projet, nous avons ainsi tenté un tour d'horizon des pratiques des historiens concernés par les corpus à numériser face aux archives en ligne. Ces travaux préparatoires ont associé dès le départ l'ensemble des membres du projet : intégrer les chercheurs a priori « destinataires » de la plateforme BHE a permis de poser la question de la finalité de la numérisation pour que l'outil à construire corresponde le plus possible aux besoins mais également pour que soient pris en compte les coûts et contraintes de chaque choix[3].

5. Les chercheurs sollicités dans le cadre du projet décrivent une pratique fréquente des archives numérisées mais non exclusive car ils combinent différents fonds sous des formats variés. Cette situation est pro-

3. En ce sens, ce texte s'inscrit dans la discussion de la table ronde « Élaborer des corpus numériques : les collaborations entre laboratoires, bibliothèques et centres d'archives » qui s'est tenue lors des rencontres DHNord 2019 (https://publi.meshs.fr/ressources/elaborer_des_corpus_numeriques).

bablement la plus fréquente pour l'ensemble des historiens, comme peut le montrer Sébastien Poublanc lors de ses travaux avec des étudiants en histoire (Poublanc 2018), mais elle s'accompagne cependant de deux autres constats. Tout d'abord, à chaque fois que la situation le rend possible, les chercheurs interrogés ont également une connaissance physique des fonds qu'ils consultent numériquement. Ensuite, le recours à la version numérisée des archives ne s'accompagne pas systématiquement de l'établissement d'une méthodologie propre à ce format. Les pratiques de recherche des collègues ne mentionnent pas d'interrogations spécifiques sur les effets produits par les outils de recherche pour naviguer dans les corpus numérisés (par opposition à une lecture continue) ou sur le recours à des filtres pour circuler dans un corpus, même si ces deux outils peuvent introduire certains biais (Rygiel 2017, chap. 3 notamment ; Gaillard 2018). Nous avons également interrogé les chercheurs sur les limites des portails auxquels ils recouraient et deux aspects se retrouvent dans tous les retours recueillis. Le premier a trait à l'accès aux métadonnées* : pour certains portails, il est difficile de connaître le périmètre de la numérisation et de savoir rapidement ce qui a été numérisé par rapport à ce qui n'a pas pu l'être. La deuxième limite soulevée concerne les outils de recherche dans les collections numérisées, parfois jugés peu performants ou tout au moins peu adaptables.

6. Établir ces différents constats visait principalement à guider la confection d'un nouveau portail. En plus de leurs pratiques actuelles, les historiens ont été invités à décrire les différents usages qu'ils souhaitaient pouvoir faire des corpus proposés à la numérisation dans le cadre du projet *BHE*, ainsi que les questions de recherche qui pourraient bénéficier de ce travail de numérisation. Cette requête a surtout rendu visible le fait que les historiens ne procèdent que rarement dans ce sens et que la constitution de questions de recherche et celle d'un corpus de documents vont de pair. Définir les questions auxquelles la numérisation d'un ou des corpus permettrait de répondre a essentiellement relevé d'une réflexion générale sur les outils à proposer en plus du texte numérisé mais n'a pas permis d'adosser le développement du projet à une recherche précise. Cette configuration nous a incité à tenter de conserver autant de modalités d'emploi du corpus que possible : lecture suivie confortable à l'écran, recherche plein texte adossée à une segmentation* fine des documents, requêtes construites, entre autres. Tenter de conserver cet équilibre fait rejouer une question fréquemment posée au sein des humanités numériques : les méthodologies mises en œuvre doivent-elles aider à renouveler nos questions de recherche ou nos résultats[4] ? En choisissant un protocole fin de documentation lors de la numérisation – sur lequel je reviendrai par la suite –, nous souhaitions faire le pari que nous maintenions ouvert le champ des possibles dans les deux cas.

7. Pour mener à bien cette ambition, nous avons choisi de restreindre dans un premier temps le travail de numé-

4. Cette question a notamment été l'un des fils communs à plusieurs interventions lors des rencontres DHNord 2017, notamment suite à l'introduction proposée par Andreas Flickers, « *Digital History: On the heuristic potential of thinkering* ».

risation enrichie à un corpus précis. L'objectif était de travailler sur un cas cohérent et complet. Le *BAIP* présente plusieurs avantages dans cette optique : la collection à notre disposition est complète, le périmètre de la publication est clairement établi et son contenu permet d'aborder un nombre important de thématiques en histoire de l'éducation. Les textes qu'il contient touchent à tous les niveaux d'enseignement et concernent aussi bien les nominations de personnels enseignants, celles du personnel administratif, l'attribution nominative de bourses, la définition des programmes ou des ouvrages acceptés, entre autres. Ce choix n'était toutefois pas exempt de difficultés : la volumétrie à numériser est importante, le papier assez fin, et la documentation nécessaire pour que chaque texte puisse être isolé, fastidieuse. De nouveau, ces contraintes ont été présentées à l'ensemble des participants au projet pour que le côté peut-être plus technique de la numérisation ne soit pas invisible et pour que la quantité de travail qu'il requiert soit prise en compte.

8. Cette première étape de définition du cadre de la numérisation dans le projet *BHE* a permis de définir des objectifs communs pour l'ensemble des participants. Pour les mener à bien, une numérisation qui aurait correspondu à une transposition numérique du *BAIP* ne suffisait pas. L'objectif du projet est également de tenter de mettre à profit les outils disponibles, notamment au sein de l'UMS Persée, pour démultiplier les pistes de lecture et d'interrogation du corpus.

L'impossibilité d'un enrichissement parfait

9. Un deuxième grand temps de définition des ambitions méthodologiques du projet s'est ainsi ouvert après le choix du *BAIP* comme corpus. L'objectif général de ce type d'opération quel que soit le projet est d'enrichir le corpus de départ. La numérisation constitue une opération modifiant le corpus de départ (Pierazzo 2015) mais il s'agit également d'ajouter des dispositifs au document initial, de profiter du travail de numérisation pour compléter les informations disponibles. Si l'intérêt de ce travail est facile à défendre, il suppose néanmoins un accord sur ce qu'il est possible d'adjoindre pour favoriser la contextualisation ou la compréhension du corpus sans toutefois entrer dans son interprétation, si tant est que cela soit possible. Trois grandes orientations ont été définies dans le cadre du projet *BHE*, qui correspondent à ce qui est généralement fait dans ce domaine.

10. Le premier enjeu est bien sûr de donner à voir le texte de départ, les différents choix opérés pour permettre l'accessibilité en ligne du *BAIP* ayant mis l'ensemble des participants au projet d'accord. Conformément à la chaîne de traitement utilisée par Persée pour l'ensemble des collections que l'UMS met en ligne, le *BAIP* a fait l'objet d'une océrisation. Il a été décidé de ne pas procéder à une correction de l'OCR* (malgré des erreurs de reconnaissance de certaines dates et certains noms propres), principalement pour des questions de faisabilité dans le temps imparti. Pour la mise en ligne, nous avons opté pour l'affichage du fac-similé adossé

à un outil de recherche plein texte. Cet affichage opère déjà un premier déplacement par rapport au texte au format papier : est visible en premier lieu la structure du fascicule et l'utilisateur doit sélectionner le titre du texte qu'il veut voir s'afficher (figure 1). Le découpage de chaque fascicule mensuel en différents niveaux en fonction du thème abordé et du jour nous a semblé correspondre le plus à la façon dont le *BAIP* est utilisé (on cherche un type d'information plutôt qu'à lire l'ensemble des textes pour un mois donné). En outre, un type de document a été assigné à chaque texte (figure 2) pour introduire un outil de recherche supplémentaire. La typologie a été définie en fonction des grilles de lecture des historiens et du mode de structuration des documents de Persée, elle a pour ambition de ne supposer aucune interprétation devant faire appel à des informations extérieures au document. Cette première grande orientation dans l'enrichissement du corpus est donc en quelque sorte nourrie par le corpus lui-même.

11. La logique qui nous a amené à choisir le *BAIP* comme corpus de travail s'explique également par le type de données qu'il contient et la possibilité de mettre en lien ces informations avec d'autres documents existant en histoire de l'éducation[5]. Par exemple, le *BAIP* liste les nominations d'enseignants par établissement, rendant possible un travail de repérage des différents postes occupés par une même personne par le recours aux entités nommées*. Nous souhaitions ainsi repérer les noms d'institutions, de personnes et de lieux pour créer trois index facilitant la navigation dans l'ensemble du corpus. Nous avons donc défini une chaîne de travail devant nous permettre de reconnaître et annoter ces entités nommées. Nous avons pour cela travaillé avec le logiciel de textométrie TXM[6] : par le biais de

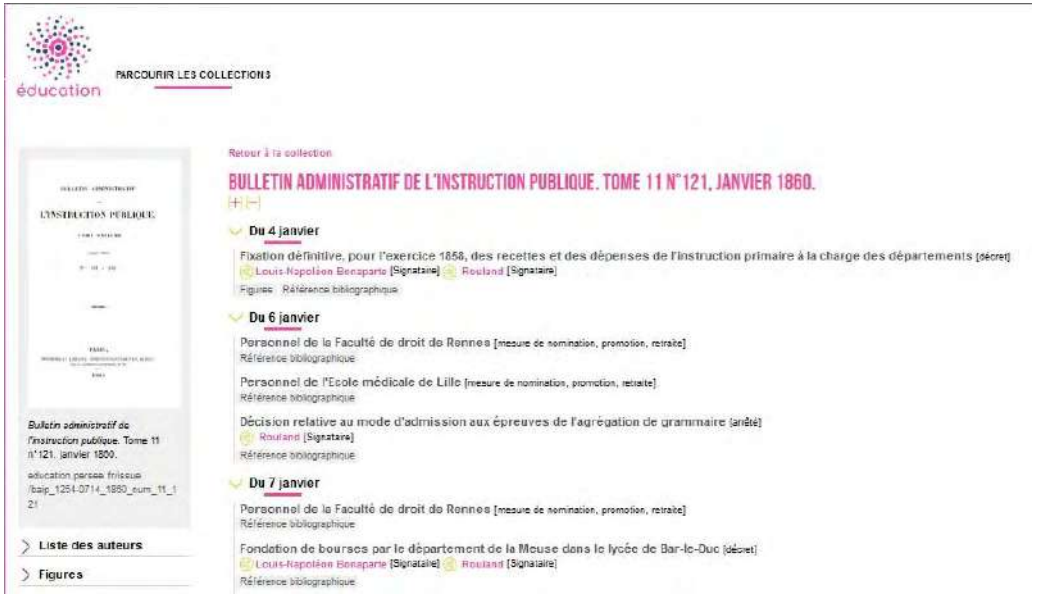


Figure 1. Affichage par défaut de la structure des fascicules du *BAIP*

Crédit : Solenn Huitric

5. Les documents qui pourraient être mis en lien avec le *BAIP* sont de natures variées. Il peut s'agit de dictionnaires biographiques, d'almanachs, de monographies d'établissements, parmi d'autres.

6. « Le projet *Textométrie* fédère les développements logiciels *open source* du domaine pour mettre en place une plateforme modulaire appelée TXM. Il s'agit à la fois d'une opération patrimoniale au rayonnement international et du lancement d'une nouvelle génération de recherche textométrique, en synergie avec les technologies de

requêtes construites, nous pouvions repérer l'ensemble des occurrences d'un même nom. L'équipe qui développe le logiciel étant partie prenante du projet, elle a accepté de travailler au développement du logiciel pour que soit possible une annotation en bloc d'une même occurrence. Cette annotation devait être faite en prenant appui sur une base de données développée au sein du LARHRA. Chaque institution mentionnée dans le *BAIP*, par exemple, serait ainsi créée dans la base de données, avec un identifiant spécifique qui serait utilisé pour l'annotation dans le *BAIP* par le biais de TXM. La base de données pourrait être mise en lien avec ces occurrences et d'autres informations concernant l'institution pourraient par ailleurs y être ajoutées. Pour favoriser une interopérabilité entre ce travail de repérage et d'autres travaux pouvant exister sur ces thématiques, nous avions choisi d'adopter les directives de la Text Encoding Initiative[7] pour l'annotation et de concevoir la création des identités nommées dans la base de données comme des fiches d'autorités. Il faut néanmoins admettre que l'ampleur de ce projet et les dispositifs qu'il requiert ont contraint à un ralentissement de cette partie du projet de la BHE pour favoriser une mise à disposition plus rapide du corpus.

Indication dans le BAIP/BAMIP	Catégorie documentaire utilisée
Loi	Loi
Décret	Décret
Projet de décret	Projet de décret
Arrêté	Arrêté
Règlement	Règlement
Ordonnance	Ordonnance
Instruction	Instruction
Circulaire	Circulaire
Interprétation de circu	Circulaire
Rapport	Rapport
Mesure de nomination, promotion, retraite, révocation, radiation (même en tant que chargé)	Arrêté, décret ou ordonnance de mesure nominative
Concession de bourse	
Nomination à une commission scientifique ou à un conseil de décision	Si précision du type (arrêté ou décret), catégorie correspondante ; sinon : Autre
Nomination à un conseil académique ou départemental	
Admission d'ouvrages	Admission d'ouvrage
Convocation	Convocation
Programme	Arrêté, décret ou ordonnance fixant programme
Varia ou toute autre information pour laquelle un type administratif n'est pas précisé	Autre

Figure 2. Typologie établie pour les textes contenus dans le *BAIP*

Crédit : Solenn Huitric

12. Enfin, une troisième grande orientation dans le travail d'enrichissement visait à mettre à profit la notion de portail. Le projet *BHE* participait en effet à l'inauguration d'un nouveau dispositif au sein de la boite à outils de Persée : les Perséïdes. Une instance spécifique de la plateforme peut être créée autour d'une thématique spécifique[8], ici l'éducation : cela permet de réunir en

corpus actuelles (Unicode, XML, TEI, outils de TAL, CQP, R). » Site internet du projet *TXM*, http://textometrie.ens-lyon.fr/, consulté le 6 janvier 2020.

7. « *The Text Encoding Initiative (TEI) is a consortium which collectively develops and maintains a standard for the representation of texts in digital form.* » Site internet du consortium de la TEI, https://tei-c.org/, consulté le 6 janvier 2020. Voir le glossaire de l'ouvrage pour plus d'informations sur le standard XML développé.

8. « Le label "Perséide" identifie des corpus sous forme numérique, accessibles en ligne et outillés. Ces corpus sont le résultat d'une collaboration entre des équipes de re-

un même endroit différents corpus liés à l'éducation. D'autres corpus d'archives pourraient ainsi être ajoutés au portail selon la même chaîne de traitement. En outre, dans le cadre du projet *BHE*, d'autres opérations de numérisation ont été menées pour les ouvrages publiés en histoire de l'éducation, notamment les outils de travail produits par le Service historique de l'éducation comme des recueils de textes de loi ou des dictionnaires biographiques. L'objectif par ce biais est de favoriser le rapprochement entre documents et de rendre identifiable un espace privilégié d'accueil des ressources documentaires en histoire de l'éducation. En fonction des développements futurs de l'outil Perséide, pourront être envisagés des liens plus formalisés entre documents (entre une même occurrence dans un dictionnaire biographique et dans le *BAIP* par exemple).

13. L'enrichissement possible d'un corpus lors de sa numérisation apparaît ainsi comme un apport important des dispositifs à notre portée. Il requiert toutefois un travail essentiel de discussion entre les participants d'un projet pour parvenir à un accord sur les objectifs assignés aux enrichissements et sur les dispositifs techniques et numériques que ces choix impliquent. Ces étapes demandent du temps mais également l'implication de tous, indépendamment de leur fonction dans le monde de la recherche (il n'est pas possible de revenir après coup sur certains choix, il vaut mieux donc participer à leur discussion). Le travail autour des modalités de création des index a également mis en avant l'importance à reconnaître dès le début du projet de numérisation d'un corpus d'archives les usages différenciés qui peuvent en être faits pour pouvoir définir des niveaux suffisamment génériques d'enrichissement. La création d'index pour rendre un corpus plus maniable suppose ainsi de se mettre d'accord sur ce qu'on dénomme institution par exemple. Enrichir un corpus d'archives nécessite donc un travail de préparation des ajouts au texte, quelle que soit leur nature. Le document de départ demeure le cœur du portail envisagé mais ce qui l'entoure vient modifier notre cadre de lecture.

La création d'archives mouvantes ?

14. L'objectif premier de la BHE demeure la mise à disposition des historiens de l'éducation de corpus cohérents numérisés et une troisième étape du projet consiste à réfléchir aux différentes façons d'y parvenir. Cette réflexion rejoint en outre une attention renouvelée au statut de l'archive dans les recherches en histoire. Deux grands ensembles d'approches ont été identifiés, en fonction de la volonté de l'historien de s'approprier la version numérique du *BAIP*.

15. Le premier usage du *BAIP* numérisé est et demeurera un recours au document numérisé presque comme si de rien n'était. Le bulletin est désormais beaucoup plus acces-

cherche qui en définissent le périmètre et les modalités d'exploitation, et Persée qui apporte sa plateforme et ses compétences en ingénierie documentaire et informatique. » Site internet de la Perséide éducation, https://education.persee.fr/, consulté le 6 janvier 2020.

sible et cela constitue malgré tout le principal apport. À cela s'ajoute la possibilité d'effectuer des recherches plein texte, qui sont toutefois tributaires de la qualité de l'OCR (pour une réflexion sur une recherche sérielle à partir d'archives océrisées, voir récemment Nelzin-Santos 2019). Nous souhaiterions néanmoins inciter les utilisateurs à réfléchir à cet usage du *BAIP* numérisé à partir de deux directions. La première reprend des considérations portées par ailleurs sur les effets de la recherche plein texte. Il n'est aucunement question de remettre en cause l'utilité de ces recherches, mais plutôt de souligner certains effets, notamment dans le cadre de recherches biographiques dans le *BAIP*. En tapant par exemple le nom d'un proviseur dont on souhaiterait reconstituer la carrière, le portail renverra l'ensemble des mentions de cette personne dans le *BAIP* : comme pour n'importe quel corpus d'archives, numérisé ou non, les éventuels trous dans la carrière ne seront pas visibles puisque ne sont consignées que les nominations, l'impression de pouvoir plus rapidement reconstituer des parcours peut être trompeuse (Milligan 2013). Les travaux de recherche produits à partir des données du *BAIP* numérisé pourraient ainsi comporter une présentation de la méthodologie numérique adoptée, au même titre que la méthodologie générale choisie pour l'ensemble de la recherche. Cette proposition va dans le même sens que notre deuxième incitation : nous citons les sources numérisées selon les normes en vigueur pour leur version papier alors même que nous pourrions modifier nos pratiques pour faire apparaître le recours aux archives numérisées ainsi qu'aux outils dont elles sont accompagnées. Cette suggestion est désormais portée par plusieurs chercheurs (Rutner et Schonfeld 2015). Étant donné que notre mode de lecture des archives numérisées ou papier n'est pas parfaitement identique, et même quand nous ne mobilisons pas d'outils spécifiquement numériques pour travailler sur des corpus numérisés, le recours aux corpus numérisés doit pouvoir se montrer dans les productions de la recherche.

16. Dans le cadre du projet *BHE*, nous avons en outre prévu à terme une mise à disposition des chercheurs le souhaitant du corpus annoté du *BAIP*. Notre intention est de permettre une annotation plus personnelle, selon les questions de recherche propres à chacun, mais qui comprendrait déjà les balises apposées pour la création des index. Cette possibilité pose toutefois la question des compétences nécessaires à l'annotation de ces fichiers, d'autant plus cruciale qu'il serait intéressant de pouvoir ensuite mettre à disposition sur le portail ces nouveaux fichiers annotés. Hors du cadre d'un projet de recherche collectif précis, comment partager des bonnes pratiques ? Comment garantir une forme de réutilisation qui suppose un travail de documentation autour des annotations réalisées ? Si nous n'avons pas encore les réponses à ces questions, il ne s'agit pas de les laisser de côté pour autant : on peut considérer en effet que pouvoir discuter des résultats d'une recherche en réinterrogeant le corpus d'archives avec la grille de lecture déjà utilisée constitue une des modalités de faire avancer plus généralement nos questions de recherche.

17. Cependant, tenter de réunir les conditions pour un enrichissement progressif de ce corpus du *BAIP* ne signifie pas que l'on chercherait à aboutir à terme à une hypothétique annotation complète. En étant attentifs à recourir, dans le cadre du projet *BHE*, à des standards déjà établis par ailleurs, il s'agit surtout de favoriser de possibles mises en réseau de sources et de bases de données. Chaque mise en lien répondrait a priori à des questions de recherche spécifiques qui guideraient les alignements réalisés et leur utilisation. En outre, un tel enrichissement progressif pose la question du statut même des documents produits : pour permettre des usages variés, il demeurerait une version de base du corpus numérisé du *BAIP* et il existerait des versions mouvantes de ce corpus, certaines pouvant être construites par plusieurs chercheurs. Ces dispositifs visent ainsi principalement à construire des configurations à même de faire émerger de nouvelles approches en histoire de l'éducation, notamment en favorisant un travail sur les grilles de lecture adoptées par les chercheurs sur un même corpus de textes réglementaires. Si on pouvait penser a priori que ces lectures se superposent sur des textes apparemment explicites, les discussions menées au cours du projet de BHE attestent que les filtres utilisés par des chercheurs sur des thématiques voisines peuvent varier de façon importante. Ce constat ne se veut pas un point d'achoppement et des façons d'établir des dialogues entre des grilles différentes peuvent être pensées (Pierazzo 2019). Il rappelle néanmoins que le corpus d'archives constitue toujours le point de rencontre principal entre chercheurs en histoire ; sa numérisation invite désormais à réfléchir aux modalités de partage de nos pratiques de lecture.

Conclusion

18. Présenter les différentes étapes de définition du cadre de numérisation du *BAIP* permet de prendre conscience de l'effet des choix effectués à chaque étape et l'intérêt d'y participer avec un regard d'historienne. Participer à ce projet m'a permis de mieux cerner le rôle des chercheurs dans le processus de numérisation et ses enjeux. En tant qu'historiens, nous ne pouvons pas nous substituer aux spécialistes des processus de documentation et des sciences de l'information, ni définir les meilleures chaînes de traitement pour la numérisation et la structuration d'un document. Par contre, à partir de notre discipline et en tant qu'utilisateurs de ces corpus d'archives, il relève de notre travail de recherche de participer aux discussions sur la numérisation de corpus d'archives, la documentation associée et les outils développés pour effectuer des recherches dans ces fonds[9]. Se mettre uniquement en position d'utilisateur final peut engendrer des situations d'incompréhension, notamment à propos du temps nécessaire entre la conception et la mise à disposition de l'outil ou concernant le volume du corpus numérisé. En outre, comprendre les défis techniques que

9. Il s'agit ainsi d'éviter ce que pointait Philippe Rygiel en 2004 : « nous sommes en permanence confrontés à un écrit-écran dont les caractéristiques ont été fixées par d'autres, qui sont rarement des historiens et prennent rarement en compte les besoins des historiens, d'ailleurs très divers d'une recherche à l'autre » (Rygiel 2004).

posent les opérations de numérisation permet probablement d'affiner nos questions de recherche.

19. Réfléchir aux modalités de mise à disposition des corpus numérisés invite également à considérer la place que nous accordons aux pratiques de travail que nous avons désormais acquises dans la façon dont nous rendons compte de nos recherches. Si la multiplication des corpus numérisés peut nous inciter à faire évoluer, a minima, nos réflexes de citation des sources, d'autres pistes peuvent émerger à la faveur de projets collectifs. L'accent mis sur le partage de méthodologies de travail lors de rencontres autour des humanités numériques ouvre d'autres possibilités relevant des modalités d'accès aux données de recherche et de la complémentarité entre les différents supports de diffusion de la recherche.

Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non uniforme ?

Alix Chagué, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie

Introduction

1. Le processus d'industrialisation a induit des mutations inédites dans les conditions des travailleurs et travailleuses, avec des effets durables sur l'organisation du travail et les niveaux de vie des ménages des couches ouvrières impliquées dans des activités productives en pleine mutation. Parmi les débats classiques sur les causes et les conséquences de la première industrialisation, l'historiographie récente a insisté sur l'importance de l'apport des femmes et des enfants aux revenus de leurs ménages (Ogilvie 2003 ; Humphries 2010).

2. Après un long silence historiographique, le rôle joué par les femmes dans le développement industriel est désormais largement reconnu (Horrell et Humphries 1992 ; Van Nederveen Meerkerk 2006 ; Van Nederveen Meerkerk et Schmidt 2008 ; Humphries et Sarasúa 2012). Dans l'un des secteurs clés de la première industrialisation, l'industrie textile, elles sont présentes dans toutes les phases du processus productif. Pourtant, les données sont lacunaires lorsqu'il s'agit de s'interroger sur leurs rémunérations, les emplois du temps, leur travail domestique et de les comparer à ceux des hommes qui travaillent dans le même secteur. Si pour le Nord de l'Europe l'historiographie s'est considérablement étoffée depuis quelques années, pour la France les données sont encore très lacunaires (Scholliers 1996 ; Heerma Van Voss, Hiemstra-Kuperus, et Van Nederveen Meerkerk 2010). Une question récurrente et incontournable dans les études qui se donnent pour but de reconstituer les activités des gens ordinaires, et des femmes tout particulièrement, est celle de savoir comment surmonter les carences des sources disponibles pour le passé (Ågren 2018). Si les données quantitatives sont souvent agrégées et ne permettent pas toujours de distinguer par sexe les chiffres fournis, d'autres informations plus qualitatives sont dispersées dans des sources multiples.

3. Comment pallier ce caractère fragmentaire des informations sur le travail des femmes ? Comment parvenir à recueillir et traiter à la fois des sources imprimées et manuscrites, les rendre accessibles dans des corpus homogènes et les rendre facilement exploitables dans une perspective évolutive des questionnements de recherche préalables à l'interprétation historique des phénomènes étudiés ? Le programme *TIME-US* a été conçu dans le but de contribuer à alimenter ce débat en

apportant des données originales pour la France[1]. Il réunit une équipe composée d'historiens et historiennes, de spécialistes du traitement informatique des documents historiques et de spécialistes du traitement automatique des langues (TAL*). Son but est d'interroger les possibilités ouvertes par l'utilisation de corpus de données semi-massifs dans le domaine de la recherche en histoire. Les données recueillies et analysées sont exploitées dans le cadre des recherches conduites en histoire économique et sociale, en histoire de la famille et du genre, en histoire des conflits du travail et de la culture des classes populaires, et plus largement par les sciences sociales concernées par des approches longitudinales sur ces questions.

4. Le programme vise à rendre accessibles des données sérielles issues de sources historiques rédigées en langage naturel pour évaluer le travail rémunéré et non rémunéré, domestique et extra-domestique des femmes dans l'économie du textile et son industrie. Pour la période concernée, l'attention est spécifiquement portée sur les activités recoupant tout le processus de fabrication de produits textiles : de la filature à la confection, en incluant l'assemblage des pièces. Les activités de distribution des produits sont écartées.

5. Deux approches sont adoptées de manières complémentaires (Moretti 2013 ; Clavert 2014) :

- une approche micro-qualitative impliquant une analyse empirique très approfondie des contextes historiques de production des sources utilisées
- une approche quantitative ancrée dans le champ des humanités numériques et mettant en relation historiens et historiennes et spécialistes des outils et méthodes informatiques

6. Nous présentons la méthodologie établie pour la constitution du corpus de textes numériques ainsi que les résultats obtenus dans le cadre de la seconde approche. Elle repose sur la collaboration étroite des membres des équipes des laboratoires LARHRA[2], ALMANACH[3] et ICT[4]. La chaîne de traitement suivie consiste à extraire le texte contenu dans des fichiers images à l'aide d'outils de transcription, afin de produire des données textuelles structurées sous la forme de fichiers XML-TEI[5]. Les fichiers ainsi générés rendent compte de la structure logique du contenu ainsi que d'éléments d'annotation sémantique. Afin d'utiliser ces données textuelles pour répondre aux hypothèses de recherche historique, une interface de requête est mise en place pour faciliter l'interrogation du corpus. Enfin, dans le cadre du travail mené par l'équipe ALMANACH, est explorée la

1. Cf. https://anr.fr/Projet-ANR-16-CE26-0018

2. Laboratoire de recherche historique Rhône-Alpes (UMR 5190), universités Lumière Lyon 2, Jean Moulin Lyon 3, Grenoble-Alpes, ENS de Lyon, CNRS.
3. Automatic Language Modelling and Analysis & Computational Humanities, INRIA, Paris.
4. Laboratoire Identités, cultures, territoires, université de Paris.
5. Cf. « TEI : *Text encoding initiative* » pour plus d'informations sur ce standard XML.

possibilité d'automatiser les tâches de traitement tout en interrogeant la pertinence, c'est-à-dire la faisabilité, le gain ou la perte de temps qu'ils induisent ainsi que la qualité des données obtenues.

7. Nous présentons, sous la forme d'un retour d'expérience, les différentes étapes de notre chaîne de traitements pour la composition du corpus de textes. Elle mêle des procédés automatisés, semi-automatisés et manuels en suivant cinq tâches principales : collecte des doubles numériques, segmentation*, transcription, uniformisation et enfin annotation.

Collecte des doubles numériques

8. Le corpus rassemblé mêle archives économiques et archives juridiques, manuscrites et imprimées, produites à différentes époques. Dans un premier temps, nous avons souhaité utiliser des sources déjà numérisées résultant de campagnes menées par d'autres institutions. Nous avons téléchargé les numérisations de 390 journaux lyonnais du XIXe siècle par l'intermédiaire de Numelyo[6], la bibliothèque numérique de la ville de Lyon, tandis qu'Internet Archive[7] nous a permis de collecter 8 000 fichiers images correspondant aux numérisations de treize volumes appartenant à la série des monographies familiales réunies dans les *Ouvriers des deux mondes* et les *Ouvriers européens* (Cardoni 2012 ; *Les Études sociales : Les monographies de familles de l'École de Le Play [1855-1930]* 2000), et versés par l'université de Toronto. Les deux ensembles ont fait l'objet d'une campagne de tri pour retirer les images non pertinentes pour nos besoins.

9. En parallèle, lors des dépouillements menés dans les centres d'archives des régions de Lyon et Paris, des doubles numériques inédits, essentiels pour le projet, ont été collectés. Initialement, cette numérisation visait à permettre la transcription manuscrite des sources. Lors des expérimentations conduites en vue de l'automatisation de la transcription, il est apparu nécessaire d'établir une véritable méthodologie pour la prise de vue afin d'obtenir des images de meilleure qualité. Il s'agissait notamment d'améliorer le cadrage, l'éclairage et la définition des images, ces deux premiers paramètres étant cruciaux pour la réussite de la transcription automatique. Par conséquent, nous nous sommes dotés d'un dispositif de prise de vue portatif, la ScanTent[8], et d'une plate-forme de stockage partagé afin de faciliter et de garantir la mise en commun des données, le service Sharedocs[9]. De plus, nous avons établi des règles pour la prise de vue, l'organisation des dossiers d'images et la constitution des métadonnées issues de la phase de numérisation. Pour l'ensemble du projet *TIME-US*, nous

6. Cf. https://numelyo.bm-lyon.fr/
7. Cf. https://archive.org/index.php
8. Projet *Read*, Computer Vision Lab, université technique de Vienne et université d'Innsbruck. 2018. « The ScanTent ». https://scantent.cvl.tuwien.ac.at/en/.
9. Cf. https://sharedocs.huma-num.fr/

avons ainsi pris et rassemblé près de 10 000 photographies (Chagué 2018).

10. Nous distinguons cinq sous-ensembles logiques (figure 1) à partir de la totalité des doubles numériques collectés, en fonction du contexte de production des documents originaux, afin d'adapter les méthodes de traitement employées par la suite :

1. Les contraventions à la police des arts et métiers de la ville de Lyon prennent la forme de registres de contraventions, produits entre 1667 et 1781[10]. Ces derniers contiennent un texte manuscrit peu structuré. Le fond a été dépouillé sur la base d'un carottage décennal concernant en particulier les années 1710, 1720, 1730, 1740, 1750, 1760, 1770 et 1780.
2. La presse ouvrière lyonnaise permet de reconstituer une partie des activités du conseil de prud'hommes de Lyon dans la mesure où les archives du Conseil ont disparu. Nous avons identifié neuf titres de journaux[11] dans lesquels sont parus, entre 1831 et 1851, en discontinu, des comptes rendus des audiences prud'homales visant en premier lieu à informer et à garder une trace de la jurisprudence. En fonction de la qualité de l'impression, le texte est parfois très bruité, rarement structuré de manière homogène d'un titre à l'autre et d'un numéro à l'autre.
3. Les minutes du conseil de prud'hommes de Paris pour les tissus (Lemercier 2007) correspondent aux comptes rendus de séances rédigés par le secrétaire de la section du conseil de prud'hommes créé en 1847 pour délibérer sur les contentieux dans le textile. Nous nous sommes intéressés aux années 1847-49, 1858, 1868 et 1878[12]. Ce sont des textes manuscrits très structurés, contenant beaucoup d'informations et dont le scribe change peu.
4. Les monographies familiales de Le Play sont des enquêtes publiées par la Société internationale des études pratiques d'économie sociale entre 1851 et 1908, sous les titres *Les Ouvriers des deux mondes* et *Les Ouvriers européens*. Le texte, imprimé, est très régulier dans sa mise en forme, mais il contient de nombreux tableaux à la mise en page difficile à appréhender informatiquement.
5. Enfin, les rapports de police de la préfecture de Lyon sont un ensemble de rapports rédigés à la suite d'observations sur les mouvements ouvriers à Lyon, en particulier à l'occasion des grèves qui touchent les industries de la soie à la fin de l'année 1894[13]. Ce sont des documents sans structure homogène.

10. Archives municipales de Lyon, HH 214 à 267.
11. *L'Avenir* (1846-1847) ; *L'Écho de la Fabrique* (1831-1834) ; *L'Écho de la Fabrique de 1841* (1841-1845) ; *L'Écho des ouvriers* (1840-1841) ; *L'Écho des travailleurs* (1833-1834) ; *L'Écho de l'industrie* (1845-1846) ; *L'Indicateur* (1834-1835) ; *Tribune prolétaire* (1845-1850) ; et *La Tribune lyonnaise* (1834-1835).
12. Archives départementales de Paris, D1 U10 379, 386, 396 et 405, respectivement.
13. Archives départementales du Rhône, 9 M5.

Ensemble	Nature de l'écriture	Mode de collecte	Nombre d'images	Méthode de transcription	Nombre d'images transcrites
Contraventions à la police des arts et métiers de la ville de Lyon	manuscrit	photo.	2 216	manuelle	1 517
Presse ouvrière lyonnaise	imprimé	téléch.	520	semi-automatique	520
Minutes du conseil de prud'hommes de Paris	manuscrit	photo.	3 131	semi-automatique	1 237
Monographies familiales de Le Play	imprimé	téléch.	6 500	automatique	6 500
Rapports de police de la préfecture de Lyon	manuscrit	photo.	451	semi-automatique	126

Figure 1. Vue d'ensemble sur le traitement des sous-corpus

Extraction du texte

11. À partir des numérisations, nous produisons un texte en langage naturel auquel sont appliqués des outils d'analyse syntaxique et textométrique permettant une lecture distante[14] du corpus (Moretti 2013). La mise en place d'une série d'étapes de traitement est nécessaire pour cela. Chacune de ces étapes nous a conduits à établir des approches heuristiques différentes, adaptées aux documents et aux contraintes rencontrées. Il faut à ce propos signaler une exception dans notre méthodologie : pour des raisons inhérentes à la spécificité de la langue du texte, l'essentiel du corpus des contraventions a été transcrit manuellement et saisi à l'aide d'un traitement de texte. Cette méthode est coûteuse en temps et non reproductible mais produit en général une transcription sans faute. Conformément à nos objectifs en revanche, les quatre autres ensembles ont été traités en suivant trois tâches : segmentation des images, transcription, et uniformisation du texte extrait.

12. La segmentation désigne la reconnaissance des zones de texte sur une image, leur typage et leur ordonnancement. Elle découle de l'analyse de la mise en page (*layout analysis**). Nous avons utilisé les solutions disponibles dans l'interface du logiciel Transkribus[15], qui propose deux options :

14. Cf. « *Distant reading* » pour une brève explication de la différence entre lecture distante et lecture rapprochée.

15. Cf. https://readcoop.eu/transkribus/

1. La première est une implémentation du logiciel de transcription automatique FineReader[16]. Ce dernier offre de bons résultats pour la détection des zones, le typage et l'ordonnancement. Dans Transkribus, FineReader n'est entraîné que pour le texte imprimé et il n'est pas possible de dissocier segmentation et transcription
2. La seconde est développée par le Computational Intelligence Technology Lab (CITlab) de l'université de Rostock (Strauß et al. 2018). Elle propose plusieurs modèles de segmentation adaptés à des mises en pages particulières, comme les cartes postales ou la presse. Même si l'option CITlab fonctionne bien pour les manuscrits, le typage et l'ordonnancement qui en résulte ne sont pas suffisants pour les documents complexes

13. Le résultat de cette segmentation nécessite parfois des corrections manuelles pour supprimer les faux positifs et compenser les faux négatifs en ajoutant des zones ou en redéfinissant les coordonnées d'une zone. Il a ainsi fallu longuement corriger le résultat de la segmentation de la presse ouvrière, que la mise en page en multicolonne rend difficile. Notons toutefois que même dans ce cas extrême, la correction manuelle de la segmentation automatique est plus rapide et moins fastidieuse qu'une segmentation entièrement manuelle.

14. Les stratégies adoptées pour la transcription ont varié en fonction de la nature des sources et de la qualité des images collectées (figure 2). Pour les deux ensembles manuscrits, plusieurs modèles d'HTR (*handwritten text recognition**, ou reconnaissance de texte manuscrit) ont été entraînés à l'aide du logiciel Transkribus, atteignant des taux d'erreurs situés entre 19 % et 5,2 %, selon les données fournies. Deux modèles sont particulièrement satisfaisants :

1. Le premier (« Prud'hommes_ 1858_M 4+ ») est entraîné sur les minutes du conseil de prud'hommes de Paris pour l'année 1858. C'est le meilleur modèle obtenu
2. Le deuxième (« Comb_French_Admin_XIX_M3+ ») résulte d'une tentative de généraliser l'entraînement en combinant des données issues des rapports de police (1894) et des années 1847-1849 et 1858 des prud'hommes parisiens

15. Pour les documents issus de la presse, l'option FineReader a permis de produire rapidement la transcription des 520 pages de bonne qualité en utilisant un modèle préentraîné pour la transcription du texte imprimé. Enfin, pour les monographies, nous avons utilisé Kraken[17], un logiciel en ligne de commande nous permettant d'entraîner notre propre modèle d'OCR*. Avec peu de données, nous sommes parvenus à un modèle (« Model_OD2M ») très efficace. En couplant la segmentation issue de Transkribus et la transcription produite avec Kraken, nous avons traduit la totalité des 6 500 pages, tout en contrôlant le jeu de caractères généré, ce qui facilite la phase suivante d'uniformisation du texte.

16. Cf. https://www.abbyy.com/fr-fr/finereader/

17. Cf. http://kraken.re/

Logiciel	Modèle	Taux d'erreur (CER*)	Quantité de vérité terrain
HTR Transkribus	Prud'hommes_ 1858_M4+	5,2 %	4 577 lignes
HTR Transkribus	Comb_French_Admin_XIX_M3+	8,8 %	20 025 lignes
OCR Kraken	Model_OD2M	2,2 %	1 300 lignes

Figure 2. Présentation des modèles de transcription

16. Après deux ans d'exploration méthodologique, de dépouillement, de collecte de numérisations et d'extraction de texte, près de 9 900 images sur les 12 818 concernées ont été transcrites automatiquement, soit près de 77 % du corpus. Quelle que soit la méthode de transcription choisie, des fichiers XML-TEI sont produits à l'issue de cette tâche, soit par l'intermédiaire de l'API* de Transkribus, soit par le biais de nos propres scripts de transformation de texte brut ou d'interaction avec Kraken[18].

Uniformisation des données textuelles

17. Uniformiser les fichiers XML-TEI et les données textuelles qu'ils contiennent permet d'aligner nos ressources afin de compenser les écarts découlant des outils et méthodes d'extraction employés tout en conservant les spécificités formelles de chaque source. L'enjeu de cette uniformisation est d'obtenir un corpus homogène compatible avec les outils d'analyse du langage déployés par la suite. Le texte doit être corrigé, les graphies alignées et la structure de chaque ensemble modélisée puis implémentée dans le schéma TEI.

18. La régularisation des graphies concerne en premier lieu les abréviations, qui sont développées, et les dates, qui sont uniformisées. Afin de reconstituer le flux des phrases et des paragraphes, les marques de césures qui ont été transcrites doivent être effacées. La stratégie adoptée pour cela repose sur un système de règles et d'expressions régulières*. Pour les césures, le meilleur scénario de résolution est déterminé en fonction des autres formes rencontrées dans le reste des documents.

19. La structuration consiste à détecter et formaliser la hiérarchie de l'information contenue dans un texte. Elle est détectée à partir d'un certain nombre de caractéristiques, parmi lesquelles des indices typographiques, des informations de mise en page (indentation, position sur la page) ou encore des formules récurrentes[19]. Cette tâche commence par une analyse des textes en vue de modé-

18. « ExportFromTranskribus », « LSE-OD2M », « TEITransformation », disponibles sur https://gitlab.inria.fr/almanach/time-us.

19. Par exemple, dans les minutes du conseil de prud'hommes parisiens, la formule « Après avoir entendu » est marque le début du jugement.

liser la structure des informations qu'ils contiennent. Pour les ensembles peu hiérarchisés, la structuration se limite au niveau des unités documentaires (par exemple, les rapports ou les affaires), composées parfois d'un titre ou d'un en-tête, de paragraphes et/ou d'une ou plusieurs signatures. L'annotation du texte se fait par l'intermédiaire de scripts basés sur des systèmes de règles. L'ensemble des monographies contient en outre des éléments liés à l'édition papier, comme les en-têtes, la pagination et les notes de bas de page, que nous souhaitons retirer afin de recomposer les paragraphes interrompus par un changement de page.

20. La structuration facilite la navigation au sein des ensembles et permet leur éditorialisation en vue d'une mise en ligne. En outre, elle rend possible de concentrer les efforts d'annotation sémantique sur les portions que nous savons susceptibles de contenir les informations ciblées pour le projet.

Annotation sémantique

21. L'annotation sémantique vise à repérer dans le texte les éléments d'information que nous souhaitons extraire et comparer dans l'optique de la recherche historique menée. L'élaboration de cette couche sémantique suppose une collaboration étroite entre les spécialistes de disciplines historiques et informatiques. Même si la structure de chaque source diffère, les informations que nous souhaitons en extraire sont similaires et doivent pouvoir être repérées grâce à un encodage commun. L'enjeu est donc de parvenir à établir un modèle d'annotation qui tient compte de la diversité des formes que prennent les informations (Le Fourner 2019). Cette modélisation est intégrée dans le schéma TEI, sous la forme d'une ODD chaînée (Burnard 2016), qui permet de moduler la spécificité des structures de chaque ensemble et l'unité de l'annotation sémantique.

22. Nous repérons trois catégories d'informations, prenant la forme de segments de taille très variables :

- les informations liées aux personnes
- les entités et segments liés au travail et aux rémunérations
- les entités et segments liés à l'expression du temps

23. Pour établir un schéma TEI compatible avec la totalité du corpus, nous partons d'un ensemble puis nous procédons par élargissements progressifs. Durant cette phase d'élaboration, des portions de texte sont annotées à la main afin de valider les choix de modélisation. Une fois le modèle stabilisé, des exemples de texte annotés serviront à fournir des données d'entraînement pour automatiser la tâche à l'aide de l'analyseur syntaxique FRMG (Villemonte de la Clergerie et al. 2009 ; Morardo et Villemonte de la Clergerie 2014), développé à l'INRIA par l'équipe ALMANACH. D'autres outils complémentaires sont utilisés : SEM (Dupont 2017) pour la détection des entités nommées*, et TXM (Heiden 2010) pour des analyses textométriques permettant d'explorer les corpus.

24. À ce stade, une première phase de traitements linguistiques a été conduite sur les ensembles des prud'hommes de Lyon et de Paris ainsi que sur les monographies. Ces trois ensembles représentent un total de 180 098 phrases. Les traitements sont réalisés à l'aide de la chaîne de traitement du français développée par l'équipe ALMANACH et en particulier à l'aide du *parser** FRMG. Les premiers résultats sont encourageants. Les taux de couverture (par phrase entière recevant une analyse complète) vont de 68 %, pour le texte extrait des contraventions, à 91 %, pour la presse ouvrière, et de 78 % à 88 % pour les monographies. L'utilisation d'outils de recherche et de traitements des erreurs devrait conduire à une amélioration de ces taux.

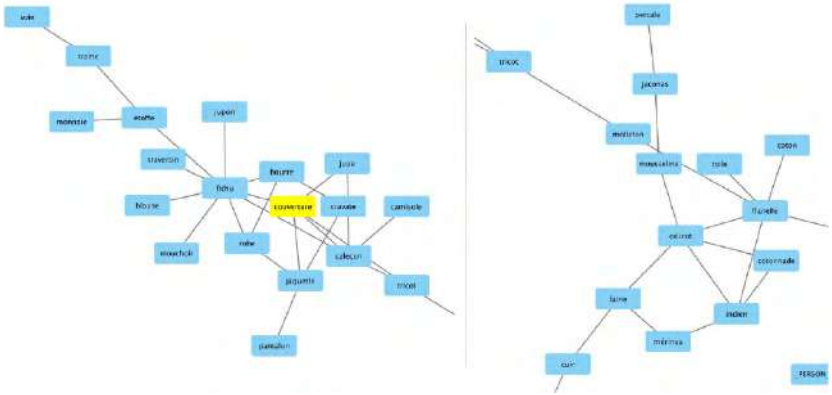


Figure 1. Réseaux sémantiques générés à partir de l'analyse du corpus

Crédit : Alix Chagué, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie

25. La prochaine étape serait d'utiliser les résultats de l'analyse syntaxique pour acquérir des connaissances dans le domaine concerné par le projet *TIME-US*. Il s'agit en particulier d'extraire des termes et des expressions multi-mots, comme « chef d'atelier » ou « paire de bas », mais aussi de construire des réseaux sémantiques (figure 1) basés sur les hypothèses distributionnelles d'Harris (les mots sémantiquement proches ont tendance à apparaître dans des contextes semblables, ici syntaxiques). Cette étape montre déjà des concepts intéressants pour le domaine :

26. Ensuite certains de ces concepts du domaine de connaissance seront utilisés pour trouver des chemins syntaxiques les reliant. Ces chemins permettront, par itération, de détecter de nouveaux concepts et de nouveaux chemins, mais aussi de trouver des occurrences où les sous-ensembles du corpus sont liés. Par exemple, les premières expériences révèlent un lien entre « sieur », « somme » et « franc » : (<root> (<subject> (<agent> sieur/nc)) payer/v (<object> (<patient> somme/nc) (<N2> de/prep (<N2> (<patient> franc/nc))))). Ce chemin est présent 12 fois dans le corpus, par exemple, dans la phase « Le sieur Tocannier payera la somme de 40 fr ». Notons que sont aussi observées et capturées des variantes de cette relation de paiement.

27. En complément, pour faciliter l'analyse du corpus dans son ensemble et obtenir un premier aperçu, l'équipe s'est dotée d'une instance web du logiciel TXM[20]. Ce portail web permet de consulter le corpus en mode lecture et de formuler des requêtes pour des explorations texto-

20. Cf. http://cref.paris.inria.fr/txm/

métriques, en synergie avec les technologies de corpus actuelles (CQP). Pour faciliter encore l'interaction avec la base de données textuelles, une interface web mieux adaptée aux besoins des utilisateurs sera développée de manière à présenter simultanément les résultats de l'analyse syntaxique et les réseaux sémantiques ainsi que les résultats des analyses textométriques. Elle permettra en outre de télécharger des sous-parties du corpus au format XML-TEI ou en texte brut[21].

Premières conclusions

28. L'hétérogénéité des sources qualitatives utilisées est une dimension incontournable des recherches sur l'histoire des femmes au travail et, de ce fait, est inhérente à l'objet de notre projet de recherche. Ainsi, chaque ensemble constituant le corpus du projet *TIME-US* a permis d'explorer un large champ de possibilités techniques. Néanmoins, nous réussissons *in fine* à constituer un corpus qu'il est formellement possible de traiter avec un seul et même outil d'annotation sémantique.

29. Il faut toutefois avoir à l'esprit que l'automatisation de ces traitements suppose d'adopter une méthodologie qui repose sur une différente répartition du temps entre les étapes du processus de traitement des sources : par rapport à une méthode de transcription manuelle, il faut ainsi prendre en compte le temps nécessaire à la préparation des numérisations, à l'entraînement d'un modèle de transcription performant, et à la correction du texte après transcription. Cette méthodologie suppose également de la part des chercheurs une formation aux outils numériques : à Transkribus, Kraken et au langage de requête de TXM mais aussi aux principes de l'annotation sémantique et de l'analyse syntaxique. Nul doute que la répétition de ce genre de protocole permettra d'éviter les écueils chronophages, réduisant le coût technique et humain de tels projets. En d'autres termes, à la question de savoir si l'automatisation permet de répondre au défi posé par le caractère fragmentaire des informations sur le travail des femmes, nous pouvons répondre affirmativement. Enfin, et cela n'est pas le moindre des apports de notre travail, nous avons obtenu un corpus textuel pérenne, distribuable, que la collectivité scientifique pourra exploiter et connecter à d'autres ressources dans le futur.

Pour consulter les données mobilisées dans le chapitre, voir **https://hnso-corpus.nakala.fr/**

21. Pour un aperçu de cette future interface, il est possible de visiter la page « *Exploring a parsed French Wikipedia* », qui permet d'explorer un *dump* du *Wikipédia* français, disponible sur : http://alpage.inria.fr/frwiki/.

Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique

Johanna Daniel

Introduction

1. À partir de l'exemple concret d'une thèse en cours[1], l'objet est d'illustrer le potentiel offert par l'ouverture des données patrimoniales pour la recherche en histoire de l'art[2] tout en soulignant la complexité de la mise en œuvre de projets individuels sur de gros corpus. De l'acquisition des données à leur analyse et leur partage, en passant par le nettoyage, le traitement et l'enrichissement, il s'agit de donner à voir la chaîne de traitement d'un corpus de recherche et de relever quelques enjeux de la collaboration entre chercheurs et professionnels des institutions patrimoniales dans la dynamique des humanités numériques.

1. Daniel, Johanna. « Les Vues d'optique, une production européenne d'estampes semi-fines (1740-1830) ». Thèse de doctorat, Université Lyon 3. http://www.theses.fr/s220481.
2. Le lecteur pourra se référer au glossaire pour plus d'informations en matière d'Open access à l'entrée correspondante.

« Un océan d'images » : étudier la vue d'optique à l'aide du numérique

2. Comme dans les autres disciplines des SHS, les pratiques des historiens de l'art ont été profondément transformées par le numérique : les photographies haute définition permettent aujourd'hui d'explorer d'infimes détails d'une œuvre d'art là où nos prédécesseurs se contentaient de reproductions imprimées d'inégales qualités ; les catalogues des musées, de plus en plus informatisés, s'interrogent depuis n'importe quel point de la planète ; de nombreuses sources textuelles (archives, imprimés) sont désormais numérisées et peuvent être, elles aussi, consultées à distance.

3. La numérisation des œuvres et leur mise en ligne, couplées au développement d'outils informatiques performants et accessibles, ouvrent, pour les historiens et historiennes de l'art, de nouvelles possibilités : faire émerger des questions de recherches inédites, changer d'échelle en travaillant sur des corpus de taille plus imposante, etc.

4. Mon objet de recherche, la vue d'optique[3], se prête justement bien à ces changements d'échelles. Les vues d'op-

3. Pour une synthèse en français sur la vue d'optique, on consultera le catalogue d'exposition disponible en *open access* : Aressy, Lorraine, Bertrand Caron, Henri De Colbert, Morgane Didier et Hélène Palouzié. 2014. *Le Monde en perspective : vues et récréations d'optique au siècle des Lumières. Les collections montpelliéraines de vues d'optique au château de Flaugergues*. Montpellier, France : DRAC du Languedoc-Roussillon. https://www.biu-montpellier.fr/sites/default/files/2019-11/duo_vues_%20doptiques.pdf.

tique sont des estampes destinées à être visionnées à travers une lentille, qui déforme l'image, donnant au spectateur l'illusion d'une profondeur de l'image (figure 1). La production de ces images a été massive, à la hauteur de leur succès commercial : entre 1740 et 1830, une quarantaine d'éditeurs, à Londres, Paris, Augsbourg et Bassano del Grappa, inonde le marché de plusieurs centaines de milliers de feuilles.

Figure 1. L'usage d'un zograscope pour visionner des vues d'optique

J. F. Cazenave d'après Louis-Léopold Boilly, *L'optique*, gravure au pointillé, vers 1794, Amsterdam, Rijksmuseum, RP-P-2015-26-2079 (CC0) https://www.rijksmuseum.nl/nl/collectie/RP-P-2015-26-2079.

5. L'un de mes axes de recherche[4] consiste à comprendre les stratégies commerciales mises en œuvre par les éditeurs, en étudiant les motifs figurés dans les vues, la relation entre l'image et le texte, les choix linguistiques effectués pour la lettre, etc. Il s'agit notamment de retracer la circulation des motifs, car une même vue a souvent été proposée par plusieurs éditeurs, soit qu'ils se copiaient les uns les autres en fonction de la demande (phénomène de contrefaçon), soit que les éléments d'impression circulaient de main en main, au rythme des mariages, associations, successions et ventes après décès.

6. Pour répondre à ces questions, mon approche se veut sérielle et doit englober, autant que possible, toute la production européenne de la seconde moitié du XVIII[e] siècle et des premières décennies du XIX[e] siècle. En l'absence de catalogue raisonné des vues d'optique, il faut au préalable constituer un corpus le plus large possible, de plusieurs milliers de pièces.

7. Fort heureusement, et en dépit de leur caractère modeste, les vues d'optique ont été relativement bien conservées, et de nombreuses institutions en possèdent. Elles sont souvent regroupées en portefeuilles, rassemblant généralement entre cent et deux cents vues, parfois jusqu'à mille[5]. Une partie d'entre elles sont cataloguées, numéri-

4. Sur les axes de recherche développés, voir mon carnet de recherche *Isidore & Ganesh* : https://ig.hypotheses.org/.

5. À ce jour, j'ai identifié plus de 13 000 tirages de vues d'optique conservées dans une soixantaine de bibliothèques et musées européens.

sées et publiées en ligne. C'est le cas notamment des vues d'optique du département des estampes de la Bibliothèque nationale de France (635 vues), du Rijksmuseum (1038 vues) ou encore du musée De Lakenhal à Leyde (677 vues).

Automatiser l'acquisition d'un corpus

8. Pour constituer mon corpus de recherche, j'aurais pu procéder de différentes façons. La première option consistait à ne travailler qu'à partir de quelques fonds institutionnels préalablement sélectionnés sur des critères objectifs (importance matérielle de la collection, représentativité de la production, accessibilité des œuvres), au risque de lacunes et de biais. La seconde était de constituer un corpus, en sélectionnant un à un le « meilleur » exemplaire de chaque vue d'optique portée à ma connaissance, c'est-à-dire le plus représentatif et le mieux conservé, comme on sélectionne une édition particulière d'un texte. Tirant parti des possibilités offertes par le numérique, j'ai opté pour une troisième méthode : accumuler, dans une base de données, un maximum de vues d'optique, sans sélection préalable au sein des séries constituées par les institutions. Le but étant, dans un second temps, de traiter ces milliers d'images pour regrouper les exemplaires similaires (les doubles sont nécessairement nombreux) et de distinguer les versions différentes d'un même motif.

9. La première étape consiste à récupérer les métadonnées* de catalogage (transcription des titres et mentions de responsabilité, indexation des éditeurs et sujets représentés, cotes des exemplaires, etc.) et les images numérisées disponibles sur les sites des institutions conservant des vues d'optique. Effectué manuellement, ce travail de copier-coller vers ma base de données personnelle aurait été fastidieux. Heureusement, il existe diverses solutions pour automatiser l'acquisition des données : interrogation via des API* (*Application Programming Interface* – interface de programmation applicative), téléchargement en CSV* ou encore recours au *web scraping**.

10. Certaines institutions proposent, en plus de l'interface classique de consultation des collections, l'accès à une API qui permet d'interroger et récupérer en masse des données. Précurseur, le Rijksmuseum a été le premier musée en Europe à proposer un tel service dès 2012. L'API de la Bibliothèque nationale de France (BNF), quant à elle, est en service depuis l'automne 2017. Dans les deux institutions, l'API permet non seulement d'extraire les métadonnées de catalogage mais également de télécharger, en haute définition, les images numérisées[6].

6. Dans le cas du Rijksmuseum, données et images sont placées sous licence Creative Commons CCo (https://www.rijksmuseum.nl/en/data/terms-of-use, consulté le 30 mars 2020). En ce qui concerne les métadonnées, la BNF a opté pour la licence Etalab 2.0 (http://api.bnf.fr/conditions-generales-dutilisation-du-site-bnf-api-et-jeux-de-donnees, consulté le 30 mars 2020). Les images quant à elles sont gratuitement réutilisables pour un usage non-commercial et, depuis octobre 2019, pour des usages académiques ou scientifiques. Les usages commerciaux sont soumis à licence

11. Pour le chercheur qui veut se confronter à ces API, le chemin est cependant semé d'embûches, car leur usage nécessite quelques préalables techniques. Il faut ainsi formuler sa requête dans une syntaxe particulière (CQL* pour l'API de la BNF), dont la documentation[7] peut, à juste titre, paraître assez aride. Nous sommes ici loin du confort ergonomique offert par les interfaces graphiques de « recherche avancée ».

12. Une fois passée cette première étape et la requête envoyée, l'usager doit aussi être en mesure de traiter la réponse formulée par l'API, le plus souvent un fichier au format JSON* ou XML* : leur manipulation pour extraire les données souhaitées exige la maîtrise d'un langage informatique (par exemple Python*) bien qu'il soit possible de transformer le fichier en tableur avec des outils plus abordables comme OpenRefine, qui dispose d'une interface graphique.

13. Certaines institutions ont compris que l'usage d'une API – qui répond à certains besoins spécifiques – n'était pas à la portée de tous les publics qu'elles visaient, du fait de cette barrière technique. Aussi quelques-unes proposent des exports de leurs données dans un fichier tabulaire au format CSV, plus maniable. La structure des informations y est moins précise que dans un fichier JSON, mais le contenu est immédiatement exploitable sans compétences informatiques particulières (Bardiot 2018).

14. C'est le choix qui a été fait sur le portail *open data* du Ministère de la Culture, où l'internaute peut interroger une extraction de la base Joconde[8]. Une interface graphique permet de formuler la requête (recherche simple et filtres), dont le résultat est présenté sous forme de tableur téléchargeable. Interrogée sur le terme « vue d'optique », la base Joconde renvoie environ 300 notices, correspondant aux collections du MUCEM, du musée des Beaux-Arts de Bernay et du musée municipal de La Roche-sur-Yon.

15. Dans le cas précis de la base Joconde, cependant, l'usager est confronté à un autre problème : la complétude des données. En effet, la plateforme *open data* du ministère de la Culture ne propose pas un accès exhaustif à la base Joconde, mais seulement un extrait. Sont notamment exclus des champs requêtables et téléchargeables les champs « commentaire » et « inscriptions »[9]. L'argument avancé est le droit d'auteur, certaines descriptions rédigées par les catalogueurs pouvant en effet relever de la propriété intellectuelle lorsqu'il s'agit d'un contenu original et non d'une simple transcription ou d'une

spécifique et payante (https://www.bnf.fr/fr/reproduction-des-documents, consulté le 30 mars 2020).

7. La documentation de l'API BNF est accessible à l'adresse : http://api.bnf.fr/api-gallica-de-recherche.

8. La base Joconde est le nom donné au catalogue collectif des collections des musées de France. Y sont décrites 600 000 notices d'artefacts conservés dans les collections publiques. Depuis 2019, cette base est interrogeable via le portail POP. Le jeu de données publié sur le portail *open data* ministériel est accessible à l'adresse : https://data.culture.gouv.fr/explore/dataset/base-joconde-extrait/information/.

9. Notons également que le téléchargement des images n'est pas prévu par l'outil.

description factuelle[10] (cependant, sur les fiches consultées dans la base Joconde, jamais le nom du catalogueur n'apparaît). Dans le cas qui m'intéresse, les informations figurant dans « inscriptions » et « commentaires » sont justement parmi les plus précieuses : c'est là que les catalogueurs ont recopié manuellement la lettre des estampes où apparaissent le nom et l'adresse des éditeurs des vues d'optique, autant d'éléments indispensables à l'histoire de l'estampe.

16. A contrario des exemples jusqu'ici développés, la plupart des sites et bases institutionnels, ne propose aucun outil d'export des résultats, sinon un téléchargement, une à une, des images numérisées (quand le clic droit n'est tout simplement pas bloqué), et éventuellement de la notice individuelle (au format TXT ou PDF). Cela s'explique parfois par un manque de moyens (interfaces vieillissantes, solutions propriétaires peu performantes) ou par une non-appréhension du besoin (seule la consultation une à une des notices a été envisagée).

17. On objectera – à juste titre – que l'usager peut directement s'adresser à l'institution pour obtenir par retour d'e-mail un export de la base de données. Si dans beaucoup de cas la réponse est positive une fois les motivations exposées, il arrive que certaines institutions ne donnent pas suite, faute de moyens humains suffisants pour traiter la demande. Il arrive aussi, mais c'est heureusement rare, que la requête suscite des crispations, l'institution étant réticence à confier à des tiers des données et des fichiers dont l'usage, lui semble-t-il, pourrait lui échapper[11].

18. Dans ces cas, le chercheur n'aura d'autres choix que de se tourner vers la recopie manuelle du catalogue en ligne qu'il consulte, ou, solution nettement plus efficace mais toujours à la limite de la légalité, le recours à un outil de *scraping* de site web, tel que WebScraper[12].

19. Grâce à ce programme, qui étend les fonctionnalités du navigateur, il est possible d'extraire des données d'une ou plusieurs pages. Dans le cas d'un usage sur un catalogue en ligne de collection, il faut d'abord repérer sur une page test les informations à extraire, puis fournir au logiciel la liste des URL à traiter, pour récupérer, à la sortie, un tableur contenant toutes les données souhaitées. Assez abordable techniquement, cette solution demeure néanmoins tributaire de la qualité du code du site à scraper : si celui-ci est mal structuré, le *scraping* sera incomplet voire inexploitable[13].

10. Sur l'ouverture de la base Joconde, voir le wiki Ouvre-Boîte (https://wiki.ouvre-boite.org/index.php?title=Joconde, consulté le 19 avril 2020) et les lettres d'informations Joconde n° 32 (mars 2018, http://www2.culture.gouv.fr/documentation/joconde/fr/apropos/lettre_info_32.pdf) ; et n° 33 (juin 2018, http://www2.culture.gouv.fr/documentation/joconde/fr/apropos/lettre_info_33.pdf).

11. Maria Vlachou a exposé le cas pour les reproductions numériques d'œuvres d'art (Vlachou 2018).

12. Logiciel propriétaire (avec version limitée gratuite), disponible à l'adresse : https://webscraper.io.

13. L'outil s'appuie sur les balises HTML pour repérer les informations à extraire. Pour un résultat satisfaisant, il est indispensable que les pages et les champs à récolter présentent une certaine homogénéité dans leur structure.

20. Au cours de ma première année de thèse, j'ai collecté sur internet près de 5000 vues d'optique provenant d'une dizaine d'institutions patrimoniales européennes[14]. Dans chaque cas, y compris lorsqu'une API était mise à disposition, il a fallu mettre en place une chaîne de traitement particulière s'adaptant aux spécificités de chaque site fournisseur pour acquérir les données et les numérisations. Autant d'étapes très chronophages qu'il convient maintenant de détailler.

Des données exploitables en l'état ? Un nettoyage nécessaire

21. Une fois les données extraites des sites institutionnels, le chercheur pourrait penser être au bout de ses peines : il dispose de données structurées (ici des tableurs ou des fichiers JSON) qu'il n'a plus qu'à intégrer dans sa propre base de données. Cependant, il lui reste encore des manipulations à faire pour disposer de données réellement exploitables. En se plongeant dans les données collectées, il apparaît vite qu'elles sont très hétérogènes, c'est-à-dire que les pratiques de catalogage varient fortement d'une institution à l'autre.

22. Prenons un cas concret, celui de trois tirages d'une même vue d'optique, la vue de la chapelle de Versailles éditée par Georg Balthazar Probst (figure 2), cataloguée par trois institutions différentes : la Bibliothèque nationale de France (figure 3), le musée De Lakenhal à Leyde (figure 4) et la bibliothèque numérique de Valenciennes (figure 5).

Figure 2. Vue d'optique de la chapelle de Versailles, éditée par Probst

Johann Friedrich Leizelt, *Vüe particulière de la Chapelle du Chateau de Versailles, du côté de la Cour*, vue d'optique éditée par Probst, seconde moitié du XVIII[e] siècle, eau-forte coloriée, Paris, Bibliothèque nationale de France, département des estampes et de la photographie, LI-72 (3)-FOL. https://gallica.bnf.fr/ark:/12148/btv1b6949131n.

23. L'estampe porte quatre titres, en français, allemand, italien et latin. La BNF a transcrit les quatre, choisissant le titre français comme titre principal et les trois autres comme titres alternatifs. Sur la bibliothèque numérique

14. Dans le même temps, 1600 notices descriptives m'ont été fournies par 5 institutions sous format tableur ou PDF (collections non disponibles en ligne) et j'ai moi-même catalogué manuellement 1300 vues d'optique conservées dans 10 institutions françaises et italiennes.

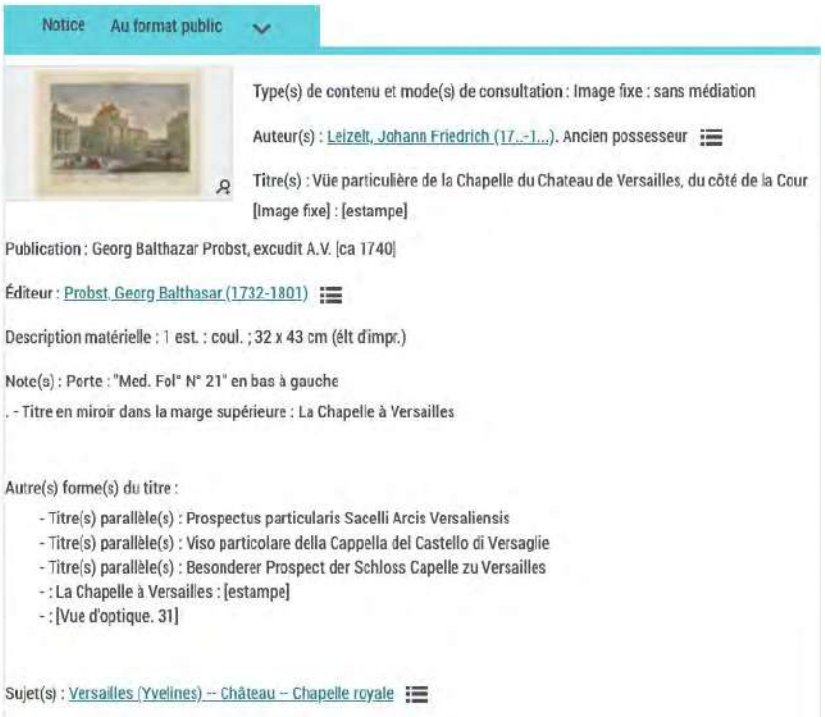


Figure 3. Capture d'écran du catalogue de la Bibliothèque nationale de France

Notice décrivant la vue d'optique *Vüe particulière de la Chapelle du Chateau de Versailles (…)*, éditée par Probst. Catalogue de la BNF : https://catalogue.bnf.fr/ark:/12148/cb41445839k.

Figure 4. Capture d'écran du catalogue des collections du musée De Lakenhal

Notice décrivant la vue d'optique *Vüe particulière de la Chapelle du Chateau de Versailles (…)*, éditée par Probst. Collections en ligne du musée De Lakenhal : https://www.lakenhal.nl/nl/collectie/3121-87.

Figure 5. Capture d'écran de la bibliothèque numérique de Valenciennes

Notice décrivant la vue d'optique *Vüe particulière de la Chapelle du Chateau de Versailles (…)*, éditée par Probst. Bibliothèque numérique de Valenciennes : https://patrimoine-numerique.ville-valenciennes.fr/ark:/29755/B_596066101_G-A18PRO0002.

de Valenciennes, seuls les titres en latin et français ont été retenus. Ils sont indiqués ensemble comme titre principal, séparés par un signe égal « = ». Dans l'interface des collections du musée De Lakenhal, c'est un titre en néerlandais qui apparaît : il s'agit d'une traduction du catalogueur, mais aucun élément ne permet de l'identifier comme un titre forgé. La lettre en quatre langues est bien transcrite, mais de façon non structurée, dans la description. Si l'on s'intéresse maintenant aux mentions de responsabilités (ici, la signature du graveur et l'adresse de l'éditeur), elles ne sont pas transcrites dans le cas de la bibliothèque de Valenciennes. Elles le sont exhaustivement sur le site du musée De Lakenhal, dans le champ « *signatuur* ». Quant à la BNF, le catalogueur a retenu la mention de l'éditeur (champ « publication »), mais pas celle du graveur. Sur cette base, les différents acteurs ont été indexés : la bibliothèque municipale de Valenciennes indique Probst comme auteur et ne fait pas mention du graveur Leizelt. Dans le catalogue de la BNF, les deux sont indiqués dans deux champs différents : Probst comme éditeur et Leizelt comme auteur, ce qui est juste. Enfin, dans le catalogue du musée De Lakenhal, les deux sont indiqués dans un même champ, intitulé « *makers* ». Leur rôle est précisé : graveur pour Leizel et « *Uitgever/drukker* » (éditeur) pour Probst. On notera ici des divergences dans l'écriture du nom du graveur (avec ou sans « t ») et dans l'indexation « Balthasar Friedrich Leizel » pour Lakenhal et « Leizelt, Johann Friedrich » à la BNF.

24. Plusieurs difficultés sont donc à relever ici. Chaque institution utilise des schémas de catalogage différents, avec des champs divergents qui ne se recoupent pas toujours. Ils dépendent de pratiques métiers parfois éloignées (bibliothèques, musées), mais aussi des solutions logicielles employées. Les choix de transcriptions, à la fois dans la nature des éléments à relever (tous les titres, un seul titre ?) et dans l'orthographe (abréviation, modernisation) varient également en fonction des institutions. Quant à l'indexation, elle repose sur des référentiels différents. Ces divergences sont accentuées lorsque l'on compare les données produites dans deux pays différents, qui ne partagent pas la même langue. Enfin, les datations qui apparaissent dans les catalogues (vers 1740 pour la BNF et Valenciennes, deuxième moitié du XVIII[e] pour Lakenhal) n'ont pas été relevées sur les estampes[15], mais sont tirées d'autres sources, non précisées, probablement bibliographiques. L'absence de source pose la question pour le chercheur de la fiabilité de ces informations.

25. Pour rendre exploitables ces données hétérogènes de provenance diverses, le chercheur doit s'atteler à un fastidieux travail de nettoyage, d'uniformisation et enfin, éventuellement, d'enrichissement. Au préalable, il est indispensable de conduire un véritable exercice d'analyse et de critique des métadonnées produites par les institutions, en s'interrogeant, notamment sur :

- la granularité du catalogage (notice sommaire ou détaillée)
- la structuration des métadonnées

15. Rares sont les vues d'optique qui portent une date gravée dans la matrice. Il faut généralement s'appuyer sur les périodes d'activité des différents éditeurs pour estimer une période de publication approximative.

- les choix de transcription (littérale ou modernisée, exhaustive ou partiel)
- les référentiels employés (thésaurus, etc.)
- la fiabilité et la traçabilité de l'information

26. Cela nécessite de comprendre le travail du catalogueur et donc, souvent, de maîtriser soi-même les bases des pratiques métiers. En effet, rares sont les institutions qui documentent clairement et précisément sur leur site les choix de catalogage, choix qui ont d'ailleurs pu évoluer dans le temps[16]. Il faut donc s'atteler à les reconstituer, à les redocumenter par une analyse critique des notices. Ici, l'aide des agents issus de ces institutions est particulièrement précieuse.

27. Une fois l'analyse des métadonnées réalisées, je commence le nettoyage de mes extractions : il s'agit d'abord de désagréger l'information (séparer deux informations qui ont pu être rassemblées, par exemple des titres alternatifs), puis de toiletter les données, en supprimant celles jugées non pertinentes ou non fiables (telles que les dates non sourcées). Il faut ensuite faire concorder les champs avec mon propre schéma de description et uniformiser les données en fonction des règles de transcription et des référentiels d'indexation que j'ai adoptés pour ma base. À ce stade, il est possible d'enrichir la notice, en transcrivant manuellement une information non relevée ou en indexant le sujet. Pour ce faire, il est souvent nécessaire de revenir à l'estampe elle-même, ce que la numérisation des œuvres facilite grandement. Cependant, encore faut-il avoir pu télécharger l'image numérisée et que cette dernière soit de qualité suffisante pour lire le texte gravé. On ne peut ici que déplorer les choix effectués par certaines institutions qui s'opposent à la mise en ligne de fichiers images en haute définition[17] ou à leur récupération par les internautes : elles rendent alors difficile, sinon impossible, le travail de recherche[18].

28. Toute cette étape de nettoyage, d'harmonisation et d'enrichissement peut en partie être automatisée, à condition de maîtriser un langage de programmation adapté. Il est aussi possible d'optimiser le traitement avec des logiciels à interface graphique, tel qu'OpenRefine[19]. C'est le choix que j'ai fait.

16. Les pages professionnelles de la BNF, ainsi que la bibliothèque numérique de l'ENSSIB offrent à tous un accès précieux à la documentation métier.

17. Certaines institutions imposent encore des limitations de quelques centaines de pixels de largeur et de hauteur pour des images mises en ligne sur le Web. Ce genre de pratiques avait une raison d'être technique à l'époque où les écrans d'ordinateur offraient une faible résolution et les fournisseurs d'accès à internet des débits limités. Aujourd'hui, ces restrictions visent surtout à empêcher une diffusion des clichés hors du contrôle de l'institution. À titre d'exemple, une des pages de documentation de Joconde, aujourd'hui hors ligne, indiquait en 2018, « Si la qualité des clichés numériques reversés sur Joconde tend à s'améliorer, bon nombre d'entre eux ne sont pas de haute qualité, notamment pour éviter le piratage éditorial ». Voir la capture de la page au 12 février 2018 sur la Wayback Machine : https://web.archive.org/web/20180212152649/http://www2.culture.gouv.fr/documentation/joconde/fr/apropos/pres_det_joc.htm.

18. À propos de l'ouverture des images numériques des institutions patrimoniales et l'impact de l'*Open Content* sur la recherche, voir (Denoyelle *et al.* 2018 ; Petermann 2018).

19. OpenRefine est un logiciel libre de nettoyage et de mise en forme de données. Il permet notamment d'effectuer des modifications en série sur des tableurs importants, de façon bien plus fine que les logiciels de tableurs traditionnels, comme Excel ou Calc. Pour télécharger OpenRefine : https://openrefine.org/. En français, on trouvera

Traiter, exploiter et analyser

29. Une fois les données acquises, nettoyées, structurées, elles sont enfin prêtes à être versées dans ma base de données. Vient alors le temps du traitement, de l'exploitation et de l'analyse. Une première opération, dans le cas de mon corpus de vues d'optique, a été de repérer combien de sujets différents apparaissent, et pour chacun de ces motifs (que j'appelle « vue type »), combien il en existe de versions différentes (c'est-à-dire de tirages distincts par des éditeurs différents – ici nommés « version »).

30. Reprenons le cas de la chapelle de Versailles (figure 6). Le motif est copié d'une estampe dite « savante » gravée par Jacques Rigaud vers 1740-1750[20] pour sa série des *Maisons royales de France* (figure 7). Jacques-Gabriel Huquier l'a reprise en vue d'optique (figure 8) vers 1760 (Version AA : exemplaires conservés à la BNF, à l'INHA et à la bibliothèque municipale d'Évreux). Probablement au moment de son émigration vers l'Angleterre, à la fin de la décennie, la matrice est passée aux mains de Basset, qui a remplacé l'adresse de son prédécesseur par la sienne (Version AB : exemplaire conservé à la BNF). Autour de 1790, la même plaque est exploitée par Jacques Chereau, après que ce dernier ait à son tour gratté l'adresse de Basset pour indiquer « A Paris, chez Chereau, rue St. Jacques (…) aux 2 colonnes, N° 257 » (Version AC : exemplaires conservés à la BNF

Figure 6. Différentes versions de la vue d'optique de la chapelle du Château de Versailles

Figure 7. Vue de la chapelle de Versailles par Jacques Rigaud

Jacques Rigaud, *Vüe particulière de la Chapelle du Chateau de Versailles, du côté de la Cour*, vers 1740-1750, eau-forte.

Cliché provenant de Wikimedia Commons, publié en CC0

d'excellents supports de formations réalisés par Matthieu Saby sur : https://msaby.gitlab.io/formation-openrefine-BULAC/.

20. La datation est proposée par Bentz et Ringot (2009).

et à Lakenhal). La version de l'éditeur allemand, Probst, évoquée plus haut, a été imprimée à l'aide d'une autre matrice. Si l'on se fie au système de numérotation de l'éditeur, cette estampe peut être placée au début de sa production de vues d'optique, vers 1765-1770. Il en existe enfin une troisième version publiée à Londres par John Tinney, probablement avant 1761, année de la mort de cet éditeur (un exemplaire est conservé au Rijksmuseum).

Figure 8. Comparaison des lettres de trois états de la matrice de la vue d'optique de la chapelle du Château de Versailles

Modification du texte gravé sur la matrice, état de Huquier (version AA), état de Basset (version AB) et état de Chereau (Version AC).

31. Pour chaque version et vue type identifiée, une nouvelle notice descriptive est générée dans la base de données. Exemplaires, versions et vues types sont liés entre elles, ce qui permet de facilement visualiser et comparer tous les tirages d'une version et toutes les versions d'une vue (figure 9).

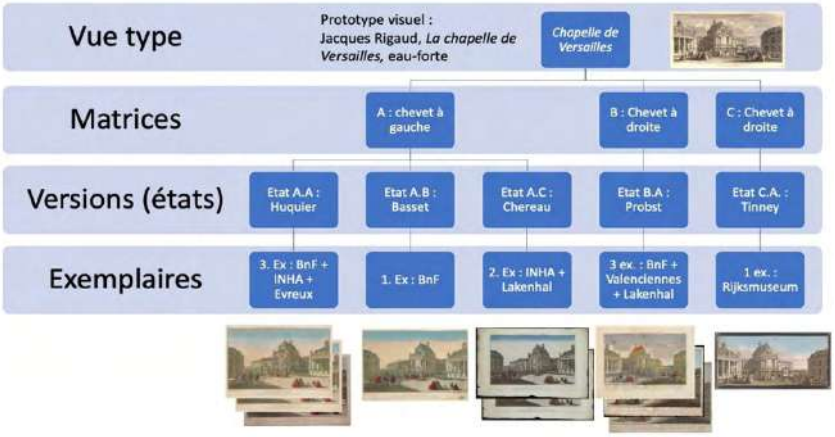


Figure 9. Modélisation des versions de la vue d'optique de la chapelle du Château de Versailles

Crédit : Johanna Daniel

32. Le traitement de chacun des tirages, à l'échelle d'une base de données contenant plusieurs milliers de notices, est complexe. Fort heureusement, le recours aux outils numériques (ici encore OpenRefine) facilite l'automatisation de certains rapprochements, par comparaison des métadonnées.

33. Une seconde phase d'opérations, actuellement en cours, consiste à appliquer des traitements statistiques pour quantifier des phénomènes, visualiser des répartitions, identifier des tendances (quels sont les motifs les plus repris par les éditeurs ? Y a-t-il des spécificités nationales dans le choix des sujets ? etc.). Par recoupement d'informations il est alors possible de produire de nouvelles connaissances sur ces vues d'optique : de reconstituer la

circulation d'un motif entre divers éditeurs, et d'affiner la datation d'un tirage sur la base des dates d'activités des éditeurs et des circulations des éléments d'impression. Ces nouvelles informations viennent enrichir la base de données et, dans un processus itératif, permettent d'affiner les traitements statistiques.

Lecture distante, traitement computationnel et regard outillé

34. Un tel travail sur un corpus d'une taille inhabituelle en histoire de l'art, qui emprunte aux méthodes quantitatives et qui revendique un traitement computationnel pourrait susciter de légitimes questionnements au sein d'une communauté dans laquelle le regard et l'approche visuelle sont au cœur de la pratique disciplinaire. Pourtant, il ne s'agit en aucun cas de « ne plus regarder les œuvres » mais d'outiller le regard, de l'enrichir de nouvelles focales. La lecture distante (*distant reading**), le traitement computationnel, l'approche statistique forment ici un point de départ et non un point d'arrivée. Ils permettent d'effectuer des observations, de relever des phénomènes, qu'il convient ensuite d'examiner, d'interroger, d'analyser, en faisant d'ailleurs appel le plus souvent aux outils « traditionnels » et éprouvés de la discipline, soit « l'œil et l'archive » (Passini 2017).

35. Regard proche et regard distant se nourrissent mutuellement, tout comme les instruments numériques ne sont qu'un enrichissement de la boîte à outils de l'histoire de l'art. S'il est beaucoup question, dans cet exposé, d'automatisation et d'obstacles au traitement computationnel, soulignons que certaines des difficultés rencontrées peuvent s'avérer profitables au travail de recherche. Le temps passé à critiquer les données, à les nettoyer, à les corriger est du temps investi à manipuler le corpus, à s'y confronter. C'est pourquoi ces tâches, certes chronophages et rébarbatives, ne doivent pas être systématiquement déléguées à des « petites mains » : c'est dans ce véritable travail de fourmi que se façonne une familiarité avec le corpus, qu'émergent de nouveaux questionnements et que se forgent des hypothèses.

Publier, partager, offrir à la réutilisation

36. Il serait dommage que ce corpus, patiemment constitué au cours des années de doctorat, dorme par la suite sur un disque dur. Aussi, je souhaite rendre publique ma base de données à l'issue de mon travail de recherche. Dans ce but, j'utilise le CMS* Omeka[21], logiciel *open source* conçu pour créer des bibliothèques numériques sur le Web.

21. Omeka est un *content management system* (CMS), c'est-à-dire un logiciel qui permet de créer et mettre à jour des sites web. Développé depuis 2008 par le Roy Rosenzweig Center for History and New Media, Omeka est *open source*. Il a été spécifiquement conçu pour mettre en ligne des bibliothèques numériques et des petits jeux de données. En France, il est notamment utilisé pour les bibliothèques patrimoniales de l'École nationale des Ponts et Chaussées (https://patrimoine.enpc.fr) et de l'École des Mines (https://patrimoine.mines-paristech.fr), mais aussi par l'Institut national d'histoire de l'art dans le cadre du projet *Digital Muret* (https://digitalmuret.inha.fr). Il est particulièrement adapté pour les corpus d'histoire de l'art, du fait de son interface graphique qui se prête bien à la présentation d'images. Sur le choix d'Omeka comme

37. Il s'agit d'abord de fournir une « annexe numérique » à la thèse qui permette aux examinateurs et aux lecteurs d'explorer le corpus, en profitant de fonctionnalités bien plus riches que celles offertes par un catalogue imprimé, nécessairement figé dans un ordre préétabli. Il sera par exemple possible d'effectuer une recherche multicritère dans le corpus ou encore de zoomer dans une image numérisée pour en discerner les détails. Il est possible d'aller plus loin et de fournir un « corpus outillé », c'est-à-dire de donner accès aux outils qui ont permis certaines visualisations cartographiques ou statistiques, et donc permettre au lecteur de « rejouer » certains calculs, d'en modifier les variables, afin de discuter des résultats obtenus, voire d'interroger le corpus à l'aune de ses propres préoccupations et objets.

38. Au-delà du cadre de ma recherche universitaire, la publication numérique du corpus est aussi pensée comme la mise à disposition d'un instrument de travail sur la vue d'optique, utile à ceux qui s'intéressent à ces artefacts. Fonctionnant comme un catalogue raisonné, il facilitera l'identification d'une vue d'optique, sa datation ou encore l'évaluation de sa rareté. Il s'agit ici, entre autres, de faciliter tout futur travail de catalogage de collections de vues d'optique. Il m'importe aussi – et surtout – de partager les résultats de mes travaux avec les institutions dont j'ai mobilisé les collections, afin qu'elles puissent en bénéficier.

CMS pour un corpus de thèse, voir mon carnet Hypothèses *Isidore & Ganesh* : https://ig.hypotheses.org/.

39. Au cours du traitement des données, de leur analyse et de leur enrichissement, j'ai en effet pu corriger ponctuellement une erreur de catalogage, compléter une transcription partielle, proposer une identification, une datation ou encore une attribution, etc. Ces informations intéressent l'institution, toujours encline à améliorer le catalogage de ses œuvres. C'est d'ailleurs l'un des arguments mis en avant dans les politiques de numérisation et d'ouverture des données patrimoniales : si la mise en ligne doit rendre accessibles à un large public les collections, elle doit aussi stimuler la recherche et favoriser la production de nouvelles connaissances. Plusieurs rapports en faveur de l'Open Data culturel[22] soulignent le potentiel offert par l'ouverture des données, qui permettrait la production de nouvelles connaissances, à leur tour réintégrables dans les systèmes d'information des institutions. *Crowdsourcing** et recherche scientifique se confondent ici dans un idéal de collaboration entre l'institution patrimoniale et ses publics (amateurs et chercheurs).

40. Ce vœu est-il cependant, d'un point de vue pragmatique, véritablement réalisable ? Dans le cas précis de ma recherche, il me semble que plusieurs obstacles existent. Le premier est technique : même si ma base de données disposera à terme d'une API, il faudra, pour les professionnels des musées, s'y adapter de la même manière que je l'ai fait avec celle fournie par leur institution. À leur tour, ils devront aligner mes référentiels sur les leurs. Autant d'opérations fort chronophages pour un gain

22. L'un des plus complets est celui de Domange (2013).

relativement faible à l'échelle de l'institution, sinon celle de la démonstration de faisabilité.

41. Comment, tout de même, porter à l'institution ces enrichissements ? Deux solutions peuvent être envisagées : la première est d'extraire de ma base, au format tableur, le jeu de données enrichies correspondant à chaque institution. Certes, les données ne seront, en l'état, pas réintégrables au sein du système d'information de la bibliothèque ou du musée, mais le fichier, intégré à leur documentation, pourra trouver des usages ponctuels. L'autre voie, plus ambitieuse, est celle du versement de ma base de données sur Wikidata : intégrées à l'écosystème global du Web sémantique*, les données produites dans le cadre de ma thèse seront facilement interrogeables et réintégrables selon des standards largement partagés par la communauté[23].

Comment mieux collaborer ?

42. L'objet de ce texte était d'illustrer, par un exemple concret, le potentiel offert à la fois par la mise en ligne et l'ouverture des données patrimoniales et par l'usage des technologies numériques dans la recherche en histoire de l'art. Le chemin est encore semé d'obstacles, notamment techniques, et je suis persuadée que c'est par la collaboration que nous les dépasserons. Aussi, j'aimerais, en conclusion de cet article, formuler trois vœux.

43. Le premier est que nous menions, collectivement et plus largement, une réflexion sur les usages possibles et autorisés des collections numérisées et des données mises en ligne. Encore trop souvent, seule la consultation « rapprochée » des documents, c'est-à-dire la lecture l'une après l'autre de chaque notice est envisagée. La lecture distante, le traitement computationnel sont en général rendus impossibles par l'absence de fonctionnalités permettant l'interrogation et le téléchargement en masse des métadonnées et images numérisées. Si l'implémentation de tels outils se révèle complexe et onéreuse, il est possible de mettre en place une solution intermédiaire : fournir des extractions CSV, indiquer clairement aux usagers que l'institution peut envoyer les données à ceux qui en font la demande. Encore trop souvent, y compris lorsque les solutions techniques existent, les démarches de lecture distante sont perçues avec méfiance par les agents (peur de la perte de contrôle des données), voire sont considérées comme non légitimes.

44. Pour faciliter l'exploitation des données produites par l'institution, il faut ensuite veiller à documenter les choix, les pratiques métiers, afin de donner à l'usager les outils pour évaluer la qualité des informations qu'il manipule. Cela passe notamment par la formation et le décloisonnement des métiers. Il est à mon sens essentiel de former les étudiants en histoire et histoire de l'art non seulement à l'usage des bibliothèques numériques et col-

23. Sur le Web sémantique et ses usages dans la recherche et les bibliothèques, voir les travaux de Gautier Poupeau, et notamment son blog *Les Petites Cases* : http://www.lespetitescases.net.

lections en ligne – ce qui est déjà le cas – mais aussi de leur donner à voir toute la chaîne de production, qui, de l'étagère des réserves à la publication sur Internet, leur permet de consulter en ligne les documents et données nécessaires à leurs recherches. Décloisonner, « donner à voir » : c'est ainsi que nous pouvons avoir des usagers véritablement informés et critiques face aux outils et données qu'ils mobilisent.

45. Enfin, continuer à favoriser l'interdisciplinarité, qui est le terreau fertile des humanités numériques. Si les réutilisations, les enrichissements de données via les API et l'exposition sur le Web de données sont souhaités par de nombreux acteurs, les réalisations concrètes de tels échanges sont encore rares. C'est pourquoi il est indispensable de les soutenir par des collaborations intéressées, afin de multiplier les preuves de concept qui ne pourront qu'insuffler une dynamique générale.

Du corpus archivistique au corpus numérique : les soubassements du Web sémantique. L'exemple des sources relatives au parlement de Flandre

Renaud Limelette

Introduction

1. Depuis une dizaine d'années le Centre d'histoire judiciaire (CHJ) s'est lancé dans la valorisation des fonds d'archives liés au parlement de Flandre. Si cette cour souveraine de justice, ancêtre de la Cour d'appel de Douai, a connu une existence relativement brève, de 1667 à 1790, ses archives offrent aux chercheurs de différentes disciplines des sources de questionnement multiples. Elles touchent de nombreux champs disciplinaires. Assurément, des recherches sur les normes applicables (législation royale, droit provincial, coutumes, etc.), sur les institutions (royales, seigneuriales, civiles et ecclésiastiques) ou les comportements sociaux (à travers particulièrement le contentieux pénal et les interrogatoires criminels) viennent immédiatement à l'esprit, mais ces archives permettent aussi d'approcher l'organisation économique de la province à travers les contentieux des communautés de métiers (privilèges économiques ou des styles des communautés).

2. Cette valorisation se décline aujourd'hui en deux projets connexes. Le plus ancien s'incarne autour de la constitution d'une base de données[1] regroupant les 30 000 dossiers de procédure civile et criminelle conservés dans la sous-série 8B1 des Archives départementales du Nord. Le développement informatique et technique a entièrement été réalisé au sein du CHJ en s'appuyant sur MySQL et le *framework* Yii[2].

3. Le second projet met l'accent, non plus sur des archives judiciaires, mais sur la littérature juridique éditée et relative au parlement de Flandre. En effet, plusieurs imprimeurs-libraires exerçant dans le ressort du parlement de Flandre, trouvèrent dans la publication de recueils d'arrêts un débouché commercial. Les Danel, Henry, Lehoucq et Mairesse proposèrent des éditions des arrêts retenus, et parfois commentés, par des magistrats du parlement de Flandre. À côté de ces recueils d'arrêts, la compagnie des magistrats était aussi chargée d'enregistrer les édits

1. Pour accéder à la base et y lancer une recherche : http://parleflandre.univ-lille2.fr/index.php/recherche.
2. Yii est un cadre d'application (ou infrastructure logicielle) destiné au développement d'application web. Il s'appuie sur un modèle-vue-contrôleur. Le modèle est la base de données, la vue est la représentation graphique (visuelle) de l'utilisateur de la base et le contrôleur fait respecter la concordance entre le modèle et la vue lorsque l'utilisateur agit sur les données (création, suppression, modification, recherche, etc.).

royaux et de prendre des décisions solennelles, appelées « arrêts de règlement », en dehors de tout contentieux. Cette activité législative a elle aussi été publiée, notamment par Derbaix et Willerval. Pour le chercheur, toute cette littérature juridique (Cazals 2018) est une source d'information précieuse. L'exposition de toutes ces sources se fait à travers les outils proposés par la TGIR Huma-Num[3], Nakala[4] et Nakalona.

4. Malgré tous les efforts menés par les ingénieurs du CHJ, aucun chercheur, quelle que soit sa discipline, ne peut lire l'intégralité des sources diffusées, pour tirer de celles-ci celles qui serviront à sa propre recherche. Pour lever cet obstacle, à savoir la sélection de données pertinente, l'informatique offre des moyens de structuration des données pour faciliter leur échange et leur compréhension entre machines. Nous proposons d'expliquer les choix opérés dans les deux projets autour de la granularité des données et des métadonnées.

La granularité des données ou comment exploiter la masse documentaire

5. Concernant la granularité des données[5], on remarquera que les deux projets touchent au même objet de recherche, le parlement de Flandre, mais la nature des sources utilisées est différente. Or c'est la nature des sources qui conditionne la manière dont elles vont être exploitées ou valorisées par les méthodologies propres aux humanités numériques.

6. Notons ici que cette démarche, qui part des sources, est inductive et non déductive. On commence par des sources pour aboutir à un moyen de les valoriser, et non d'une idée de valorisation pour y soumettre des sources[6]. Cette démarche ne s'entend que si l'on veut, à travers les humanités numériques, expliquer, c'est-à-dire déplier, les composantes d'un corpus, pour en présenter le contenu. Au contraire, si l'on veut d'abord utiliser un média numérique bien précis, pour valoriser un corpus, on choisira une sélection des contenus du corpus qui coïncide assez bien avec le média, mais qui ne reflète pas tous les aspects du corpus. Par exemple, si l'on choisit d'utiliser un film documentaire pour présenter le fonds du parlement de Flandre, on choisira des affaires judiciaires qui ont un potentiel visuel assez fort, et on laissera de côté l'énorme masse de la banalité des affaires.

7. Cette digression posant le sens de déductif par rapport à inductif, revenons au choix opéré de manière inductive pour le parlement de Flandre et à la granularité des données. Nous verrons que c'est la nature du corpus qui conditionne la solution numérique.

3. Cf. https://www.huma-num.fr
4. Cf. https://documentation.huma-num.fr/nakala/#utilisation-de-nakala
5. Toute cette partie repose sur la notion de granularité. La granularité s'entend par le contenu du corpus mis en avant et la manière dont le corpus est sectionné en unités logiques, voir par exemple (Delmotte 2009).
6. Pour une plus longue explication de cette démarche et une mise en pratique, on se reportera à (Limelette 2020).

30 000 dossiers de procédure ou 450 mètres linéaires d'archives manuscrites : le projet *ParleFlandre*

8. Le premier corpus concernant le parlement de Flandre est la sous-série 8B1 des Archives départementales du Nord. Matériellement, cette sous-série contient près de 30 000 dossiers et s'étale sur 450 mètres linéaires. Y sont conservés tous les dossiers de procédures de 1667 à 1790.

9. Pour mettre en valeur ces dossiers, le Centre d'histoire judiciaire (UMR 8025) déposa un projet à l'Agence nationale de la recherche sous l'acronyme « Parle-Flandre ». Le but du projet était de présenter au public ces 30 000 dossiers de procédure. Devant ce genre de projet, où la masse documentaire est importante, il faut commencer par analyser un échantillon des dossiers pour concevoir leur exploitation numérique. Si aux Archives départementales du Nord, il existe encore aujourd'hui un inventaire des 30 000 dossiers sous forme de fiches analytiques, ces dernières sont tellement lacunaires qu'elles restent quasiment inexploitables. On y trouve seulement, quand elles sont bien remplies, un ou deux noms de famille ou d'institution, une date et un mot-clé.

10. L'échantillon a vite révélé une diversité de données tant sur leur forme que sur leur nature. Sur la forme, les dossiers sont d'une épaisseur inégale, certains ne contiennent que quelques pièces d'autres sont un enchevêtrement touffu de papiers de dimensions irrégulières, parfois reliés par une simple cordelette. La nature des pièces contenues dans les dossiers présente une grande variété qui déconcerte au premier abord. Il peut s'agir de plans figuratifs, de contrats de toute sorte, de livres de compte et bien entendu d'actes rédigés par le personnel du parlement, le tout sans logique apparente.

11. Néanmoins, un dossier type est apparu. Si celui-ci était constitué de nombreuses pièces, elles se présentaient toujours de la même manière. Ainsi, le dossier type contient un inventaire des pièces (figure 1) produites par le procureur des parties. Cet acte est adressé au conseiller-commissaire (Limelette 2018) du parlement ou son équivalent dans une juridiction inférieure, ce qui permet d'identifier rapidement le niveau hiérarchique juridictionnel. Le procureur prend soin de coter d'une lettre, A pour la première, B pour la deuxième, et ainsi de suite, toutes les pièces transmises. Chaque pièce de l'inventaire est nommée selon le vocabulaire procédural[7] et judiciaire de l'époque. On trouve donc souvent une « commission de procureur », c'est-à-dire l'acte par lequel une partie choisit son procureur et lui donne mandat d'agir pour elle. Puis, dans l'ordre procédural et selon que la partie soit demanderesse ou défenderesse, on trouve la requête introductive, un placet[8], des contredits, des répliques et enfin des dupliques.

7. Pour l'analyse de cette procédure, voir (Limelette 2018, sect. 2.2.1).
8. Acte par lequel le requérant demandait au conseiller-commissaire de fixer un jour de comparution.

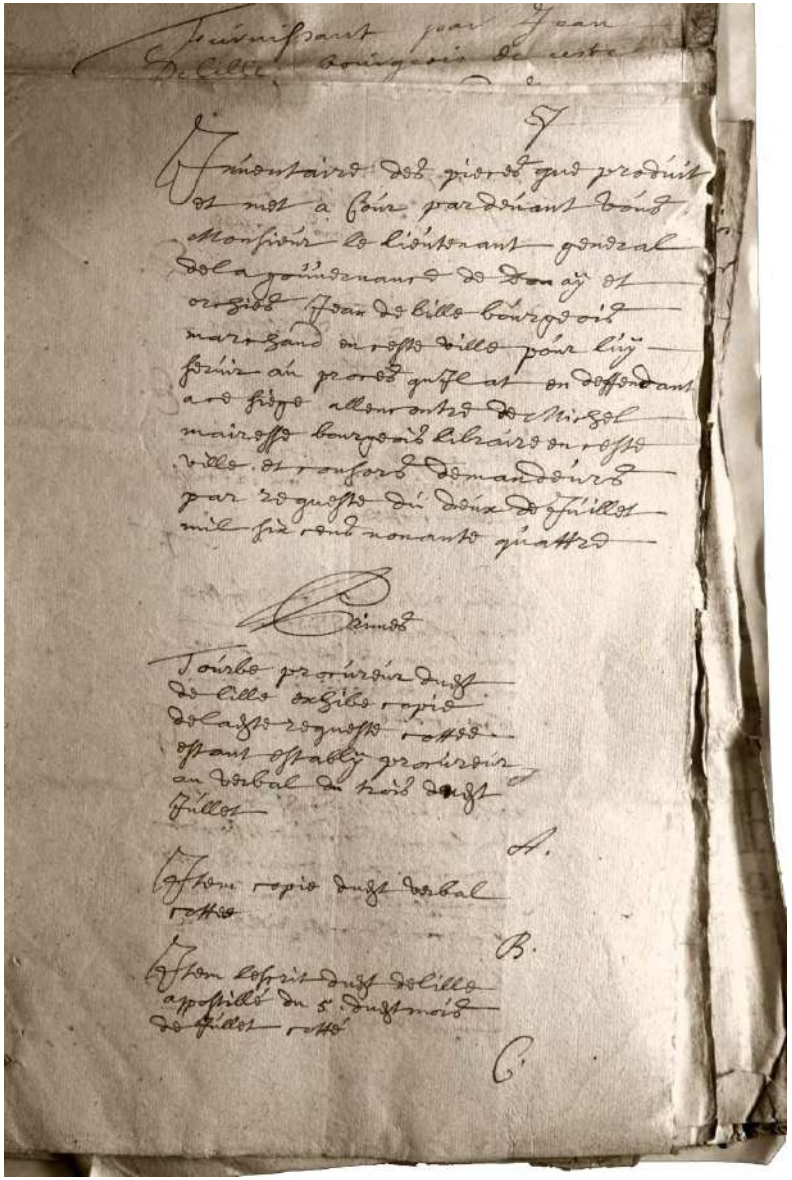


Figure 1. Extrait de l'inventaire du dossier 8B1/1846 des Archives départementales du Nord

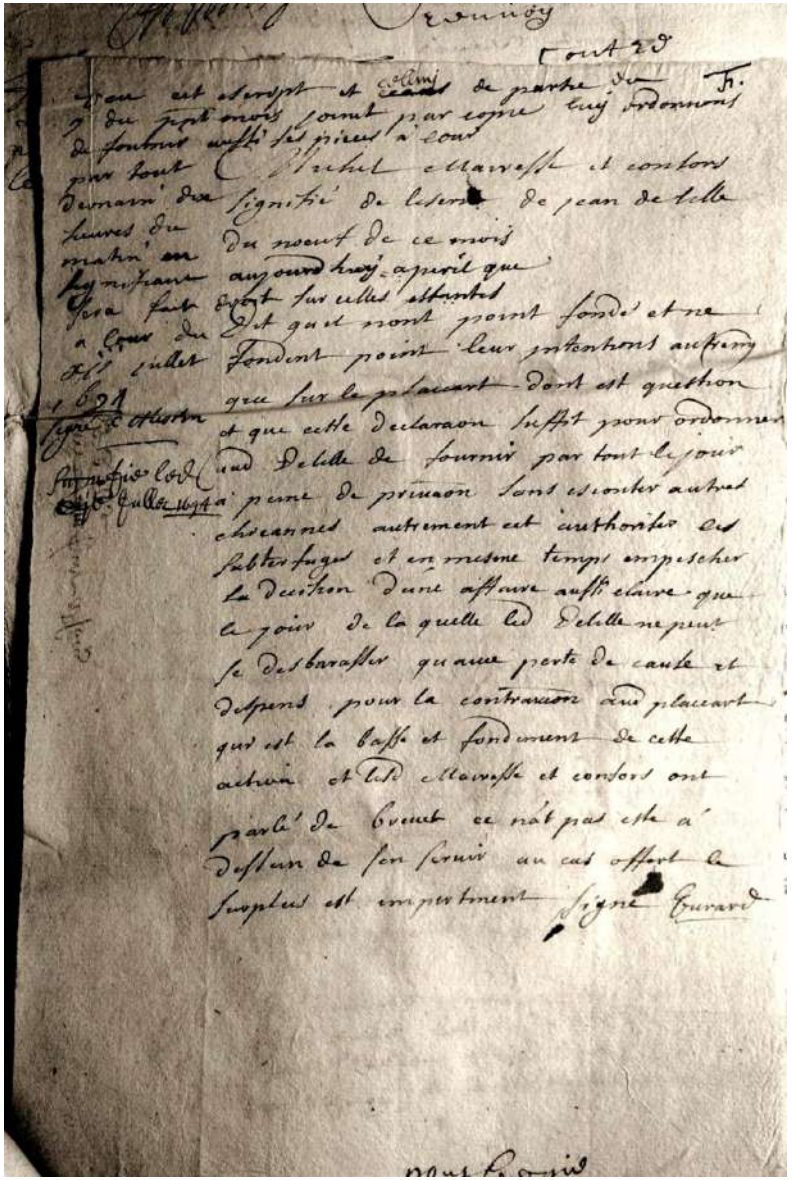


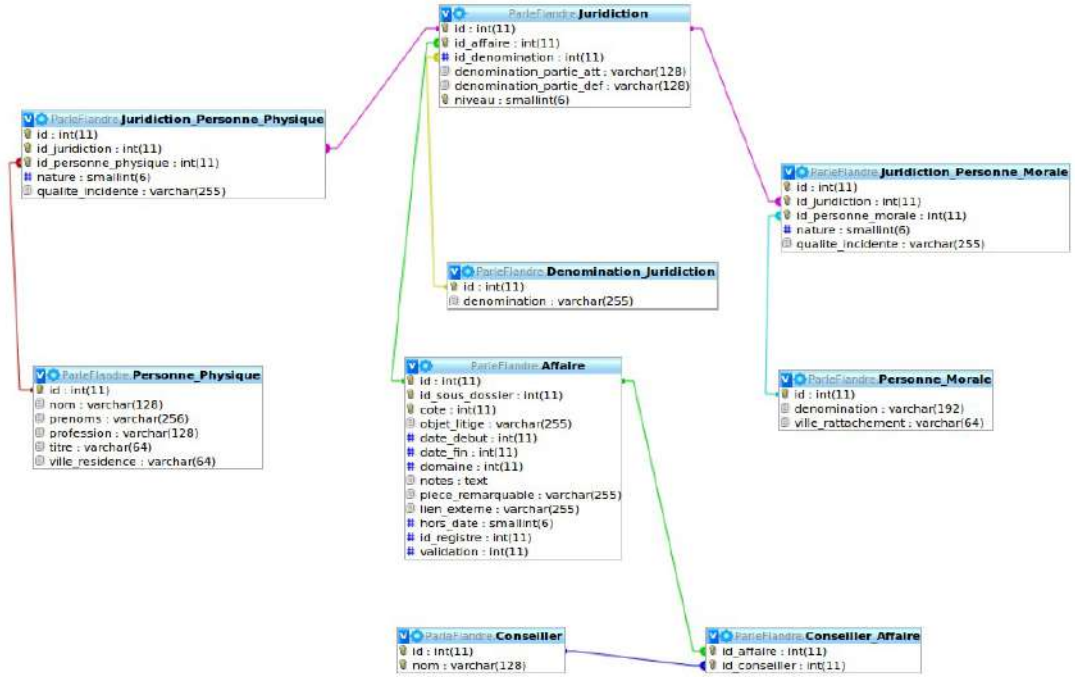
Figure 2. Exemple d'apostille en marge d'un acte, dossier 8B1/1846 des Archives départementales du Nord

12. À côté de ces pièces procédurales, on trouve également bien souvent des pièces probatoires, servant d'appui à l'argumentation des parties pour soutenir juridiquement leurs dires. Là, la nature des pièces est très variable, elle dépend du type de contentieux. C'est ici qu'on trouve, sans être exhaustif, des testaments, des contrats de toutes espèces ou des registres d'imposition. Bref tout ce qui traduit la vie ordinaire aux XVII^e^ et XVIII^e^ siècles.

13. En marge de ce dossier type, d'autres dossiers sont beaucoup moins structurés. Soit ils ne renferment qu'une à plusieurs pièces, parmi celles précédemment évoquées, et on peut alors les rattacher à un autre dossier mieux structuré, soit les pièces sont éparses et parfois forts anciennes, si bien qu'il est difficile d'identifier dans quel autre dossier elles s'insèrent.

14. Pour être complet, on ajoutera que parfois une pièce peut en contenir une autre intrinsèquement. C'est le cas des apostilles (figure 2), qui sont la marque, et en même temps un acte procédural distinct, que la cour a pris en compte la pièce fournie. Par exemple, le conseiller peut écrire en marge « Reçu le [...] transmis le », où un greffier atteste du paiement des frais de procédure (Fréger 2009).

15. Ainsi donc, c'est un écheveau de pièces plus ou moins structurées, et dont la nature est plus ou moins disparate, qu'il a fallu présenter numériquement. C'est le but du projet *ParleFlandre*, sous-titré « constitution d'un outil adapté à une exploitation scientifique » (Limelette et Michel 2014). Très vite la solution qui s'est dégagée fut la constitution d'une base de données, dont il fallait dresser le schéma (figure 3).

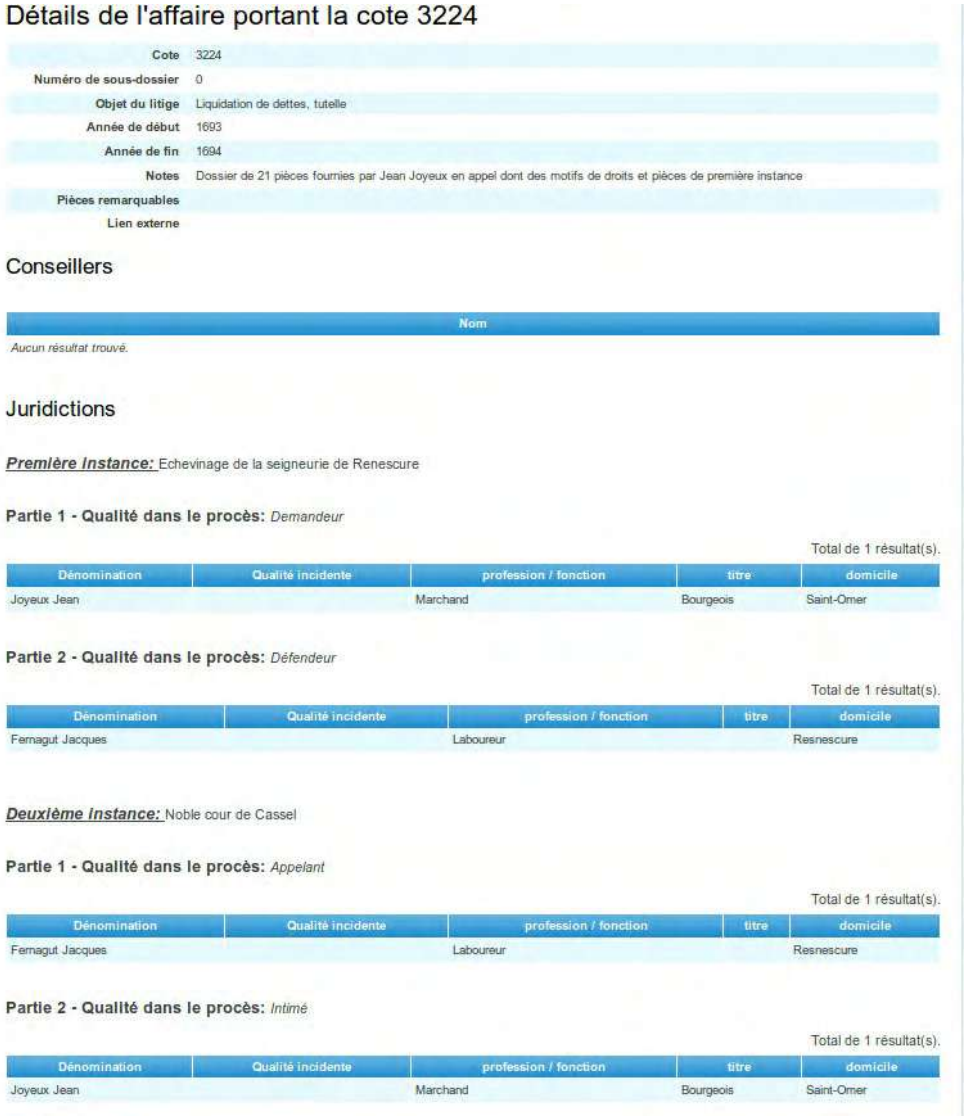


Graphe des dépendances réalisé par François Périchon, IR au CRI de l'Université Lille 2, avec le concours de Mehdi Aksil, AI contractuel au CRI, et Renaud Limelette, IE en analyse de sources au CHJ.

Figure 3. Graphe des dépendances – Base de données ParleFlandre
Crédit : Renaud Limelette

16. Celui-ci, outre ses aspects informatiques, traduit la granularité qui a été retenue. Pour des raisons de temps de dépouillement, d'analyse et de mise en forme numérique, le parti a été pris de concevoir les tables et les champs de la base à partir d'un dossier type, en laissant de côté les apostilles mais en conservant la possibilité d'insérer dans la base les dossiers moins structurés (figure 4) qu'un dossier type.

17. Reste sur ce premier projet à dresser deux remarques conclusives. Un financement par L'Agence nationale de la recherche ne dure que quelques années. Or, pour ce projet, la seule analyse de l'échantillon et la création de la base de données ont épuisé ces maigres années. Pourtant un nombre conséquent de personnel[9] a été mobilisé, mais l'exploitation d'un tel corpus dépasse l'entendement, on comprend mieux pourquoi ces archives ont dormi pendant une centaine d'années (Cleyet-Michaud 2009 ; Demars-Sion 2014). La dernière remarque porte sur la nature des pièces et l'objet de la base de données. La base de données n'expose pas les arrêts prononcés par le parlement de Flandre, car les pièces conservées dans la sous-série 8B1 sont antérieures au prononcé de l'arrêt, et ces arrêts sont conservés dans la sous-série 8B2.



Détails de l'affaire portant la cote 3224

Cote	3224
Numéro de sous-dossier	0
Objet du litige	Liquidation de dettes, tutelle
Année de début	1693
Année de fin	1694
Notes	Dossier de 21 pièces fournies par Jean Joyeux en appel dont des motifs de droits et pièces de première instance
Pièces remarquables	
Lien externe	

Conseillers

Nom
Aucun résultat trouvé.

Juridictions

Première Instance: Echevinage de la seigneurie de Renescure

Partie 1 - Qualité dans le procès: *Demandeur*

Total de 1 résultat(s).

Dénomination	Qualité incidente	profession / fonction	titre	domicile
Joyeux Jean		Marchand	Bourgeois	Saint-Omer

Partie 2 - Qualité dans le procès: *Défendeur*

Total de 1 résultat(s).

Dénomination	Qualité incidente	profession / fonction	titre	domicile
Fernagut Jacques		Laboureur		Resnescure

Deuxième Instance: Noble cour de Cassel

Partie 1 - Qualité dans le procès: *Appelant*

Total de 1 résultat(s).

Dénomination	Qualité incidente	profession / fonction	titre	domicile
Fernagut Jacques		Laboureur		Resnescure

Partie 2 - Qualité dans le procès: *Intimé*

Total de 1 résultat(s).

Dénomination	Qualité incidente	profession / fonction	titre	domicile
Joyeux Jean		Marchand	Bourgeois	Saint-Omer

Figure 4. Exemple de vue d'une affaire de la base de données ParleFlandre, 8B1/3224
Crédit : Renaud Limelette

9. Sur la genèse du projet et les difficultés rencontrées par le personnel affecté, voir (Baudens et Jeannin 2009).

De plusieurs dizaines de livres papier à plusieurs milliers de ressources numériques : le projet *Pandectes des Flandres*

18. Le second projet reposant sur le parlement de Flandre s'intitule « Pandectes des Flandres[10] » (anciennement « Meta-parleFlandre »). Il touche à l'exploitation de la littérature juridique imprimée et produite par les magistrats du parlement. Cette littérature juridique se décline en divers recueils. On pense de suite, pour les habitués du monde judiciaire moderne aux recueils d'arrêts[11]. À côté de ceux-ci, il est aussi tous les recueils sur l'enregistrement des édits royaux et tous les arrêts de règlement produits par le parlement de Flandre.

19. Par rapport au projet *ParleFlandre*, dans le projet *Pandectes des Flandres* les sources mises en valeur sont livresques. Ainsi la première étape de ce projet a été d'identifier les ouvrages afin de les numériser. Or en 2009, la Bibliothèque nationale de France menait ses deuxièmes journées professionnelles sur la numérisation et la valorisation concertées en sciences juridiques. Elle proposait, pour compléter ses propres collections, de financer une part de la numérisation de fonds locaux. Ainsi à l'époque, la bibliothèque numérique scientifique de la réserve patrimoniale des ex-universités de Lille, nommée Pôlib, avait obtenu le financement de la numérisation de 43 ouvrages conservés dans ses fonds et ceux de l'Institut catholique de Lille. Parmi ces ouvrages repérés et numérisés, de nombreux ouvrages touchent au parlement de Flandre et forment le corpus du projet *Pandectes des Flandres*[12]. Une phase essentielle dans ce travail de numérisation fut de coupler la numérisation en mode image avec la possibilité d'interroger le texte de chaque document. Pour ce faire l'océrisation (OCR*) a été complétée par une correction manuelle pour obtenir une version texte fiable à 99 %. En effet, les graphies particulières des imprimeurs de l'époque sont mal prises en compte par une simple océrisation, et si l'on veut un grain assez fin pour rechercher dans ce corpus d'ouvrages il faut passer du temps à reprendre chaque mot pour avoir une version texte fiable. Présenté d'abord sur Pôlib, ce corpus sur les sources du droit en Flandre est dorénavant accessible sur NordNum[13] et sur Gallica.

20. Si la numérisation des ouvrages était une étape essentielle, elle n'en était pas l'aboutissement. En effet, navi-

10. Cf. https://parleflandre.nakalona.fr
11. On lira à propos sur ces recueils d'arrêt (Cazals 2018).
12. Willerval, Jacques François, éd. 1730. *Recueil des édits déclarations arrests et réglemens qui sont propres et particuliers aux Provinces du ressort du Parlement de Flandres*. Douai, France. https://nordnum.univ-lille.fr/ark:/72505/a0115224001366EQToQ/from/a011522400136C4fuAQ. Derbaix, Jean-Pierre-Joseph, éd. 1785. *Recueil des édits, déclarations, lettres patentes, etc. enregistrés au Parlement de Flandres*. 12 vol. Douai, France. https://nordnum.univ-lille.fr/ark:/72505/a011522400136C4fuAQ. Pollet, Jacques. 1716. *Arrests du parlement de Flandre sur diverses questions de droit, de coutume, et de pratique*. Édité par Liévin Danel. Lille, France. http://nordnum.univ-lille.fr/ark:/72505/a0115217019356v106i/from/a011522400136C4fuAQ. Henry, Jean-Baptiste-Joseph, éd. 1773. *Recueil d'Arrêts du parlement de Flandres*. 2 vol. Lille, France. Pillot, Louis. 1849. *Histoire du Parlement de Flandres*. Édité par Adam D'Aubers. 2 vol. Douai, France. https://nordnum.univ-lille.fr/ark:/72505/a011544092129DYRL7H.
13. L'intégralité du projet négocié avec la BNF est accessible sous l'URI suivant : http://nordnum.univ-lille.fr/ark:/72505/a011521028444a7MYpO/from/a011522400136C-4fuAQ.

guer à l'intérieur d'un recueil d'arrêts ou d'édits de plusieurs centaines de pages n'est pas chose aisée, même si le site de consultation dispose d'un bon moteur de recherche. En fait, l'intérêt de Pandectes des Flandres est de faciliter la recherche pour le lecteur. Au lieu de consulter un ouvrage dans son entièreté, l'ouvrage est divisé en unités logiques selon la nature des données. Ainsi on trouve sur Pandectes des Flandres des arrêts rendus par le parlement de Flandre, des édits enregistrés au parlement de Flandre et des arrêts de règlement pris par le parlement de Flandre. En d'autres termes, la numérisation des ouvrages avait abouti à constituer des grains numériques trop gros, il a fallu moudre ce grain pour séparer chaque arrêt, chaque édit et chaque règlement de l'ouvrage initial dans lequel ils étaient rassemblés. Profitant de la numérisation de chaque ouvrage en PDF, le projet *Pandectes des Flandres* se construit autour de la dissociation des pages des PDF pour les réassembler en unités juridiques ou judiciaires[14]. Une fois le lourd travail de définition de la granularité des données terminé, une autre tâche est à aborder : celle de la conception de la diffusion des données.

Les métadonnées : enjeu de diffusion

21. L'ancienneté relative du projet *ParleFlandre* (2007) n'a pas poussé les porteurs de ce projet ANR à se questionner sur la nécessité de mettre en place des métadonnées, si bien qu'à ce jour aucune métadonnée n'est associée aux affaires judiciaires alimentant la base de données. À l'inverse, le projet *Pandectes des Flandres* a de suite pris en compte cette problématique. S'il y a encore quelques années les solutions trouvées pour les projets en sciences humaines sociales étaient souvent particulières à chaque projet, aujourd'hui depuis le déploiement de la très grande infrastructure de recherche (TGIR) Huma-Num les solutions sont plus standardisées.

22. Si la grille de service a permis de solutionner les questions relatives au stockage, à la diffusion, au signalement et à l'exposition des données issues, au centre de toutes ces problématiques il est un dénominateur commun : la description des données selon un référentiel de métadonnées. Pour Pandectes des Flandres nous verrons qu'à chaque donnée sont associées des métadonnées au format Dublin Core*, et que ces métadonnées sont au cœur de la synergie du projet.

Le vocabulaire Dublin Core et les données de Pandectes des Flandres

23. Comme je l'ai déjà énoncé précédemment, quatre natures de données (figure 5) alimentent le site Pandectes des Flandres : des fiches biographiques des magistrats, des arrêts de règlement produits par le parlement de Flandre, des édits ou autres normes enregistrés au parlement, et de la jurisprudence commentée.

14. C'est-à-dire que maintenant il y a autant de PDF que d'arrêt, d'édits et de règlements.

Figure 5. Exemple de représentation des quatre natures de données dans Pandectes des Flandres

Crédit : Renaud Limelette

24. Pour un humain, un lecteur commun du site, il est presque instinctif de reconnaître la nature de chaque document : une lecture rapide, même sans grande attention, des caractères gras suffit souvent pour identifier un document. Mais dès que l'on commence à vouloir affiner l'analyse, l'humain n'est plus capable d'embrasser d'un seul coup toutes les informations, et plus il y a de données plus l'affaire se complique.

25. C'est ici que le passage au numérique prend tout son intérêt. Encore faut-il ne pas se limiter à numériser les ouvrages anciens accueillant les fiches biographiques, les arrêts de règlement, les édits et la jurisprudence de la cour. En plus de l'océrisation et de la division des ouvrages en autant d'unités logiques qu'il y a d'arrêts, d'édits ou de magistrats, la description précise de chaque unité logique par un référentiel de métadonnées* permet à la machine d'interpréter très rapidement chaque élément[15].

26. Ainsi dans le projet *Pandectes des Flandres* chaque objet numérisé est décrit selon le vocabulaire Dublin Core. Parmi les éléments et qualificatifs du vocabulaire Dublin Core, nous avons retenu le schéma suivant : titre, sujet, type de ressource, source, couverture, créateur, contributeur, identifiant et date, avec les qualificatifs date de création et date de parution.

27. Arrêtons-nous maintenant aux solutions élaborées par ordre de difficulté :

1. Pour le « titre », aucune difficulté n'a été rencontrée dans la mesure où chaque donnée avait déjà un titre dans sa version bibliographique[16].

2. L'URI (*uniform resource identifier*) de consultation de l'ouvrage d'où est tirée la ressource sur la plateforme NordNum a servi à remplir l'élément source[17].

15. Sur l'intérêt des métadonnées et des vocabulaires contrôlés, voir également (Broughton 2010).

16. Cf. http://purl.org/dc/terms/title

17. La Dublin Core Metadata Initiative (ci-après DCMI) recommande d'identifier la ressource au moyen d'une chaîne de caractères conforme à un système d'identification formel (voir http://purl.org/dc/elements/1.1/source). Or NordNum remplit cette recommandation en utilisant pour chaque ressource des URI de type ARK : voir http://

3. Comme toutes les données sont tirées de livres publiés, l'élément « type de ressource » n'a pas donné lieu à discussion[18].

4. Nous avons choisi d'utiliser l'élément « contributeur » pour y mettre le nom du laboratoire porteur du projet, sous la forme[19] « Centre d'Histoire Judiciaire (CHJ) – UMR 8025 » ; le but est de bien dissocier dans la vie de la donnée les créateurs originaux de la version bibliographique des créateurs secondaires qui en font une édition numérique.

5. La distinction de l'élément « date » par les raffinements « date de création » et « date de parution » permet de différencier la date originelle de la ressource avec la date de mise en ligne[20].

6. Si l'élément « couverture » pouvait poser quelques problèmes, compte tenu de la variabilité du ressort du parlement de Flandre (Limelette 2009), la difficulté a été vaincue en retenant simplement « Parlement de Flandre » comme le permet la Dublin Core Metadata Initiative[21].

7. L'élément « créateur » a pu sembler facile à utiliser puisque chacun des ouvrages d'où proviennent les données est identifiable par des auteurs. Pour autant la manière de référencer les auteurs restait incertaine, la Dublin Core Metadata Initiative laissant beaucoup de place à l'interprétation[22]. Pour identifier chaque auteur nous avons retenu la dénomination retenue par l'International Standard Name Identifier[23] que suit la Bibliothèque nationale de France dans ses notices d'autorité[24]. Ainsi chaque auteur est identifié par son nom, son prénom et ses dates de vie et de mort entre parenthèses.

8. L'un des éléments qui a demandé le plus de réflexion est l'élément « sujet ». Instinctivement on comprend que sous l'élément « sujet » se trouvent des mots-clés[25]. Or, on conviendra que, si on laisse à plusieurs spécialistes de la même discipline le soin de définir par des mots-clés les pans de leur discipline, on ne trouvera pas au final une liste de mots-clés arrêtée[26]. Au contraire, chaque spécialiste ayant sa conception et sa représentation de

n2t.net/e/ark_ids.html. Pour une présentation succincte des URI et du format ARK, voir les entrées correspondantes dans le glossaire.

18. Nous avons retenu à chaque fois l'étiquette « Text » comme le recommande la DCMI à propos des livres (voir http://purl.org/dc/dcmitype/Text).

19. La DCMI reconnaissant qu'il faut entendre comme contributeur « l'entité chargée d'apporter des contributions à la ressource » (voir http://purl.org/dc/terms/contributor).

20. Cf. http://purl.org/dc/terms/date

21. Cf. http://purl.org/dc/terms/coverage

22. Cf. http://purl.org/dc/elements/1.1/creator

23. Cf. http://www.isni.org

24. Cf. https://catalogue.bnf.fr/recherche-autorite.do

25. Cf. http://purl.org/dc/terms/subject

26. Nous reprenons à notre compte la réflexion de Michel Villey (1995, 176-177) sur le sens des mots dans ses *Carnets* : « Nous voici placés devant le monde des abominables mots qui poursuivent à travers le temps leur tourbillon désordonné. La plupart de nos contemporains nagent à leur aise dans cette extrême confusion des mots, dont le sens n'est plus définissable. Je ne puis construire sur cette glaise, et sens le besoin de restituer aux mots leur sens qui est objectif, assis sur l'étymologie qui est le témoin de leur formation spontanée conforme à l'ordre naturel. »

sa discipline, on tombera inévitablement dans une cacophonie littérale, une logorrhée plus ou moins subjective.

28. Prenons à dessein un exemple juridique, que chacun comprendra aisément. Si on vous demande de décrire par un seul mot-clé la situation pénale suivante, « M. X, après mûre réflexion, a tué sa mère », il y a de grandes chances que les réponses fournies soient très différentes les unes des autres. J'aurais sans doute : meurtre, homicide, assassinat, homicide volontaire, crime, matricide. C'est la même chose avec les visiteurs d'un site internet, quand ils utilisent un moteur de recherche ils n'ont pas tous le même référentiel de mots, et donc faute d'avoir rendu public ce référentiel de mots-clés, les résultats proposés seront incertains ou incomplets. Il en est de même pour une machine qui doit interpréter et classer des données du web.

29. À travers cet élément « sujet » c'est tout le Web sémantique* qui transparaît. C'est pourquoi les recommandations de la DCMI prennent soin d'insister sur la nécessité d'utiliser un vocabulaire contrôlé. Reste ici à préciser qu'un vocabulaire contrôlé ne se développe pas seul, au contraire il doit être mis en place par des spécialistes chargés de choisir les termes descripteurs[27] les plus représentatifs d'un domaine. Il fallait donc pour le projet *Pandectes des Flandres* trouver un vocabulaire contrôlé le plus adéquat. Après de nombreuses recherches, nous avons retenu celui proposé par le ministère de la Culture pour décrire et indexer les archives historiques locales[28].

9. Enfin vient l'élément « identifiant » qui permet d'identifier la donnée. La DCMI recommande ici[29] d'utiliser une chaîne de caractère reconnue dans un système d'identification formel. Pour Pandectes des Flandres l'identifiant retenu est la fin d'un URI de type Handle[30]. L'URI provient du réservoir de données de Pandectes des Flandres qu'est Nakala. En effet chaque donnée enregistrée dans Nakala se voit attribuer une référence unique et pérenne. L'intérêt est de pouvoir consulter la donnée sans aller dans les pages publiques de Nakala ou de l'exposition du projet. Ainsi chaque donnée de Pandectes des Flandres a un URI construit de cette façon : « http://www.nakala.fr/data/identifiant de la donnée » et « http://hdl.handle.net/identifiant de la donnée ». Ces deux URL permettent d'accéder directement à la donnée, c'est-à-dire au PDF décrit à travers le vocabulaire Dublin Core.

30. On peut se demander quel est le but de toutes ses métadonnées. L'un des avantages du Dublin Core est de rendre le corpus de données interopérable. De plus, le moteur de recherche du site, où sont exposées publiquement et éditorialisées les données[31], s'appuie sur

27. Pour une approche synthétique on se reportera à (Coustaty *et al.* 2012).

28. Cf. https://francearchives.fr/article/37828#

29. Cf. http://purl.org/dc/elements/1.1/identifier

30. Cf. https://www.handle.net/index.html

31. La TGIR Huma-Num permet de coupler le réservoir de données Nakala à un système de gestion de contenu appelé Omeka (appelé aussi Nakalona quand Omeka est « bridé » par Huma-Num).

les champs du Dublin Core pour cibler précisément les résultats. En effet, l'effort investi par l'utilisation de vocabulaires contrôlés donne ici toute son efficacité. La recherche avancée sur les contenus enregistrés dans Omeka permet de lancer une requête à travers les champs du Dublin Core retenus en configurant la recherche avec des opérateurs de type : « contient », « ne contient pas », « est exactement », « est vide », « n'est pas vide », « commence par » ou « fini par ». On peut bien entendu croiser la recherche avec un autre champ du Dublin Core en appliquant un opérateur booléen de type « et » ou « ou »[32]. C'est là un des gros avantages des métadonnées quand leur schéma est bien conçu[33].

La synergie du projet *Pandectes des Flandres* et le Web sémantique

31. Quand on construit un projet numérique[34], outre les aspects de recherche scientifique attachés au projet, on désire que celui-ci soit visible sur la toile, sans avoir à fouiller le tréfonds des pages de résultats de son moteur de recherche. Ainsi pour remonter dans les pages de résultats, il fallait, il y a encore quelques années, multiplier les liens hypertextes pointant sur son propre site.

32. Aujourd'hui, on conviendra que dans nos disciplines scientifiques ce n'est pas le PageRank[35] qui devrait compter, mais la qualité scientifique du site. Ainsi, dès lors que le PageRank n'apporte qu'une vue quantitative des personnes visitant un site il ne devrait pas être un motif d'évaluation d'un projet scientifique. Autrement dit, ce n'est pas le nombre de visiteurs qui compte mais la motivation qui les a poussés à visiter le site[36].

33. Pour faire valoir la qualité scientifique de son projet numérique, encore faut-il avoir les outils adéquats. Dès l'initiation du projet *Pandectes des Flandres*, l'interopérabilité des données a été prise en compte. Pour cette raison, nous avons opté pour le pack Nakalona proposé par la TGIR Huma-Num. Le système de gestion de documents numériques Omeka permet de nous concentrer sur l'édition des données. Déjà sous Omeka les données peuvent être affichées sous des formats facilitant la compréhension du corpus par une machine et l'indexation dans les moteurs de recherche[37]. En effet, Omeka propose des sorties dans différents formats[38] dont on mesurera l'intérêt.

32. Cf. https://omeka.org/classic/docs/GettingStarted/Searching

33. À l'inverse, à défaut de se conformer à au moins un vocabulaire contrôlé, une recherche dans la base de données ParleFlandre renvoie des résultats partiellement inexploitables si on n'y prend pas garde. Par exemple, la requête « dette » dans le champ « Objet du litige » agrège tous les objets du litige où il a été saisi soit « dette » soit « dettes », alors que la requête « dettes » ne renvoie que les objets du litige où seul « dettes » a été saisi.

34. Un guide pratique accompagne les chercheurs et les ingénieurs dans les étapes à suivre pour mener à bien un projet numérique : voir (Ingarao et Saïdi 2011).

35. Sur le rapport entre le PageRank et le référencement des moteurs de recherche, voir (Andro, Chaigne et Smith 2012, sect. 2.2).

36. On lira ici avantageusement (Cardon 2013).

37. Bien qu'Omeka s'appuie essentiellement sur le Dublin Core, un article récent montre l'écart qu'il peut y avoir dans la conception du Dublin Core entre créateurs de la norme et utilisateurs de celle-ci, voir (Maron et Feinberg 2018).

38. C'est grâce à la fonction ContextSwitch du *framework* Zend Framework, sur lequel s'appuie Omeka, qu'une ressource peut avoir plusieurs représentations. Voir (Pauli et Ponçon 2008).

34. Sans être du Web sémantique, JSON (figure 6) est bien connu comme format d'échange de données (Sobréro 2014). Dans le cas ci-dessous, rien n'est très explicite sans savoir que l'« id 707 » est l'item intitulé « Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune ». Pour plus d'expressivité et dans une approche de Web sémantique on préférerait une exposition en JSON*-LD (Lanthaler et Gütl 2012), le nouveau standard du W3C[39].

Figure 6. Exemple de données au format JSON exposées dans Omeka, relatif à un « Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune »
https://parleflandre.nakalona.fr/items/show/707?output=json

35. On reconnaît parfaitement les éléments du Dublin Core (figure 7) repérables à travers les balises forgées sur <dcterms : élément-Dublin-Core></dcterms : élément-Dublin-Core>. En d'autres termes, c'est à la fois compréhensible pour un humain, déjà initié au Dublin Core, et facilement interprétable par une machine[40].

36. Dans l'exemple ci-après (figure 8), le format est plus explicite, car il détaille pour chaque terme du Dublin Core sa définition entre les balises <description></description>. C'est plus compréhensible pour un humain, même non initié au Dublin Core, mais trop volubile pour une machine.

37. Ainsi donc, Omeka fournit des formats de sortie qui selon leur fonction auront un intérêt spécifique. Par exemple, la sortie en XML diffère de la sortie RDF*. La première met en avant la structure des données alors que la seconde s'attache aux liaisons logiques des données (Poupeau 2010).

38. Grâce aux extensions « Import » et « Export », Omeka et Nakala sont reliés entre eux. Ainsi, avec Nakala, les métadonnées sont présentées publiquement sous un modèle RDF avec une sortie Turtle[41] (figure 9) ou RDF/XML[42] (figure 10).

39. Toutes ces vues ou expositions de formats différents nous permettent de comprendre comment une machine, un ordinateur, arrange[43] les données. Certes, le

39. Cf. https://www.w3.org/2018/jsonld-cg-reports/json-ld

40. C'est le sens du RDF, il facilite « l'exploitation et le traitement automatique ». Voir : « Le guide des bonnes pratiques numériques ». 2015. Huma-Num. https://www.culture.gouv.fr/Media/Thematiques/Enseignement-superieur-et-recherche/Archives-anciens-programmes/Numerique/Numerisation-du-patrimoine-Archives/La-numerisation-en-pratique/Guide-des-bonnes-pratiques-numeriques-TGE-Adonis-septembre-2011. Voir également l'entrée de glossaire correspondante pour une présentation succincte du RDF.

41. Turtle est une syntaxe textuelle pour RDF, elle permet de décrire un graphe RDF sous la forme d'un texte compact et naturel, voir https://www.w3.org/TR/2014/REC-turtle-20140225.

42. RDF/XML est une syntaxe qui permet de décrire un graphe RDF à l'aide de balises XML, voir https://www.w3.org/TR/rdf-syntax-grammar.

43. C'est à dessein que nous utilisons ce verbe, plus représentatif que « structurer » car il introduit l'idée de passage d'un état à un autre, d'une forme à une autre, tout en

Figure 7. Exemple de données exposées au format DC-RDF, relatif à un « Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune »

https://parleflandre.nakalona.fr/items/show/707?output=dc-rdf


```
−<rdf:RDF>
 −<rdf:Description rdf:about="https://parleflandre.nakalona.fr/items/show/707">
  −<dcterms:title>
     Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune
   </dcterms:title>
  −<dcterms:subject>
     <a href="/items/browse?advanced%5B0%5D%5Belement_id%5D=49&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=Huissier+de+justice">Huissier de justice</a>
   </dcterms:subject>
  −<dcterms:subject>
     <a href="/items/browse?advanced%5B0%5D%5Belement_id%5D=49&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=Organisation+judiciaire">Organisation judiciaire</a>
   </dcterms:subject>
  −<dcterms:subject>
     <a href="/items/browse?advanced%5B0%5D%5Belement_id%5D=49&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=R%C3%A9mun%C3%A9ration">Rémunération</a>
   </dcterms:subject>
   <dcterms:creator>Plouvain, Pierre-Antoine-Samuel-Joseph (1754-1832)</dcterms:creator>
   <dcterms:creator>Six, Philippe-Josse-Auguste-Joseph (1732-1793)</dcterms:creator>
  −<dcterms:source>
     http://nordnum.univ-lille.fr/ark:/72505/a0115224001360lm5Ol/from/a011522400136ypxWl2
   </dcterms:source>
  −<dcterms:created>
     <a href="/items/browse?advanced%5B0%5D%5Belement_id%5D=56&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=14%2F05%2F1725">14/05/1725</a>
   </dcterms:created>
  −<dcterms:issued>
     <a href="/items/browse?advanced%5B0%5D%5Belement_id%5D=60&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=01%2F07%2F2019">01/07/2019</a>
   </dcterms:issued>
  −<dcterms:issued>
     <a href="/items/browse?advanced%5B0%5D%5Belement_id%5D=60&advanced%5B0%5D%5Btype%5D=is+exactly&advanced%5B0%5D%5Bterms%5D=1790">1790</a>
   </dcterms:issued>
   <dcterms:contributor>Centre d'Histoire Judiciaire (CHJ) - UMR 8025</dcterms:contributor>
   <dcterms:type>Text</dcterms:type>
   <dcterms:identifier>11280/afe3b32f</dcterms:identifier>
   <dcterms:coverage>Parlement de Flandre</dcterms:coverage>
 </rdf:Description>
```


Figure 8. Exemple de données au format Omeka-XML, relatif à un « Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune »

https://parleflandre.nakalona.fr/items/show/707?output=omeka-xml


```
−<itemContainer xsi:schemaLocation="http://omeka.org/schemas/omeka-xml/v5 http://omeka.org/schemas/omeka-xml/v5/omeka-xml-5-0.xsd" uri="https://parleflandre.nakalona.fr/items/browse?output=omeka-xml"
 accessDate="2019-10-03T14:52:17+02:00">
 −<miscellaneousContainer>
  −<pagination>
     <pageNumber>1</pageNumber>
     <perPage>50</perPage>
     <totalResults>486</totalResults>
   </pagination>
 </miscellaneousContainer>
 −<item itemId="707" public="1" featured="0">
  −<fileContainer>
   −<file fileId="432">
    −<src>
       https://parleflandre.nakalona.fr/files/original/f05e055089c302280c6d26c88290f0d7.pdf
     </src>
     <authentication>db2722c362fd519f5b8bd66a98e3a3d6</authentication>
   </file>
  </fileContainer>
 −<collection collectionId="15">
  −<elementSetContainer>
   −<elementSet elementSetId="1">
     <name>Dublin Core</name>
    −<description>
       The Dublin Core metadata element set is common to all Omeka records, including items, files, and collections. For more information see, http://dublincore.org/documents/dces/.
     </description>
    −<elementContainer>
     −<element elementId="50">
       <name>Title</name>
       <description>A name given to the resource</description>
      −<elementTextContainer>
       −<elementText elementTextId="9471">
        −<text>
           Ordonnances et arrêts de règlement du parlement de Flandre
         </text>
       </elementText>
      </elementTextContainer>
     </element>
    −<element elementId="43">
       <name>Identifier</name>
      −<description>
         An unambiguous reference to the resource within a given context
       </description>
```

Figure 9. Exemple de métadonnées publiques exposées dans Nakala au format Turtle, relatif à un « Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune »

```
@prefix ore:   <http://www.openarchives.org/ore/terms/> .
@prefix geonames: <http://www.geonames.org/ontology#> .
@prefix void:  <http://rdfs.org/ns/void#> .
@prefix owl:   <http://www.w3.org/2002/07/owl#> .
@prefix xsd:   <http://www.w3.org/2001/XMLSchema#> .
@prefix yago:  <http://localhost:8080/class/yago/> .
@prefix skos:  <http://www.w3.org/2004/02/skos/core#> .
@prefix ir:    <http://www.ontologydesignpatterns.org/cp/owl/informationrealization.owl#> .
@prefix rdfs:  <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dbpedia: <http://localhost:8080/resource/> .
@prefix units: <http://dbpedia.org/units/> .
@prefix nkl:   <http://localhost:8080/pubby/> .
@prefix geo:   <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix p:     <http://localhost:8080/property/> .
@prefix rdf:   <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix prv:   <http://purl.org/net/provenance/ns#> .
@prefix doap:  <http://usefulinc.com/ns/doap#> .
@prefix meta:  <http://example.org/metadata#> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix prvTypes: <http://purl.org/net/provenance/types#> .
@prefix foaf:  <http://xmlns.com/foaf/0.1/> .
@prefix dc:    <http://purl.org/dc/elements/1.1/> .

<https://www.nakala.fr/data/resource/11280/afe3b32f?output=ttl>
        rdfs:label          "RDF description of Afe3b32f" ;
        foaf:primaryTopic   <https://www.nakala.fr/resource/11280/afe3b32f> .

<https://www.nakala.fr/resource/11280/afe3b32f>
        a                   foaf:Document ;
        dcterms:creator     <https://www.nakala.fr/user/98839290-95d0-4ae7-8dee-f0c0b566efd1> ;
        dcterms:extent      "891179"^^xsd:long ;
        dcterms:format      "PDF" ;
        dcterms:identifier  "11280/afe3b32f" ;
        dcterms:issued      "2019-07-03T10:48:11.096+02:00"^^xsd:dateTime ;
        dcterms:modified    "2019-07-03T11:02:11.865+02:00"^^xsd:dateTime ;
        dcterms:publisher   <https://www.nakala.fr/account/11280/5f539340> ;
        ore:isAggregatedBy  <https://www.nakala.fr/collection/11280/7681d5cf> ;
        skos:altLabel       "2056.pdf" ;
        foaf:primaryTopic   <https://www.nakala.fr/data/11280/afe3b32f> ;
        foaf:sha1           "adba0df29e78a1249b7adea51c2b9d7f00ece70f" .

<https://www.nakala.fr/collection/11280/7681d5cf>
        ore:aggregates  <https://www.nakala.fr/resource/11280/afe3b32f> .
```

Figure 10. Exemple de métadonnées publiques exposées dans Nakala au format RDF_XML, relatif à un « Arrêt du parlement, concernant les salaires des Huissiers pour les devoirs qu'ils font au profit de leur bourse commune »

https://www.nakala.fr/data/resource/11280/afe3b32f?output=xml

```
−<rdf:RDF>
  −<rdf:Description rdf:about="https://www.nakala.fr/data/resource/11280/afe3b32f?output=xml">
     <rdfs:label>RDF description of Afe3b32f</rdfs:label>
    −<foaf:primaryTopic>
      −<foaf:Document rdf:about="https://www.nakala.fr/resource/11280/afe3b32f">
         <dcterms:publisher rdf:resource="https://www.nakala.fr/account/11280/5f539340"/>
         <foaf:sha1>adba0df29e78a1249b7adea51c2b9d7f00ece70f</foaf:sha1>
         <dcterms:identifier>11280/afe3b32f</dcterms:identifier>
         <dcterms:modified rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2019-07-03T11:02:11.865+02:00</dcterms:modified>
        −<ore:isAggregatedBy>
          −<rdf:Description rdf:about="https://www.nakala.fr/collection/11280/7681d5cf">
             <ore:aggregates rdf:resource="https://www.nakala.fr/resource/11280/afe3b32f"/>
           </rdf:Description>
         </ore:isAggregatedBy>
         <dcterms:format>PDF</dcterms:format>
         <skos:altLabel>2056.pdf</skos:altLabel>
         <dcterms:extent rdf:datatype="http://www.w3.org/2001/XMLSchema#long">891179</dcterms:extent>
         <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">2019-07-03T10:48:11.096+02:00</dcterms:issued>
         <foaf:primaryTopic rdf:resource="https://www.nakala.fr/data/11280/afe3b32f"/>
         <dcterms:creator rdf:resource="https://www.nakala.fr/user/98839290-95d0-4ae7-8dee-f0c0b566efd1"/>
       </foaf:Document>
     </foaf:primaryTopic>
   </rdf:Description>
 </rdf:RDF>
```

porteur d'un projet numérique n'est peut-être pas amené à mettre ses mains dans cet « arrangement », mais aujourd'hui, il doit, nous semble-t-il, comprendre l'intérêt de ces formats[44].

40. Au-delà de ces expositions de données dans différents formats, Omeka propose un entrepôt de métadonnées, interopérable grâce au protocole OAI*-PMH* (Prime-Claverie et Mahé 2017). Cette solution donne la possibilité de diffuser les enregistrements du projet *Pandectes des Flandres* sur la plateforme Isidore (figure 11). Ainsi Pandectes des Flandres est aussi une collection à part entière dans Isidore[45].

41. Outre la diffusion du corpus numérique vers une communauté spécifique aux sciences humaines et sociales, le projet *Isidore* enrichit les données à travers plusieurs référentiels comme HAL[46], OpenEdition[47] ou Rameau[48].

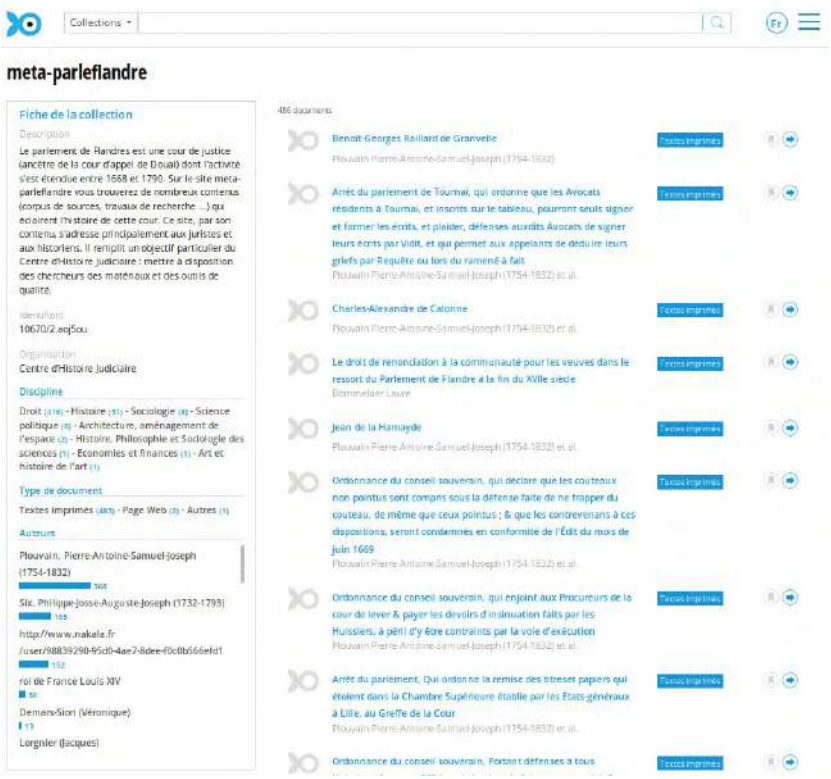


Figure 11. Vue de la collection Pandectes des Flandres sur Isidore
Crédit : Renaud Limelette

Conclusion

42. En fin de compte, à partir de deux projets connexes sur le fonds, en tant qu'ils touchent à valoriser la connaissance du parlement de Flandre, les réalisations numériques diffèrent largement. Dans un court laps de temps, moins d'une dizaine d'années, les outils numériques dis-

faisant référence à l'idée d'organisation. Il englobe alors à la fois la structuration d'un document XML et la logique du Web sémantique.

44. Sur ce questionnement on se reportera à (La Barre 2010 ; Bachimont *et al.* 2011).

45. Collection disponible sur https://isidore.science/collection/10670/2.aoj5ou.

46. Le référentiel de HAL est disponible ici : https://halshs.archives-ouvertes.fr/browse/domain.

47. Le référentiel d'OpenEdition est disponible ici : https://journals.openedition.org/index/49.

48. Le référentiel Rameau est disponible ici : http://rameau.bnf.fr.

ponibles en sciences humaines et sociales, et par là les possibilités techniques pouvant être facilement mises en œuvre, ont largement permis aux porteurs de projets de se libérer de contraintes techniques en s'appuyant sur des solutions évaluées par la très grande infrastructure de recherche qu'est Huma-Num.

43. Ainsi, le projet *ParleFlandre*, antérieur au projet *Pandectes des Flandres*, n'a pas pu dès sa conception s'appuyer sur la TGIR Huma-Num. Si des solutions techniques existaient déjà au moment du montage du projet *ParleFlandre*, l'accent a de préférence été mis sur l'analyse des dossiers d'archives afin de remplir la base de données. Mais cette analyse n'a pas anticipé l'utilisation d'un vocabulaire contrôlé, notamment pour la description de l'objet du litige, ce qui rendrait aujourd'hui l'interrogation des données de la base plus efficace. Heureusement, tout le projet *ParleFlandre* a été conçu avec des outils libres et donc facilement adaptables.

44. Cette constatation faite, ceci nous amène à la réflexion suivante. Le porteur d'un projet numérique doit rester vigilant sur l'évolution des humanités numériques. Aujourd'hui, tout bouge très vite et il ne faudrait pas que les choix faits dans la précipitation de la conception d'un projet soient bloquants ou obsolètes très peu de temps après. Pour se prémunir de telles déconvenues, nous ne pouvons que rappeler que le suivi de bonnes pratiques, étape par étape, et l'utilisation de solutions libres sont des éléments essentiels dans la conduite de projet. La très grande infrastructure de recherche Huma-Num, par les services qu'elle propose, facilite le tournant numérique de la recherche en sciences humaines et sociales. Tout un dispositif technologique, appelé grille de services, est mis en place pour aider les porteurs de projet à traiter, conserver, donner accès et rendre interopérables les données de la recherche. Assez récemment en France, avec le développement de la science ouverte[49], chaque projet financé par des fonds publics doit maintenant structurer et ouvrir ses données selon les principes du FAIR*[50] et prévoir un plan de gestion des données[51]. Or, mettre en place des données *Fair* ou un plan de gestion ne peut se faire seul, et donc des institutions reconnues, comme des très grandes infrastructures de recherche ou des réseaux spécialisés, sont des structures incontournables pour qui veut mener à bien son projet numérique.

49. Le Plan national pour la science ouverte a été dévoilé le 4 juillet 2018 par le ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation, http://corist-shs.cnrs.fr/sites/default/files/ressources/plan_national_scienceouverte_2018.pdf. Pour une brève présentation générale de l'Open access, le lecteur pourra se référer au glossaire.

50. Acronyme pour « Facile à trouver », « Accessible », « Interopérable » et « Réutilisable ».

51. Un plan de gestion des données accompagne dans le temps la vie des données, il conduit le chargé de projet à prévoir des standards et des formats autour de la qualité des données, leur description, stockage et sauvegarde, un questionnement éthique et juridique, leur partage et conservation à long terme, ainsi que les responsabilités et les ressources allouées à la gestion des données.

Les technologies du Web pour la valorisation d'un patrimoine industriel textile en mouvement dans les Hauts-de-France

Éric Kergosien et Mathilde Wybo

Introduction et objectifs

1. Une question sociale importante dans le domaine du patrimoine culturel est liée à la collecte, l'analyse, la publication et la mise en valeur de l'histoire et de la mémoire collectives des acteurs du domaine, à l'oral comme à l'écrit. En ce sens, la formalisation de l'information sur le patrimoine culturel représente un véritable défi en raison de la diversité et de l'incomplétude des données. De plus, ces données sont hétérogènes et peuvent être trouvées dans différentes sources, en ligne ou hors ligne : bases de données, bibliothèques, musées, dossiers de presse, expertise des intervenants, etc. Cette diversité des ressources pose de nombreux problèmes tels que la documentation des données, la représentation, l'intégration et l'interopérabilité au sein d'une même base de connaissances. La plupart des tentatives pour résoudre les problèmes d'interopérabilité sémantique se concentrent sur la standardisation, voire la normalisation de formalismes communs tels que FRBR, FRBRoo, CIDOC CRM, etc. Parmi ces modèles, le CIDOC CRM est un modèle conceptuel de référence conçu pour modéliser les domaines du patrimoine culturel.

2. Afin d'aider les experts du domaine à valoriser des collections numériques disponibles sur le domaine textile, les chercheurs des laboratoires GERIICO (sciences de l'information et de la communication), STL (linguistique) et IRHIS (histoire), collaborant dans le projet *DENIM* (Données numériques, langages et représentations du patrimoine textile en région Nord-Picardie : quelles compréhensions réciproques ? – financement ministère de la Culture), proposent une méthodologie semi-automatique visant à construire une base de connaissances, que l'on nomme également ontologie de domaine. Les difficultés dans ce travail sont nombreuses, à commencer par la construction d'un vocabulaire structuré décrivant le domaine de l'industrie textile, puis la mise en place d'une méthodologie semi-automatique permettant d'extraire les informations pertinentes relatives au domaine à partir du vocabulaire défini, et de structurer en une ontologie les connaissances identifiées dans le corpus analysé. L'ontologie produite est formalisée en OWL CIDOC CRM, offrant ainsi une première représentation sémantique du domaine implicitement décrite dans le corpus analysé. Des compétences en analyse et traitement automatique des langues, en extraction et

structuration de données et dans le domaine du Web sémantique sont mobilisées à cette fin[1].

3. Avant de présenter la méthodologie mise en place dans le projet, nous présentons le besoin de valoriser les données en exploitant les nouvelles technologies du Web sémantique. Nous concluons en explicitant les perspectives du projet.

Contexte

Le Web sémantique et les données ouvertes

4. Le Web sémantique ou Web 3.0 est une technologie numérique qui intéresse les organismes de grande taille confrontés à la multitude et à la diversité des données car il permet de lier et de faire communiquer des données d'origines très différentes. Celles-ci doivent être traduites dans un langage informatique spécifique propre aux données ouvertes (*open data**) afin de répondre aux exigences techniques exprimées par les besoins d'accès automatique au sens des informations plutôt qu'à leur forme : XML*, RDF*, OWL* ou SPARQL* en sont quelques exemples emblématiques. Mais si ces langages offrent bien les fonctionnalités nécessaires pour la mise en œuvre d'outils de traitement du sens, ils ne préjugent cependant pas de l'angle sous lequel sont abordées les données et leur signification, et ils laissent toute latitude pour en évoquer la logique. En effet, un seul mode de représentation du sens n'est pas capable à ce jour – et ne sera sans doute jamais capable – de prendre en charge la description universelle des données dans toutes leurs dimensions.

5. Si aucune structure informationnelle spécifique ne semble avoir été conçue pour décrire le patrimoine industriel textile, il existe cependant plusieurs exemples de formalismes créés pour décrire les objets culturels. Les principaux formalismes sont des modèles complexes qui permettent de décrire les objets culturels tout en exprimant les relations pouvant exister entre eux soit explicitement, soit en facilitant l'utilisation d'outils du Web sémantique pour dépasser l'implicite. Il s'agit des modèles FRBR (Le Bœuf 2013), CIDOC CRM (Doerr 2003), et FRBRoo (Doerr, Bekiari et Le Bœuf 2008).

6. FRBR (Functional Requirements for Bibliographic Records) est un modèle de description qui distingue quatre niveaux d'information portant sur un même objet (initialement bibliographique) depuis ses caractéristiques physiques qui doivent être distinguées pour chaque exemplaire (« item ») jusqu'aux spécificités les plus abstraites de sa conception (« œuvre ») en passant par les spécifications de sa mise à disposition d'un public (« manifestation ») et celles de son contenu intellectuel (« expression »). À chaque niveau de description – du plus matériel au plus conceptuel – le renseignement des champs informationnels n'est pas forcément opéré par

1. Le lecteur pourra se référer au glossaire de l'ouvrage pour toutes les définitions relatives aux FRBR, au langage OWL, au TAL, au Web sémantique, à une ontologie, ainsi qu'au principe d'interopérabilité.

une explicitation locale, mais autant que faire se peut par une référence au modèle FRAD (pour les personnes physiques et morales) ou au modèle FRSAD (pour les lieux, événements, concepts et objets). Un dense réseau de relations se construit dès lors entre les œuvres, entre les autorités et entre les descripteurs qui y sont attachés, sortant des limites classiques de la fiche descriptive.

7. Le modèle conceptuel de référence (*conceptual reference model*) CIDOC CRM est un modèle de représentation de données conçu par le Comité international pour la documentation du Conseil international des musées pour permettre l'interopérabilité des référencements des objets de musées, puis par extension de tout objet de patrimoine culturel physique ou non, selon la définition proposée par l'UNESCO. Il vise à dépasser les incompatibilités sémantiques et structurales des nombreuses sources d'informations hétérogènes portant sur des réalités patrimoniales et culturelles pour faciliter l'échange de documentations et la recherche dans ces documentations. La version actuelle (ISO 21127:2014) intègre 86 classes (acteurs, lieux, événements ou entités temporelles…) qui sont reliées entre elles par 137 propriétés distinctes. Le modèle est assorti de plusieurs outils, dont des implémentations OWL et RDF et des utilitaires de *mapping* avec d'autres formalismes (UNIMARC, EDM…).

8. FRBRoo est une évolution « orientée objet » imaginée à partir de FRBR et de CIDOC CRM. Reprenant les quatre niveaux de description de FRBR, il fait des entités originelles des conteneurs chargés d'intégrer les classes CIDOC CRM pour assurer l'interdépendance entre la richesse des deux modèles. Très ambitieuse, l'ontologie FRBRoo est conçue pour prendre en charge, décrire et mettre en relation toute réalité de l'univers culturel. Le modèle dans son état actuel n'est pas encore stabilisé, et toutes les questions conceptuelles qu'il soulève n'ont pas encore obtenu de réponse. Son développement est néanmoins organisé de manière à ce qu'il puisse être instancié automatiquement par des données issues de ses modèles « parents », CIDOC CRM et FRBR.

9. Du fait de son niveau élevé de maturité et de sa stabilité, de son adéquation avec les données du projet ainsi qu'avec son objectif d'agrégation de données hétérogènes, nous avons choisi de mettre en œuvre l'ontologie CRM CIDOC. Bien entendu, les outils déjà proposés, de même que son interopérabilité planifiée avec son évolution que constitue FRBRoo, nous ont également guidés dans ce choix.

10. L'état de l'art met en avant différents travaux pour la valorisation du patrimoine culturel en proposant de construire une ontologie de domaine au format CIDOC CRM. Nous pouvons notamment citer les travaux pour la construction d'une ontologie du patrimoine bâti (Messaoudi 2017), des chercheurs proposant un enrichissement du modèle pour une description fine du bâti (Billen *et al.* 2018). D'autres travaux se sont également appuyés sur cette norme pour proposer une modélisation des objets du patrimoine antique (Sza-

bados et Letricot 2014), ou encore de la relation entre objets culturels et connaissance associée (Possompès 2017). Des travaux intéressants présentent également une démarche semi-automatique pour modéliser le patrimoine matériel à partir de connaissances orales (Du Château, Boulanger et Mercier-Laurent 2012). Des groupes de réflexion travaillent également à l'évolution de ce modèle et à sa mise en relation avec les autres modèles sémantiques pour la description du patrimoine culturel à l'échelle européenne notamment (Doerr *et al.* 2013 ; Isaac et Charles 2015). Cependant, il n'existe pas de travaux à notre connaissance visant à modéliser les domaines du patrimoine industriel textile et minier, en prenant notamment en compte les aspects immatériels, ce qui est l'un de nos objectifs ici.

Améliorer la lisibilité du patrimoine textile régional

11. Un constat identique a été fait dans les études antérieures menées par les chercheurs du projet : celui d'une immense variété et richesse du patrimoine, liée à l'ancrage historique et géographique de cette activité dans l'Europe du Nord-Ouest et en particulier, pour la France, en région Hauts-de-France (Nord, Pas-de-Calais, Picardie), mais d'une faible visibilité.

12. La richesse et la diversité des ressources sont incontestablement un atout pour le territoire. En effet, les ressources (fonds, collections d'objets et de machines, témoignages) sont de natures très diverses et présentes dans de nombreux lieux (musées, bibliothèques, associations, entreprises, etc.) (Hurez 2015). Il apparaît cependant que ces ressources sont peu visibles, et dispersées. En effet, malgré des initiatives en faveur de leur valorisation depuis la fin des années 1970 et l'existence de musées à Fourmies, Calais, Caudry, Roubaix, Fresnoy-le-Grand notamment, ce constat semble s'inscrire dans celui plus général d'un manque de lisibilité du domaine de la culture scientifique, technique et industrielle en région Hauts-de-France. Les collections techniques, industrielles et scientifiques sont assez mal connues et peu appropriées (Wybo 2017).

13. Pour être étudiés et valorisés, les savoirs dispersés liés au textile ont besoin d'être connus, identifiés et mis en lien. Le besoin « d'établir une documentation raisonnée à disposition des chercheurs et autres publics, pour favoriser une meilleure connaissance du domaine et favoriser la diffusion de celle-ci auprès d'un large public » avait déjà été souligné par Jean-Charles Leyris dans une note d'intention de recherche dès 2011. Le projet de la « Cité régionale de l'histoire des gens du textile », porté par l'Union des gens du textile (association loi 1901), prévoyait, à partir de ce premier constat, la création d'un centre de ressources permettant de conserver (mais aussi collecter) les sources, notamment orales et audiovisuelles. L'objectif de ce centre de ressources était de mettre en valeur le patrimoine, les recherches et les publications liées à l'histoire textile ainsi que de permettre d'effectuer des recherches sur des entreprises et d'anciens salariés, grâce à une banque de données qui serait complétée par les archives

du personnel, dans le respect des délais légaux de consultation. Le projet de *learning center*, porté depuis 2014 par l'association Les Amis du Ciretex, semble rejoindre cet objectif. Pour cette association, ce dispositif doit être « un lieu pédagogique en direction des jeunes, un outil contribuant à promouvoir la recherche, un projet scientifique d'études et de médiation défini avec l'ensemble des acteurs concernés ». À noter qu'à la demande de cette association, la possibilité pour la région de « mener une réflexion sur l'opportunité de créer un *Learning Center* textile ou mémoires du textile et, le cas échéant, d'en assurer la préfiguration » a été inscrite au contrat de plan État-région (CPER) 2015-2020. Ce projet s'inscrirait dans le cadre du développement du *learning center* de l'université de Lille.

14. Cet objectif repose sur un important travail de repérage des ressources, des collections et des travaux de recherche à l'échelle régionale. Il fut initié au sein des laboratoires IRHIS et GERIICO en proposant une approche permettant de cartographier les acteurs du domaine ainsi que leurs corpus disponibles, combinant des entretiens semi-directifs ainsi qu'une cartographie du Web issue d'une veille automatisée (Kergosien, Severo et Berthelot 2019). Ce travail a permis d'identifier 169 acteurs du domaine (et ressources associées) sur le territoire des Hauts-de-France.

La collecte et l'analyse des données

La collecte de documents hétérogènes

15. À partir de cette liste d'acteurs produisant ou diffusant du contenu sur le thème du patrimoine industriel textile, nous avons collecté un premier corpus constitué de 6 000 documents hétérogènes : images avec notices descriptives XML de la bibliothèque Georges Lefebvre à Lille 3, articles de presse de *La Voix du Nord* et de *Nord Éclair*, témoignages retranscrits, documents du Service commun de la documentation de l'université de Lille, notices et Plan local d'urbanisme disponibles sur le portail de la MEL, notices de l'Inventaire de la Région. Sur cette base de documents collectés, nous nous sommes concentrés dans ces premiers travaux sur trois types de documents pour la phase d'extraction et de structuration des connaissances liées au domaine :

1. 142 articles de presse (*La voix du Nord*, *Nord Éclair*) collectés entre 2004 et 2016 par l'AASPT (Association des anciens salariés du Peignage de la Tossée) et numérisés via l'Agence nationale de reproduction des thèses (ANRT, Lille 3) (figure 1 et 2)
2. 59 témoignages retranscrits auprès des anciens du domaine par des étudiants de l'Institut social de Lille en 2012
3. des ouvrages anciens sur l'industrie textile et sur l'exposition internationale de Roubaix de 1911 (bibliothèque Georges Lefebvre, université Lille 3) numérisés et océri-

sés via l'ANRT. Ces ouvrages ont été mis à disposition du public sur le portail NordNum (bibliothèque numérique de l'université Lille 3 consacrée à l'histoire du Nord et du Pas-de-Calais)

Quatre cents élèves sont venus visiter le mini-musée des Anciens de la Tossée

Le gamin montre fièrement son écharpe 100 % acrylique. « ça c'est de la laine ! » Ou presque ! Quatre cents enfants d'écoles primaires de Tourcoing ont visité le mini-musée des Anciens de la Tossée au Pont-rompu. « *En consultant les registres, un gamin a reconnu la photo de son grand-père* », raconte Maurice Vidrequin, vice-président de l'association des Anciens de la Tossée. Pour ces anciens du textile, la transmission de l'histoire est indispensable. « *Les jeunes de Roubaix, Tourcoing et Wattrelos en ont besoin pour mieux comprendre le monde de demain. Et souvent ils sont déjà étonnés que la laine vienne du mouton !* » Pour inciter plus de jeunes à venir découvrir ce lieu de mémoire, l'association a conclu un partenariat avec la fédération des parents d'élèves PEEP. « *Notre rôle sera de diffuser l'information* », précise Rabah Mezine, président départemental. Philippe Vrand, président national, a promis un reportage dans le magazine national de la PEEP. « *On relaiera auprès du ministère de l'Éducation nationale cette initiative. Il est important de montrer aux jeunes les métiers pour les aider dans leur orientation.* » ■ A.CL

Figure 1. Un article de presse numérisé

Crédit : Éric Kergosien et Mathilde Wybo

Quatre cents élèves sont venus visiter le mini-musée des Anciens de la Tassée
Le gamin montre fièrement son écharpe 100 % acrylique. « ça c'est de la laine ! » Ou presque ! Quatre cents enfants d'écoles primaires de Tourcoing ont visité le mini-musée des Anciens de la Tos-sée au Pont-rompu. « *En consultant les registres, un gamin a reconnu la photo de son grand-père* », raconte Maurice Vidrequin, vice- président de l'association des Anciens de la Tossée. Pour ces anciens du textile, la transmission de l'histoire est indispensable. « *Les jeunes de Roubaix, Tourcoing et Wattrelos en ont besoin pour mieux comprendre le monde de demain. Et souvent ils sont déjà étonnées que la laine vienne du mouton !* » Pour inciter plus de jeunes à venir découvrir ce lieu de mémoire, l'association a conclu un partenariat avec
la fédération des parents d'élèves PEEP. « *Notre rôle sera de diffuser l'information* », précise Rabah Me- zine, président départemental. Philippe Vrand, président national, a promis un reportage dans le maga
zine national de la PEEP. « On relaiera auprès du ministère de l'Éducation nationale cette initiative. Il est important de montrer aux jeunes les métiers pour les aider dans leur orientation. » • A. CL.

Figure 2. Le texte après océrisation

Crédit : Éric Kergosien et Mathilde Wybo

16. Une fois les documents collectés, l'étape suivante consiste à extraire les informations pertinentes de façon semi-automatique. L'extraction semi-automatique d'informations se réfère au fait d'interroger des documents textuels bruts provenant de différentes sources (et donc de nature hétérogène ; cela peut être des ouvrages, des articles de presse, des témoignages, etc.) avec des outils informatiques de façon à en extraire les informations qu'ils contiennent et à les visualiser. Une dernière étape s'est ensuite concentrée sur le volet structuration des informations extraites en une ontologie de domaine décrivant le patrimoine industriel du textile décrit dans le corpus documentaire traité. L'ensemble de la méthodologie est présenté figure 3.

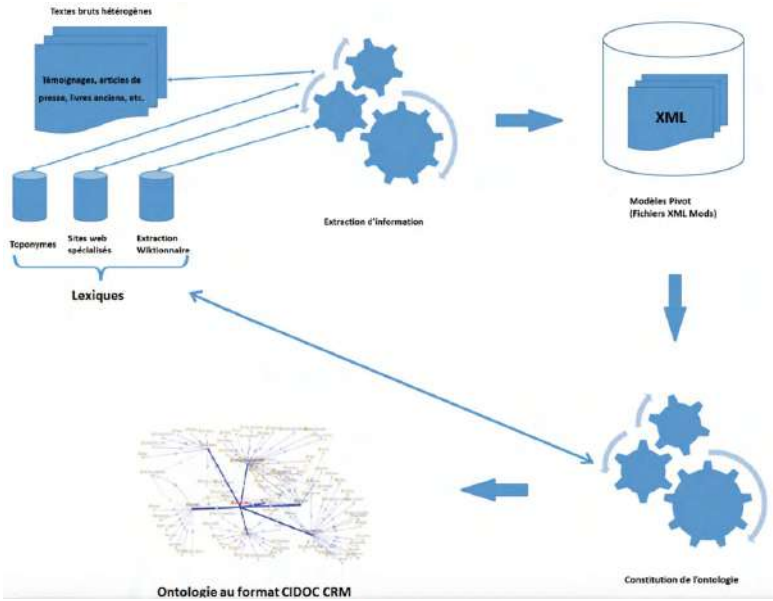


Figure 3. Démarche globale du traitement des documents pour la construction d'une première base de connaissance

Crédit : Éric Kergosien et Mathilde Wybo

La construction du vocabulaire expert de domaine

17. Parmi l'ensemble des descripteurs caractérisant le domaine du patrimoine de l'industrie textile, nous nous concentrons dans nos travaux sur les informations de type acteur, lieu et thématique qui sont présentes dans le corpus analysé. Le lexique des acteurs est constitué des 169 acteurs identifiés lors de la phase de cartographie des acteurs du domaine, lexique enrichi par la liste des acteurs interrogés dans le cadre du projet *DENIM*. Le lexique des lieux est constitué de l'ensemble des communes de la région Hauts-de-France. Ces listes sont disponibles sur *Wikipédia* et ont été extraites à partir des pages web, comme par exemple dans le cas des départements du Nord[2] ou du Pas-de-Calais[3].

18. Pour aborder la tâche de création d'un lexique thématique, correspondant à l'un des enjeux principaux du projet, il est important de préciser ici qu'il n'existe pas de ressource lexicale caractérisant de façon précise le domaine de l'industrie textile en assurant une couverture importante. Les acteurs de la société civile partenaires du projet (différentes associations et institutions spécialisées autour du domaine) déplorent ce manque d'un lexique unifié qui couvrirait le domaine du textile.

19. Il a d'abord été envisagé d'utiliser des ressources spécialisées déjà existantes. Ces ressources incluent par exemple le lexique Rameau, produit par la BNF ou le lexique Joconde, produit par les musées de France. Ces lexiques couvrent plusieurs domaines. Ils sont également bruts, c'est-à-dire que les termes qu'ils contiennent ne sont associés à aucune information qui permettrait par exemple d'identifier le domaine auquel ils sont liés. Les utiliser nécessite donc de les filtrer pour identifier les termes liés au domaine du textile. Mettre en place une méthode automatique pour arriver à cette identification représenterait un long travail et qui restera limité en portée car ces lexiques généralistes ne comptent que peu de termes du domaine. Le choix a donc été de partir de la version *XMLisée* du *Wiktionary* français : GLAWI. Le *Wiktionary* est une ressource grand public facilement disponible et libre d'utilisation. Sa version *XMLisée* facilite grandement sa manipulation.

20. De plus, le *Wiktionary* contient de nombreuses balises qui marquent les domaines auxquels se rattachent les termes spécialisés, que ce soit au niveau des articles eux-mêmes ou de leurs définitions. Les choses ont encore été facilitées grâce à l'intervention de Franck Sajous – un des concepteurs de GLAWI (Hathout et Sajous 2016). GLAWI est ainsi à notre connaissance le lexique disponible en ligne le plus représentatif du domaine (300 termes) parmi les lexiques existants (Rameau, Joconde, etc.) qui sont soit trop généraux, soit trop orientés sur l'un des sous-thèmes du domaine. Nous l'avons donc sélectionné comme ressource pivot pour la construction de notre lexique de domaine.

2. Cf. https://fr.wikipedia.org/wiki/Liste_des_communes_du_Nord
3. Cf. https://fr.wikipedia.org/wiki/Liste_des_communes_du_Pas-de-Calais

21. Dans un premier temps, nous avons recherché les ressources existantes sur le Web pour proposer une première couverture sémantique du domaine. La liste des lexiques identifiés est présentée figure 4.

Source	Type d'acteur	Adresse	Type	Domaine
Aux Tissus de Roubaix	Boutique grand public	http://auxtissusderoubaix.fr/pages/le-vocabulaire-du-textile	Dictionnaire	Textile
Agent Textile	Boutique professionnels	http://www.agent-textile.com/index_fichiers/lexiques.htm	Lexique FR ↔ EN	Textile
metierdart.com	Formation et consultation	http://intranet.metierdart.com/documents/lexique_textile_2009.pdf	Dictionnaire	Textile
Musée La Fabrique	Musée	http://www.museelafabrique43.free.fr/wt/?page_id=67	Dictionnaire	Technique
Peter Hahn	Boutique grand public	http://www.peterhahn.fr/lexique-textile-et-mode_a	Dictionnaire	Textile
Marquage Textile ZAM	Boutique professionnels	http://www.marquage-textile-zam.com/lexique-textile.html	Dictionnaire	Textile
Wikipédia – Glossaire du Tissage	Fondation	https://fr.wikipedia.org/wiki/Glossaire_du_tissage	Dictionnaire	Industrie textile
Garnier Thiébaut	Boutique grand public	http://www.garnier-thiebaut.fr/content/12-lexique-textile	Dictionnaire	Textile
Le Sentier	Place de marché professionne	http://www.le-sentier.com/dico/	Lexique FR/EN/DE/ES/I	Textile
De Fursac	Boutique grand public	http://www.defursac.fr/fr/vetement-homme-abecedaire.html	Dictionnaire	Textile
Bugis	Créateur de mode	http://www.bugis.fr/bugis-mode/lexique-glossaire-definitions-.php	Dictionnaire	Textile
Michèle LARDY	Universitaire (anglicisté)	http://langues.univ-paris13.fr/glossaire/patrimoine/francais-anglais.pdf	Lexique FR ↔ EN	Patrimoine
Joconde	Institution publique		Lexique	Patrimoine/Textile

Figure 4. Lexiques collectés sur le Web

Crédit : Éric Kergosien et Mathilde Wybo

22. Afin d'intégrer les termes récoltés à notre lexique déjà existant, les lexiques provenant du Web ont subi différents traitements pour être unifiés. En effet, les lexiques téléchargés prennent de nombreuses formes, aussi bien au niveau de l'organisation de leurs informations que de leur organisation technique. Dans la liste des lexiques, nous avons des lexiques multilingues, des dictionnaires, ainsi que des listes de mots bruts qui nous intéressent particulièrement pour l'intégration à notre chaîne de traitements.

23. Cependant, lors de la conversion, il peut être intéressant de conserver les autres informations présentes pour d'éventuelles tâches ultérieures. Nous ne détaillerons pas ici les méthodes de conversion des lexiques, ces méthodes n'étant pas intéressantes en dehors des cas spécifiques où elles ont été utilisées. Nous évoquerons cependant les différentes étapes nécessaires selon les cas, et les outils ayant servi à les réaliser :

- Téléchargement des lexiques : récupération automatique des données via une approche de moissonnage des données
- Conversion des lexiques dans un format unique : certains lexiques étaient au format PDF, l'utilitaire PDF2Text a été utilisé pour les convertir. La qualité de conversion des caractères a nécessité une correction manuelle fastidieuse car relativement longue. Chaque lexique a été converti dans un format XML simple, où chaque terme est contenu dans une balise <terme>. Pour les lexiques qui contiennent plus d'informations qu'une simple liste de mots (définitions, équivalents traductionnels, synonymes, hypéronymes, etc.), ces informations sont conservées dans les attributs de l'élément <terme>
- Regroupement des termes des différents lexiques via les synonymes, les termes génériques et spécifiques (règles de construction du thésaurus)

24. Dans cette étape, nous passons d'une liste d'un peu plus de 300 termes à un lexique enrichi contenant plus de 2 000 termes, dans lequel sont maintenant détaillées les matières premières existantes, les techniques de production, etc.

L'extraction des informations pertinentes

25. Dans le même registre que le projet *DENIM*, la bibliothèque de l'université de Yale, aux États-Unis a mené un projet autour du traitement automatique du langage

(TAL) et du textile, sur la version locale du magazine *Vogue* : *Robots Reading Vogue*. Un des sous-projets, intitulé *FabricSpace : Clustering*, offre des pistes pour ce que nous cherchons à faire, à savoir acquérir de nouveaux termes spécialisés à partir des corpus eux-mêmes.

26. Leur but est de faire du *clustering*, c'est-à-dire de classer des documents selon des catégories de thèmes. Ce but ne nous intéresse pas directement, mais la description de l'outil qu'ils utilisent, Wod2Vec, et plus généralement du plongement lexical (ou *word embedding**) – à savoir l'identification de termes proches de thématiques de départ selon le contexte dans lequel ils se trouvent – offre des perspectives intéressantes pour nous[4] :

27. *One way to begin with Word Embedding is to gather a set of terms to explore. Unlike some other machine learning methods, Word Embedding allows for more control over the terms that are important to you:*
1. You can use your own domain expertise to select words you know will be significant in the corpus.
2. You can leverage the ability of Word Embedding Models to show similar terms to those you think are meaningful. This computationally-assisted, but human-mediated, form of constructing lists of meaningful words has definite advantages over techniques such as standard topic modeling. You're deeply involved in the choice of words that are meaningful to your investigation, rather than just hoping that those words turn up somewhere in a cluster or topic.

28. Les auteurs soulignent qu'ils peuvent utiliser leurs propres connaissances sur le textile et que cela représente un avantage. La situation de laquelle nous partons, avec un lexique compilant des termes liés à un domaine spécifique et que nous cherchons à enrichir semble en adéquation avec cet aspect du plongement lexical. De plus, le volet de ces travaux qui nous intéresse particulièrement est la récupération des termes similaires à ceux que nous estimons déjà comme pertinents. Nous avons choisi d'explorer cette technique pour réaliser la tâche qui nous occupe.

29. En résumé, l'approche consiste à rechercher dans le corpus documentaire, pour chacun des 2 000 termes constituant notre corpus, l'ensemble des termes présents dans les documents qui apparaissent régulièrement à proximité de celui-ci. Ces nouveaux termes sont alors candidats à enrichir le lexique. Nous filtrons les termes extraits en supprimant ceux qui apparaissent moins de 5 fois à côté du terme pour limiter le bruit. L'approche, s'appuyant sur un algorithme nommé Word2Vec issu du TAL, est détaillée dans le mémoire de Rémi Cardon (2017). Nous obtenons dans nos résultats un taux de précision[5] de 91,5 %. 48 termes candidats peuvent être ajoutés à notre lexique pour l'enrichir, ce qui représente donc

4. Pour plus d'informations, voir également : « Robots Reading Vogue. FabricSpace: Clustering ». *Digital Humanities at Yale University Library*. Consulté le 30 avril 2021. http://dh.library.yale.edu/projects/vogue/fabricspace/.

5. Un positif étant un terme lié au domaine qui nous intéresse, qu'il soit présent ou non dans notre lexique de départ.

un accroissement d'un peu plus de 10 % de la part du lexique effectivement présente dans le corpus.

30. Pour illustration (figure 5), nous reproduisons les listes des termes renvoyés pour « laine », « galon » et « gros » ; en noir, les termes déjà présents dans notre lexique de départ ; en bleu, les termes qui peuvent être ajoutés au lexique ; en orange ceux qui devraient être validés par quelqu'un de compétent sur le domaine ; et en rouge le bruit.

Figure 5. Extrait de la liste de termes candidats résultante de notre approche

Crédit : Éric Kergosien et Mathilde Wybo

31. Pour chaque document indexé, l'ensemble des données extraites sont structurées au format XML MODS, format d'indexation de documents créé par la bibliothèque du Congrès aux États-Unis.

La construction d'une première base de connaissances

32. Une fois les données extraites des documents, l'étape suivante consiste à construire une première base de connaissances permettant de structurer les informations liées au patrimoine industriel textile implicitement décrit dans les documents du corpus. En résumé, le travail consiste d'une part à formaliser chacune des informations extraites (acteurs, lieux, thématiques entités temporelles et références aux documents) dans une même base formalisée en OWL CIDOC CRM, puis d'autre part à les analyser/valider via un logiciel permettant de visualiser l'ontologie produite. Le détail de l'approche est présenté dans le mémoire de Kaouther Ben Smida (2016).

33. Pour expliciter le résultat produit, nous décrivons ici un exemple représentant l'ontologie produite pour quatre des documents de notre corpus. Notre objectif étant de traiter sémantiquement l'information issue de sources hétérogènes, nous avons conservé des documents hétérogènes issus de sources distinctes. Il s'agit de quatre documents très distincts, tant par leur forme, par leur contenu, que par leur producteur. Deux d'entre eux sont des descriptifs d'objets du patrimoine issus d'acteurs institutionnels (une fiche extraite de l'Inventaire général et une autre issue de la base « Images » du laboratoire de recherche IRHIS), se présentant sous la forme de fichiers XML, mais dont la structure informationnelle est radicalement différente. Un troisième est un article de presse de *La Voix du Nord*, en texte brut non

structuré, sans vocation descriptive spécifique ; et le dernier est un document PDF de la MEL. Tous les quatre répondent à la requête géographique, et deux d'entre eux répondent au lexique du bâti industriel (usine textile, filature, lainerie etc.).

34. L'ensemble des informations pertinentes collectées dans le corpus de test a été intégré au modèle comme instances de classes, et les propriétés qui les relient ont été générées soit directement par le modèle, soit par un moteur d'inférences intégré au modèle. La figure 6 présente une projection d'un extrait de l'ontologie peuplée par notre corpus de test, visualisée via le logiciel Protégé (Musen *et al.* 1995). Outre les possibilités d'édition et de visualisation de l'ontologie produite, Protégé permet de valider le bon formatage en langage XML OWL CIDOC CRM. Aussi, comme nous l'avions prévu, la structure informationnelle de l'ontologie est bien adaptée à la description des objets de patrimoine que nous cherchons à mettre en évidence. On notera également et surtout que les quatre documents tests traités dans cet exemple sont bien mis en relation dans le modèle. De plus, de nouvelles propriétés, absentes des sources d'information originelles, leur sont associées, soit par la puissance du modèle (dans la classe E53 Place, Roubaix est une ville du Grand Lille, lui-même agglomération du Nord, etc.), soit grâce au moteur d'inférences qui crée de nouvelles relations (un événement « E5 Event » tel que l'Exposition internationale de Roubaix est une entité temporelle – « E2 Temporal entity » – qui a forcément un début et une fin). Dans cet exemple, un premier document intitulé « IRHIS_FL1269145.xml » relate la participation du président de la République de l'époque à l'exposition internationale du textile en 1911, et un second document intitulé « MEL_Roubaix_AVA.pdf » précise que l'événement a eu lieu le long du Parc Barbieux à Roubaix, commune du nord de la France. L'ontologie produite fut présentée aux experts du domaine participant au projet, et une évaluation qualitative précise doit encore être réalisée avec la collaboration des experts.

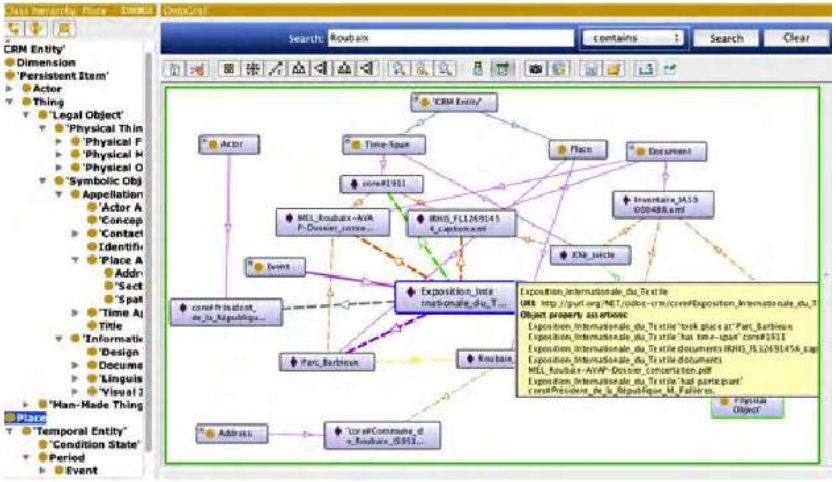


Figure 6. Extrait de l'ontologie produite
Crédit : Éric Kergosien et Mathilde Wybo

Conclusion

35. Ce travail de construction d'un premier lexique de domaine s'est appuyé sur de nombreux échanges entre les acteurs du domaine durant les réunions d'échanges

organisées dans le cadre du projet *DENIM*. Une proposition a d'ailleurs été formalisée pour constituer un groupe de travail à l'échelle de la région en incluant une perspective comparative franco-belge.

36. La construction de ce lexique doit en effet venir en appui du travail d'inventaire réalisé par les associations et les institutions patrimoniales. Éric Fossey, directeur adjoint de l'écomusée de l'Avesnois à Fourmies, a proposé l'idée d'une démarche participative afin que les publics ou visiteurs contribuent à critiquer et à enrichir les terminologies proposées (rencontre régionale textile, le 20/01/2017). Dans le même sens, Olivier Clynckemaillie, conservateur au musée de La Rubanerie de Comines, a évoqué le travail réalisé au sein de la fédération Wallonie-Bruxelles sur la réalisation d'un thesaurus* sur le textile. Des difficultés ont été rencontrées au niveau de l'appréhension scientifique du thesaurus et de la prise en compte, parallèlement, des réalités vernaculaires. C'est important, rappelle-t-il, que des mots employés par les acteurs du textile, et qui ont donc du sens, apparaissent aussi. Ils sont notamment présents dans les enquêtes ethnologiques. Il y a aussi des mots qui n'ont pas la même signification dans le Sud ou dans le Nord du pays. C'est important de construire ce thesaurus pas à pas, notamment au moment de remplir les fiches d'inventaire (rencontre régionale textile, le 20/01/2017).

37. Un des résultats du projet est donc la volonté partagée d'enrichir ce premier lexique produit en faisant intervenir différents acteurs qui constitueraient un groupe de travail. L'objectif étant à terme ici de produire une ressource la plus exhaustive possible, prenant en compte l'évolution des termes dans le temps, et également les différentes façons de nommer des sujets/objets selon les territoires.

38. Le projet *DENIM* a ainsi croisé les apports et les méthodologies de plusieurs disciplines scientifiques et des acteurs du domaine, dans leur diversité, avec des perspectives heuristiques très intéressantes qui nécessiteraient d'être approfondies. Il a constitué un pont entre des projets, des acteurs, des structures, des disciplines durant une année (Kergosien et Wybo 2017).

39. Au niveau technique, ce travail a abouti à la constitution d'une chaîne de traitements semi-automatisée complète et des ressources nécessaires à son fonctionnement, qui prend en entrée des textes bruts résultants de la phase de collecte des documents du domaine et qui produit des documents XML MODS condensant les informations recherchées (thèmes, lieux, acteurs) de manière structurée. En perspectives à ces travaux, nous souhaitons étendre la phase de construction d'une base de connaissances OWL CIDOC CRM à un volume plus important de documents. Une seconde action sera de proposer un moteur de recherche permettant de naviguer à travers le corpus indexé en s'appuyant sur la base de connaissances produite. Une action importante à mener est enfin la mise en place du processus d'évaluation de l'ontologie produite avec les experts du domaine participant au projet.

Méthodologie de validation et d'enrichissement d'une ontologie minière fondée sur le CIDOC CRM

Amélie Daloz

Introduction

1. Depuis la fin des années 1970 et principalement depuis l'inscription du bassin minier du Nord–Pas-de-Calais au patrimoine mondial de l'UNESCO en 2012, de nombreuses initiatives de valorisation des ressources liées à l'activité minière sont mises en place. Cependant, la dispersion de celles-ci sur plusieurs lieux de mémoire où les systèmes d'information et langages d'indexation utilisés sont hétérogènes empêche leur visibilité. De ce fait, la sauvegarde du patrimoine, et notamment de la culture scientifique, technique et industrielle est en danger (Chaudiron, Jacquemin et Kergosien 2019, 313).

2. Pour répondre à cette problématique, le projet *Mémo-Mines*[1], dans lequel s'inscrit notre recherche, vise à mettre en place un cadre conceptuel et des dispositifs innovants pour permettre l'analyse, la valorisation, et la sauvegarde du patrimoine minier et plus spécifiquement la mémoire minière.

3. Dans notre travail de recherche, nous abordons cette problématique par l'identification et la structuration des connaissances du domaine. Nous créons et interrogeons conjointement trois systèmes d'organisation des connaissances (SOC*) dont une terminologie*, un thésaurus* et une ontologie*. La problématique du chapitre porte quant à lui sur l'approche de validation et d'amélioration du dernier système, l'ontologie du domaine minier.

4. Dans une première partie, nous présentons le cadre théorique de notre étude en trois points : l'ontologie comme système d'organisation des connaissances, le concept de patrimoine utilisé pour l'étude et sa modélisation et enfin la justification de l'utilisation du modèle ontologique du CIDOC CRM (Doerr 2003). Dans une deuxième partie, nous présentons notre méthodologie en deux étapes ; tout d'abord, la constitution d'un corpus (non exhaustif) du domaine minier constitué d'instances contextualisées pour peupler notre ontologie. Nous décrivons ensuite notre approche de sélection des entités candidates du modèle CIDOC CRM, basée sur l'analyse des définitions du patrimoine de l'UNESCO et du TICCIH au regard de notre domaine (les instances contenues dans notre corpus). Dans une troisième partie, les résultats valident et enrichissent l'approche par une illustration précise issue du patrimoine minier puis nous concluons avec quelques pistes de perspectives.

1. ANR-16-CE38-0001 : https://anr.fr/Projet-ANR-16-CE38-0001. Les recherches présentées dans ce chapitre sont partiellement financées par le projet ANR.

Cadre théorique

L'ontologie comme SOC

5. Notre recherche s'inscrit dans le cadre de l'organisation des connaissances telle qu'elle est définie par Claudio Gnoli (2012, 52). Pour cet auteur, le champ d'étude de l'organisation des connaissances se subdivise en quatre niveaux : la théorie, les systèmes, la représentation et les applications. Notre intérêt porte sur les systèmes d'organisation des connaissances (SOC). Ceux-ci sont des vocabulaires plus ou moins structurés, considérés par Gail Hodge comme des dispositifs d'organisation de l'information et se plaçant ainsi au cœur de toute bibliothèque, musée ou service d'archives[2]. Pour Manuel Zacklad, les SOC sont de différentes natures et regroupent les « langages documentaires, les schémas de classification [..., les] langages de représentation des connaissances issus de l'intelligence artificielle » mais aussi les « index de moteurs de recherche » (Zacklad 2010, 135). Dans le cadre du projet *Mémo-Mines*, nous avons créé deux types de SOC, appartenant à la catégorie des langages documentaires, correspondant aux objectifs suivants :

1. Une terminologie pour la mise à disposition d'un langage commun entre les acteurs sur la base de 33 ressources lexicales du domaine (Daloz 2018) et actuellement riche de 2400 termes intégrant également le dialecte des anciens mineurs du Nord–Pas-de-Calais
2. Un thésaurus pour la désambiguïsation du langage naturel, l'indexation de tout document du domaine et l'accès à ceux-ci. Il est actuellement composé d'environ 500 termes descripteurs et a été créé sur la base de la terminologie et de différentes structures classificatoires du domaine (Daloz et Chaudiron 2019)

6. Enfin, une ontologie de domaine, appartenant à la catégorie des langages de représentation des connaissances, est en voie de réalisation. Contrairement aux ontologies de haut niveau qui contiennent des concepts très généraux, elle est centrée sur un domaine particulier, ici, le domaine houiller du territoire du Nord-Pas-de-Calais et représente l'approche patrimoniale que nous avons de celui-ci. La représentation se fait à l'aide de concepts qui sont des catégories et des propriétés les reliant entre elles. Le domaine en question se compose d'instances de ces concepts, où l'instance correspond à un élément spécifique du domaine étudié ; cela peut être une entité nommée*, un terme. Par exemple, le concept « édifice » peut être instancié par le « Chevalement du Vieux 2, de la Compagnie des Mines de Marles[3] », et celui d'« outil » par « rivelaine[4] ».

2. « *Because knowledge organization systems are mechanisms for organizing information, they are at the heart of every library, museum, and archive* » (Hodge 2000, 9).

3. Cf. http://www.bassinminier-patrimoinemondial.org/mediation/le-chevalement-du-vieux-ii-de-marles-les-mines, consulté le 28/02/2020.

4. Cf. https://andredemarles.skyrock.com/3036489017-La-rivelaine-outil-du-mineur-haveur-vers-1900.html, consulté le 28/02/2020.

7. L'ontologie de domaine peut avoir plusieurs rôles au-delà de la mise en relation des concepts. Nous énumérons ceux-ci ci-dessous dans une liste non-exhaustive :
   - valoriser (sémantiquement) des objets, à partir d'une indexation par concepts plutôt que de termes
   - participer à l'uniformisation d'un langage
   - mettre en place un cadre de référence pour favoriser la communication entre les différents acteurs du domaine
   - inférer des connaissances

8. Dans le cadre du projet *Mémo-Mines*, l'ontologie produite servira dans un premier temps à modéliser les connaissances patrimoniales du domaine défini, c'est-à-dire à représenter par un modèle notre point de vue scientifique sur ce domaine précis. L'uniformisation du langage et la mise à disposition de cette ressource au plus grand nombre devront participer à l'amélioration de la visibilité des éléments de patrimoine. Les résultats produits dans ce chapitre s'inscrivent dans une première étape de validation méthodologique de constitution ontologique.

Modéliser le patrimoine

9. Afin de délimiter le périmètre de ce que nous appelons le patrimoine minier, nous nous appuyons sur les « éléments de cadrage » du patrimoine donnés par les institutions suivantes :
   - pour le patrimoine culturel et naturel, la *Convention concernant la protection du patrimoine mondial culturel et naturel* de l'UNESCO (1973) et son enrichissement définitoire dans la *Déclaration de Mexico sur les politiques culturelles (1982)*
   - pour le patrimoine culturel immatériel, la *Convention pour la sauvegarde du patrimoine culturel immatériel* de l'UNESCO (2003)
   - pour le patrimoine industriel, la *Charte Nizhny Tagil pour le patrimoine industriel* du TICCIH (2003)

10. Par « éléments de cadrage », et non « définition » comme cela est mentionné dans les différentes conventions, nous soulevons deux points. Le premier concerne la nécessaire interprétation non restrictive des définitions, c'est-à-dire que l'UNESCO laisse une part d'interprétation dans la mise en œuvre de la patrimonialisation. Ceci se justifie par le fait que, comme énoncé dans le cadre du patrimoine immatériel notamment, ce sont « les communautés, les groupes et, le cas échéant, les individus[5] » qui reconnaissent ce qui fait partie de leur patrimoine. Le deuxième se rapporte au caractère nécessaire et suffisant des définitions, c'est-à-dire que le patrimoine d'un peuple n'est pas défini strictement par toutes les conditions énoncées dans les définitions. Pour ces raisons, nous n'utilisons pas le terme de défi-

5. Cf. la *Convention pour la sauvegarde du patrimoine culturel immatériel* de l'UNESCO (2003, 2).

nition pour désigner ce que nous considérons en fait comme un cadre posant des limites.

11. L'utilisation de ces éléments de cadrage se justifie par le statut international et juridique de leurs auteurs, par leur prise en compte dans la construction du modèle ontologique du CIDOC CRM et enfin par leur nécessaire mention dans tout projet visant la sauvegarde d'un patrimoine culturel ou industriel en danger.

12. Par rapport à ces cadrages de la notion de patrimoine, nous avons sélectionné des sources primaires (des archives de la Mission Bassin minier, des œuvres littéraires, des articles de presse, des témoignages d'acteurs, des productions audiovisuelles) et des sources secondaires (des ressources lexicales définissant les termes techniques employés lors de l'activité minière et des plans de classement). L'ensemble de ces ressources constitue le matériau à partir duquel et sur lequel l'ontologie de domaine est construite. Précisons enfin que par « ontologie de domaine », nous entendons le domaine du patrimoine minier (exploitation du charbon) circonscrit au territoire des départements du Nord et du Pas-de-Calais en France et sur une période allant du début de l'exploitation (vers 1750) jusqu'à nos jours avec par exemple la réhabilitation des anciens sites de production et des habitats liés à la mine.

Le CIDOC CRM

13. L'arrivée du Web sémantique* dans le cadre des activités du W3C, organisme de normalisation du Web, a donné la possibilité aux structures patrimoniales de penser et manipuler autrement le patrimoine. Outre apporter une nouvelle contextualisation des données, les modèles ontologiques tendent à rassembler les connaissances sur un espace unique, avec un langage commun, pour permettre aux différentes structures patrimoniales de requêter leurs bases de données souvent hétérogènes. Ainsi, des modèles ontologiques tels que le CIDOC CRM ou le FRBR (Le Bœuf 2013) et leur migration vers FRBRoo (Doerr, Bekiari et Le Bœuf 2008) tendent à permettre, avec des objectifs différents, l'un centré sur les objets muséaux, l'autre sur les notices bibliographiques des bibliothèques, « l'intégration de n'importe quel échantillon de savoir, qu'il s'agisse d'un document archivé, de spécimens naturels ou d'artefacts exposés dans un musée, les galeries d'art, les expositions, et peut-être même les organisations qui traitent de l'information thématique. » (Gnoli 2012, 53).

14. Dans le cadre du projet *Mémo-Mines*, le modèle du CIDOC CRM a été choisi pour les six raisons énoncées ci-après. Premièrement, il s'agit d'une norme stable et mature, dont la version officielle (disponible en français et en anglais) est un standard ISO (ISO 21127), version ayant été déjà mise à jour une fois (publiée en 2006 puis mise à jour en 2014). Deuxièmement, le modèle permet une manipulation facilitée par la mise à disposition d'outils et d'une

documentation détaillée[6]. Troisièmement, le modèle a été construit sur la base d'une centaine de formats de métadonnées, y compris les informations patrimoniales UNESCO, nous verrons d'ailleurs plus bas comment retrouver les traces de cette organisation dans le modèle. Quatrièmement, le modèle possède une implémentation déjà effective dans les langages en vigueur, notamment OWL* pour sa version officielle, qui permet de faire des inférences, c'est-à-dire de déduire des nouvelles connaissances qui se réalisent par de nouvelles relations entre les objets. Cinquièmement, un *mapping* vers le modèle FRBRoo*[7], son successeur, et pour l'instant toujours en cours d'évolution, est déjà effectué. Enfin et sixièmement, son développement est spécifiquement prévu pour gérer des informations issues de sources hétérogènes, ce qui dans notre cas, est un critère positif pour le balayage le plus exhaustif possible du domaine.

15. Pour illustrer l'utilisation du modèle comme dispositif de médiation, nous présentons une requête aléatoire sur l'interface ResearchSpace[8] dédiée à la manipulation des fonds patrimoniaux du Musée de l'histoire et de la culture humaine à Londres, le British Museum. Ici, la médiation est prise au sens de « procédé de communication et de transmission qui utilise un ou plusieurs intermédiaires, [et permettant] de rendre accessibles des informations par différents processus de codage-décodage » (Azémard 2013, 124), avec comme intermédiaire, une interface utilisateur et les processus de codage-décodage, le modèle ontologique du CIDOC CRM.

16. L'interface ResearchSpace (figure 1) permet à tout utilisateur d'interroger les données des sept millions d'objets faisant partie de la collection du musée, à l'aide des entités et des propriétés sélectionnées à partir du modèle. Ainsi, les objets en bois d'acacia trouvés ou acquis en Égypte peuvent être retrouvés par la requête suivante : « *Thing – has material type – Concept: acacia wood AND Thing – found or acquired at – Place: Egypt* » (figure 2), où « *Thing* », « *Concept* » et « *Place* » sont les classes, « *acacia wood* » et « *Egypt* » sont les instances de ces classes et « *has material type* » ou « *found or acquired at* » sont les relations reliant sémantiquement le tout. Le fruit de cette requête consiste en quatre-vingt-deux résultats affichés sous la forme de petites vignettes (figure 3) présentant en un clic les objets dans leur contexte (titre, description, type d'objet, identifiant, matériau, dimensions, documentation, période de production, date d'acquisition, lieu d'extraction ou de trouvaille, propriétaire actuel, etc.). Une partie graphique (figure 4) rend visuellement et statistiquement compte de la comparaison entre la requête et l'ensemble des données du musée, en fonction de plusieurs propriétés du modèle. Les possibilités d'analyse des données illustrées par cette application permettent de faire émerger de nouvelles pistes d'utilisation et d'analyse de ses propres données. Enfin, le silence de certaines requêtes est une donnée à part entière. Celui-ci peut permettre de favoriser la communication entre usagers et struc-

6. Cf. http://www.cidoc-crm.org, consulté le 28/02/20.
7. Le *mapping* est un langage qui permet de mettre en correspondance (plus ou moins automatiquement) les instances de plusieurs ontologies.
8. Cf. https://public.researchspace.org/resource/Start, consulté le 28/02/2020.

Figure 1. Interface d'accueil ResearchSpace du British Museum
© BritishMuseum https://public.researchspace.org/resource/Start

Figure 2. Requête conceptuelle sur l'interface ResearchSpace du British Museum
© BritishMuseum https://public.researchspace.org/resource/Start

Figure 3. Échantillon des 82 résultats de la requête sur l'interface ResearchSpace du British Museum
© BritishMuseum https://public.researchspace.org/resource/Start

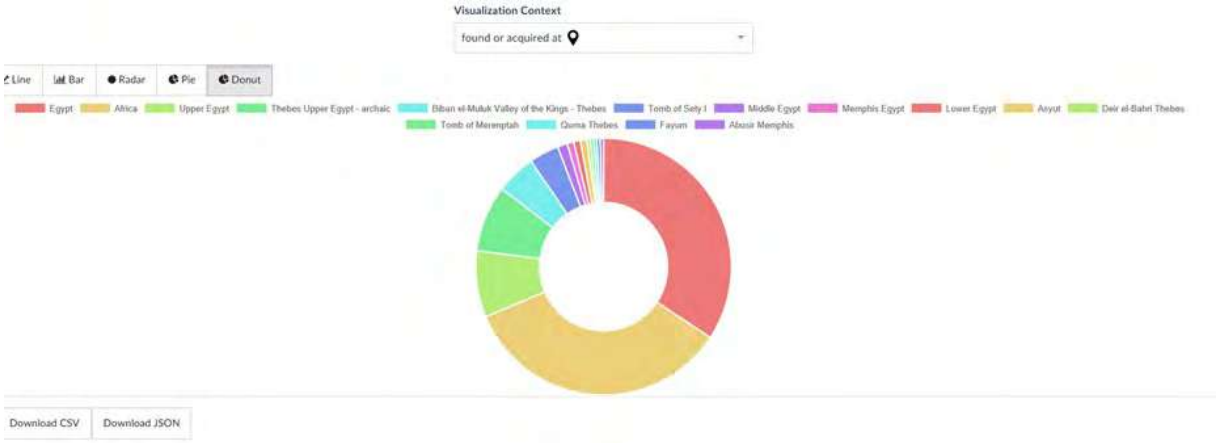


Figure 4. Diagramme résultant de la requête sur l'interface ResearchSpace : lieu de la découverte ou de l'acquisition des objets en bois d'acacia du British Museum
© BritishMuseum https://public.researchspace.org/resource/Start

tures patrimoniales en mettant au jour les besoins des utilisateurs, tout en pointant les lacunes dans certains domaines ou sur certains objets.

17. Suite à cette rapide revue théorique concernant l'ontologie comme SOC, notre cadrage patrimonial et l'intérêt du modèle du CIDOC CRM, nous présentons notre méthodologie de constitution de l'ontologie.

Méthodologie de constitution de l'ontologie

18. La constitution d'une ontologie peut être effectuée selon différentes approches. Notre méthode peut être qualifiée de mixte, dans le sens où l'approche est tantôt ascendante, tantôt descendante (en anglais *bottom-up* et *top-down*). L'approche ascendante fait référence à la relève des instances spécifiques au patrimoine minier dans le corpus et à leur conceptualisation, tandis que l'approche descendante part des concepts du modèle ontologique, étudié préalablement dans son ensemble, pour indexer les instances relevées. Les deux approches sont effectuées en parallèle. La figure 5 illustre cette méthodologie. Elle distingue les instances, des classes, des concepts. La différence entre la classe et l'instance est que l'instance est unique tandis que la classe peut avoir plusieurs instances. La différence que nous faisons entre la classe et le concept est que la classe appartient au domaine tandis que le concept appartient au modèle ontologique. Sur la même figure, les feuilles correspondent au corpus dont nous décrivons la constitution dans la partie suivante.

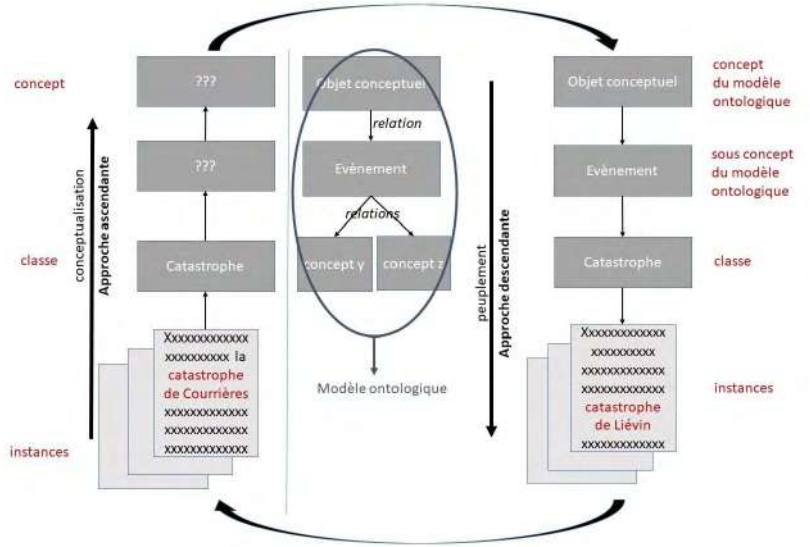


Figure 5. Approche ascendante-descendante
Crédit : Amélie Daloz

Constitution du corpus

19. Le critère d'éligibilité des documents à intégrer notre corpus correspond à la volonté de créer une ontologie de domaine. Tout document concernant le domaine houiller du territoire du bassin minier du Nord–Pas-de-Calais est éligible à intégrer notre corpus. L'objectif est de relever un maximum d'instances du domaine, pour obtenir une saturation conceptuelle, c'est-à-dire que toutes les nouvelles instances identifiées dans les documents produits ou identifiés dans le cadre du projet doivent pouvoir être indexées par les classes sélectionnées du

modèle ontologique. Le corpus est constitué de deux types de documents : des articles de presse et des captations audiovisuelles (témoignages d'anciens mineurs, de collectionneurs, de passionnés…).

20. Concernant le corpus presse, la période de parution des articles prise en compte est celle de l'après-mine, de 1995 à 2018. Un total de 542 articles provenant des journaux cités ci-dessus ont été scannés puis océrisés*. Afin de caractériser le corpus presse, nous listons ci-après les thématiques récurrentes principales, notamment identifiées pour la création du thésaurus du domaine mentionné plus haut : anniversaire de la fin de l'exploitation minière ; immigration dans les mines ; histoire ; classement UNESCO ; reconversion ; vacances des mineurs ; transport ; musique ; catastrophes ; luttes syndicales ; habitat minier ; séries TV liées à la mine ; sport ; régime social des mineurs ; témoignages ; objets et anecdotes ; archives du monde du travail ; bâti minier ; filmographie ; et tourisme.

21. Pour chacune de ces thématiques, nous avons identifié manuellement dans les articles tous les termes correspondant aux instances ou entités nommées (noms de lieux, d'organisations…). Nous avons ensuite procédé de la même façon sur le corpus audiovisuel. Celui-ci est composé d'une vingtaine d'heures de témoignages, correspondant à des récits de vie et des visites sur site de dix anciens mineurs respectivement captées en 2013 par la communauté d'agglomération de la Porte du Hainaut (CAPH) puis en 2018 par l'équipe du projet *Mémo-Mines*.

22. En ce qui concerne les thématiques, les entretiens portent sur la terminologie spécialisée ; la signification de la pratique du métier de mineur avec l'explication des techniques spécifiques, des machines et des outils ; l'habillement ; les lieux ; et les risques du métier. Mais ces entretiens portent également sur ce qui relève du patrimoine immatériel, à savoir les coutumes, l'histoire personnelle, les émotions, les évènements marquants, les expressions patoisantes ainsi que l'importance de la sauvegarde du patrimoine et des sites en danger. Dans ce cas, pour un accès plus aisé au corpus, les termes spécifiques au domaine minier (venant enrichir directement notre terminologie) et les entités nommées ont été transcrits. Ces deux premières étapes participent à la représentation du patrimoine minier dans une approche dite ascendante, c'est-à-dire en partant des instances présentes dans le corpus vers la conceptualisation.

Modélisation

23. La seconde approche consiste à s'approprier les concepts et les propriétés du modèle ontologique du CIDOC CRM à partir de la lecture de la version officielle 5.0.4 (Crofts *et al.* 2011) en anglais et de sa version non officielle française (Crofts *et al.* 1999). Nous nous sommes également inspirée des travaux d'Éric Kergosien *et al.* (2015) et de Kaouther Ben Smida (2016) qui ont utilisé le même modèle ontologique dans le cadre du patrimoine textile. La méthode consiste ensuite à lister toutes les entités du modèle susceptibles de cor-

respondre aux définitions de catégories sélectionnées par l'UNESCO et du TICCIH puis de les aligner avec les instances du corpus. Cette approche est qualifiée de descendante dans le sens où l'on part des concepts les plus génériques en allant vers les plus spécifiques jusqu'à rejoindre les instances. Les résultats sont présentés et discutés dans la partie suivante.

Résultats et discussion

24. Les résultats illustrent l'approche énoncée ci-dessus en donnant un exemple précis issu du patrimoine minier. Dans cet exemple, une douzaine d'entités du modèle ontologique du CIDOC CRM ont été sélectionnées.

Connaissances explicites ou implicites des corpus : le cas de la Sainte-Barbe

25. La sainte Barbe est la patronne des mineurs. En considérant les définitions que donne l'UNESCO de la notion de patrimoine, la sainte Barbe se réfère à plusieurs types de patrimoines.

26. En tant que monument de type sculpture, elle fait partie du patrimoine culturel matériel, comme par exemple dans un article de *La Voix du Nord* du samedi 25 décembre 2010. En tant que représentation (ici d'une sainte, d'un symbole), elle s'inscrit dans le champ du patrimoine immatériel. L'article mentionne en effet, en légende d'une photo : « Mineurs remontant la statue de leur patronne sainte Barbe lors de la fermeture de la fosse 4/5 de Méricourt, en 1982 ». Dans un autre article, celui du dimanche 19 décembre 2010 du même journal, la mention d'une autre statue de sainte Barbe justifie le fait de prendre en compte cet objet comme faisant partie du patrimoine : « Comme tout le monde je me suis arrêtée devant la sainte Barbe à la sortie de la cage, pour lui demander d'avoir un œil sur notre descente. La statue était toute propre, derrière sa vitre, les mineurs y veillaient ».

27. L'analyse de ces passages textuels fait émerger un autre type de patrimoine immatériel, une connaissance, en partie implicite, qui est la suivante : il existe plusieurs statues de sainte Barbe, qui sont des représentations d'une sainte. Cette sainte est considérée comme la patronne des mineurs. Les statues étaient généralement situées au fond de la mine lors de l'activité minière et sont des symboles physiques de protection des mineurs.

28. En tant que patrimoine immatériel, la Sainte-Barbe se manifeste également comme un évènement festif. L'article du 25 décembre 2010 et des entretiens, avec Jacques Potier, ancien mineur (cf. extraits 1 et 2 ci-dessous), font émerger explicitement ou implicitement quatre autres connaissances sur cet évènement festif. Celles-ci sont à la fois de type temporel : la Sainte-Barbe est fêtée tous les ans, le 4 décembre ; de type légende : la Sainte-Barbe est une légende basée sur la vie d'une sainte du nom de Barbara ; de type rituel : l'évènement festif est partagé avec

d'autres corps de métiers comme les pompiers, les artilleurs ; et enfin cet évènement est composé d'un rituel de type bal et de musique traditionnelle polonaise.

- Extrait 1 de l'article de *La Voix du Nord* du 25 décembre 2010 : « [...] la tradition la plus forte et spécifique à la mine était la Sainte-Barbe fêtée le 4 décembre. Basée sur la légende de la vie de Barbara, érigée en patronne protectrice de "tout ce qui tonne et détone" : pompiers, artilleurs et mineurs. »
- Extrait 2 de l'entretien de Jacques Potier [37 : 29] : « [...] et tous les ans, par contre ce qui continue, c'est notre bal de Sainte-Barbe, ce bal de Sainte-Barbe il est sacré. À Sainte-Barbe, on fait la fête avec un orchestre polonais et on est bien en arrivant mais on n'est pas bien quand on s'en va. »

Modélisation ontologique des connaissances extraites

29. La figure 6 ci-après récapitule à l'aide d'un schéma ontologique le lien entre les éléments de cadrage UNESCO et certaines instances du corpus liés aux multiples manifestations de la Sainte-Barbe.

30. Les relations sont exprimées en encadré blanc tandis que les éléments de cadrage de l'UNESCO sont marqués « UNESCO ». Les connaissances apportées par les corpus correspondent au reste des encadrés (classes et instances). On peut par exemple y « lire » que l'instance « Sainte-Barbe » s'apparente à la classe « statue » qui elle-même s'apparente à la classe « sculpture » telle que la définit l'UNESCO qui, à son tour, relève de la classe « monument » qui fait partie de ce que l'UNESCO définit comme « patrimoine culturel matériel ». Il en est de même pour le « patrimoine culturel immatériel » qui se manifeste par des « évènements festifs » et qui est composé par des « représentations » et des « connaissances ». À partir de l'exemple montré dans la figure 6, nous constatons que l'enjeu est de définir les classes et les propriétés du modèle permettant de décrire les instances telles qu'elles « fonctionnent » sémantiquement dans le corpus, tout en tenant compte du cadrage de la notion de patrimoine fournie par l'UNESCO.

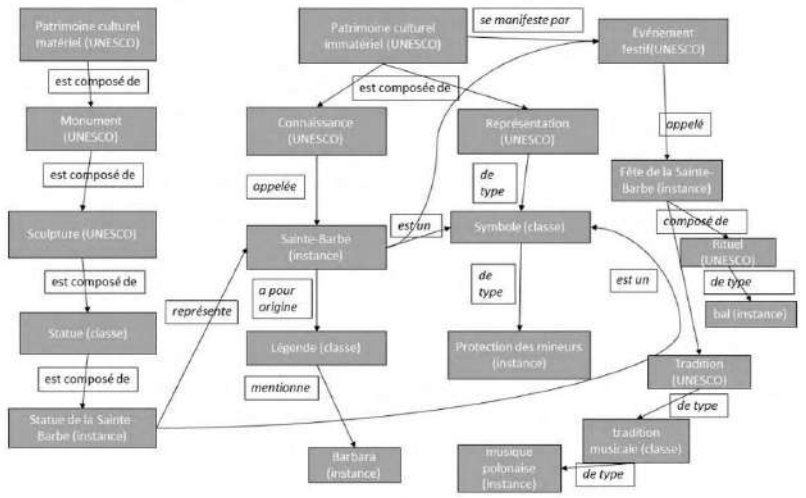


Figure 6. Illustration du schéma ontologique pour la Sainte-Barbe
Crédit : Amélie Daloz

31. La figure 7 présente des exemples des classes du CIDOC CRM qui correspondent aux éléments de cadrage de l'UNESCO.

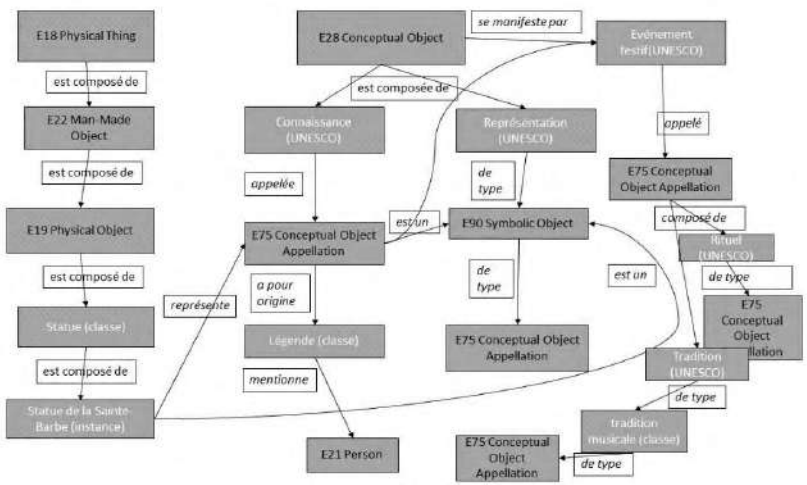


Figure 7. Alignement de l'ontologie du patrimoine minier avec les entités du CIDOC CRM

Crédit : Amélie Daloz

32. Ainsi, les classes indiquées en noir sur fond gris dans les encadrés (de type « *E18 Physical Thing* ») désignent celles qui expriment les réalités patrimoniales issues des éléments de cadrage. À l'inverse, les encadrés gris en lettres blanches, de type « Connaissance (UNESCO) », désignent les réalités patrimoniales listées par l'UNESCO qui n'ont pas de correspondants immédiats dans le modèle ontologique. Cela s'explique par le caractère très générique des concepts du CIDOC CRM.

Conclusion

33. Comme nous l'avons montré dans ce chapitre, la construction du modèle ontologique CIDOC CRM pour le domaine minier (restreint selon la définition que nous en avons donnée précédemment) s'appuie sur une double démarche, ascendante et descendante. Cette approche permet tout à la fois de définir les classes et les propriétés du modèle utiles et nécessaires pour décrire le monde de connaissance défini dans le cadre du projet *Mémo-Mines* puis de peupler le modèle.

34. Cette approche contribue également à formaliser la notion de patrimoine minier. Alors que nous nous sommes initialement appuyée sur la définition du patrimoine donnée par l'UNESCO et le TICCIH, le choix des classes et des propriétés ainsi que des instances à représenter dans le modèle précise le domaine de connaissances à formaliser et, ce faisant, contribue à définir notre approche du patrimoine minier.

35. Les perspectives immédiates de la recherche concernent en priorité la poursuite de la construction de l'ontologie, à partir du corpus à notre disposition, mais aussi à partir du thésaurus du domaine minier que nous avons réalisé et qui sera très prochainement accessible au format SKOS. Ce thésaurus est un premier système d'organisation des connaissances du domaine de la mine et propose une modélisation que nous prévoyons d'intégrer dans l'ontologie.

36. L'exemple de la Sainte-Barbe illustre la complexité à construire une ontologie de domaine et souligne l'importance des choix qui sont opérés dans l'alignement entre les classes du modèle et les réalités patrimoniales. Construire une telle ontologie implique de procéder à des arbitrages qui, à terme, contribuent à façonner une vision spécifique de ce que nous entendons par domaine minier, et qu'il conviendra d'expliciter.

Augmenter et publier des corpus en ligne

Le programme des registres de la Comédie-Française : un corpus numérique en extension

Agathe Sanjuan

Introduction

1. La Comédie-Française, théâtre fondé en 1680, a toujours conservé ses archives, stockées sans grande précaution dans un premier temps, puis inventoriées et étudiées à partir du milieu du XIXe siècle, lorsqu'un service dédié s'est structuré autour du patrimoine théâtral de l'institution, la bibliothèque-musée de la Comédie-Française. Depuis les années 1960, un personnel formé s'est occupé du fonds et un conservateur d'État en est responsable depuis 1980. Les archives ont toujours été communiquées mais la salle de lecture, inexistante jusqu'en 1999, petite encore aujourd'hui, ne permettait certainement pas de donner à ces documents une audience à la hauteur de leur qualité d'information. Au cours des années 2000, l'arrivée du numérique et l'informatisation des collections au sein de la base de données La Grange, ont permis une diffusion plus importante de fonds jusque là réservés à un public confidentiel.

2. Le programme des registres de la Comédie-Française a débuté en 2008 par un premier projet centré sur les registres de recettes, grâce à une équipe de chercheurs motivée, internationale, familière des pratiques collaboratives – certains avaient déjà mis en ligne leurs corpus respectifs dans la base CÉSAR, pour Calendrier électronique des spectacles sous l'Ancien Régime – Christian Biet (université Paris-Nanterre), Georges Forestier (université Paris-Sorbonne), Pierre Frantz (université Paris-Sorbonne), Jeff Ravel (MIT) et Sara Harvey (post-doctorante à la Sorbonne). L'historien américain Jeff Ravel avait étudié ce corpus (Ravel 1999) et évalué sa richesse et son potentiel. S'associant à des historiens du théâtre et à des historiens de la littérature, il a alors proposé à la Comédie-Française de collaborer avec son laboratoire d'humanités numériques au sein du MIT, Hyperstudio, pour se pencher sur cette source somme toute peu exploitée en regard de ses possibilités.

3. Les acteurs de ce programme, de par leurs statuts et leurs missions respectives, se trouvaient avoir des objectifs différents : pour la bibliothèque-musée primait la mission fondamentale de sauvegarde du patrimoine par sa numérisation, pour les chercheurs celle de la facilitée d'accès à la ressource et de son exploitation quantitative facilitée. Enfin, le théâtre de la Comédie-Française lui-même s'est assez vite intéressé au programme dans la mesure où il permettait une approche renouvelée du répertoire historique, ce que les chercheurs eux-mêmes appelaient

de leurs vœux[1]. Ces deux types de publics ont pu faire converger leurs intérêts : d'une part les utilisateurs scientifiques qui ont pu, grâce au programme et à ses outils, accélérer et faciliter les dépouillements quantitatifs et faire émerger des hypothèses qualitatives difficiles à obtenir auparavant, d'autre part les praticiens, comédiens et le grand public, attirés par une histoire plus factuelle permettant d'expliquer les ressorts dans le temps de ce répertoire de plus de 3 000 pièces qui semblaient parfois inexplicables. Le spectre de l'historiographie se trouvait donc balayé dans ses deux extrêmes macro et micro-historique, en regard des publics concernés. Avec le temps et l'augmentation du programme, cette profondeur de l'analyse s'est trouvée augmentée, matériellement, de l'ajout de corpus documentaires complémentaires. L'extension du programme agit donc dans toutes ses dimensions, bien au-delà de ce qui était prévisible il y a dix ans.

Les contours du programme

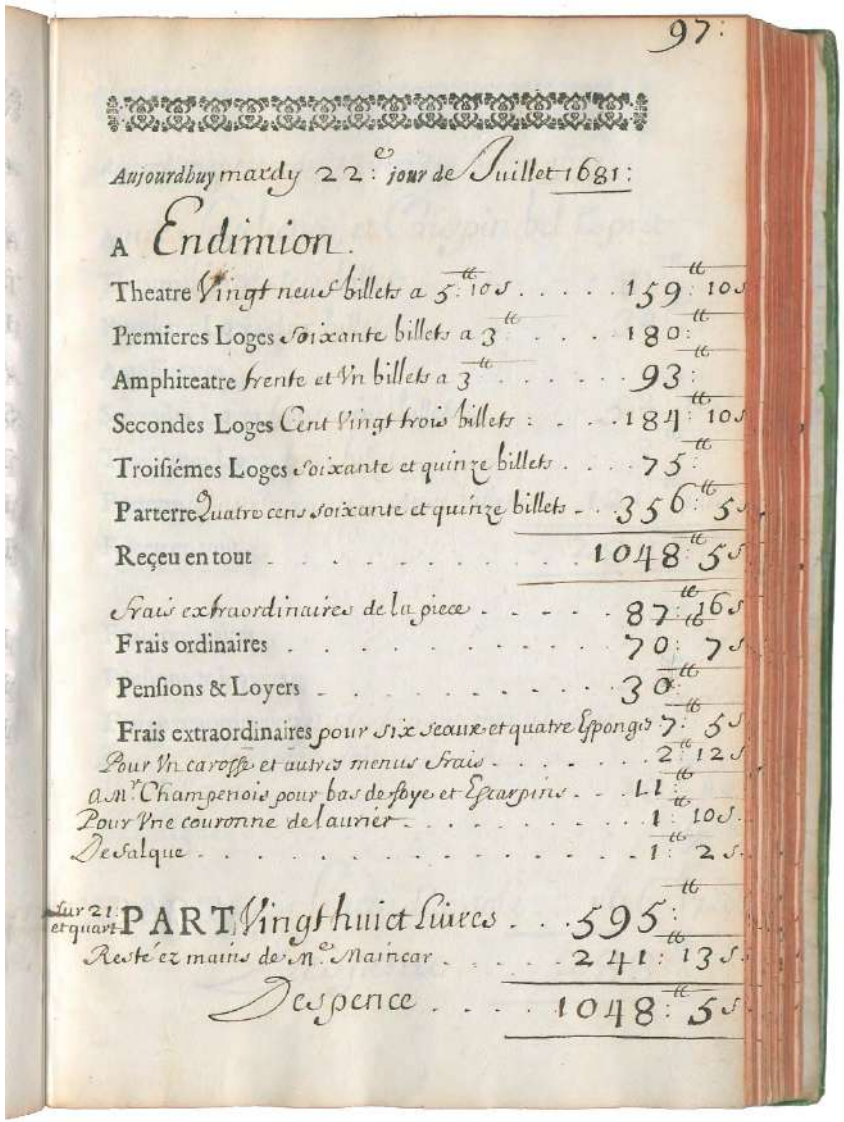
4. Le *Projet des registres de la Comédie-Française* (*PRCF*) s'est développé en trois phases correspondant à des domaines d'intérêt des différents chercheurs associés, en lien avec les financements apportés.

1. La Comédie-Française, dotée d'une troupe permanente, est chargée de mettre en œuvre l'enrichissement de son répertoire ainsi que le maintien de son répertoire ancien par les représentations qu'elle en donne (décret n° 95-356, du 1er avril 1995, modifiant le régime administratif de la Comédie-Française, conférant à la Comédie-Française le statut d'établissement public national à caractère industriel et commercial, article 2).

Premier programme (*PRCF1*) : recettes, programmation, publics (1680-1793)

5. Le corpus d'archives comprenait environ 140 registres journaliers manuscrits (figure 1) portant sur 113 saisons théâtrales et plus de 30 000 soirées de représentation, sur la période 1680-1793 – de la fondation de la Comédie-Française à sa fermeture sous la Révolution. Pour chaque saison théâtrale, la troupe tenait un registre destiné à enregistrer la comptabilité journalière indiquant la date, le titre des pièces jouées, le nombre de places vendues dans chaque catégorie de place (théâtre, parterre, loges), la recette du jour et les dépenses du jour (fig. 1).

6. Au sein de ce corpus papier, manquant de forces et de moyens, nous avons dû restreindre le corpus de données aux seules recettes du théâtre. Cette catégorie de données avait l'avantage d'être relativement stable et complète sur la période considérée. Si ce premier ensemble pouvait nous donner des informations précieuses sur les stratégies de programmation, le succès des pièces, des auteurs, ainsi des outils pour étudier l'affluence, et donc permettre une première approche du public, nous ne pouvions raisonnablement envisager d'aborder les questions purement économiques sans avoir en machine les données liées aux dépenses du théâtre. Il a néanmoins fallu nous contenter de ce corpus de données – déjà important – dans un premier temps, à savoir 120 000 images et 700 000 données.


97:

Aujourdhuy mardy 22e jour de Juillet 1681:

A Endimion.

Theatre Vingt neuf billets a 5tt 10s 159tt 10s
Premieres Loges Soixante billets a 3tt . . . 180tt
Amphiteatre trente et Vn billets a 3tt 93tt
Secondes Loges Cent Vingt trois billets . . . 184tt 10s
Troisiémes Loges Soixante et quinze billets . . 75tt
Parterre Quatre cens Soixante et quinze billets . 356tt 5s
Reçeu en tout 1048tt 5s

Frais extraordinaires de la piece 87tt 16s
Frais ordinaires 70tt 7s
Pensions & Loyers 30tt
Frais extraordinaires pour Six seaux et quatre Eponges 7tt 5s
Pour Vn carosse et autres menus frais 2tt 12s
a Mr Champenois pour bas de soye et Escarpins . 11tt
Pour Vne couronne de laurier 1tt 10s
De Salque 1tt 2s

Sur 21 et quart PART Vingt huict Sieres . . 595tt
Resté es mains de Mr Maincar 241tt 13s
Despence 1048tt 5s

Figure 1. Registre journalier, 22 juillet 1681

© Comédie-Française https://flipbooks.cfregisters.org/R13/index.html#page/197/mode/1up.

7. Ce premier programme[2] a été riche d'enseignements quant à la méthode adoptée (Harvey, Sara et Agathe Sanjuan 2016) : collaborations pluridisciplinaires, constat réaliste des limites matérielles à prendre en compte, compromis sur les outils à développer en fonction des intérêts divergents des partenaires. Les photographies et les données ont été rassemblées et exploitées via un site internet bilingue doté de trois outils de recherche principaux[3].

Deuxième programme (*PRCF2*) : dépenses, feux, assemblées, critique théâtrale (1680-1793)

8. L'objectif était bien sûr d'augmenter ce corpus de celui des dépenses[4], que l'on pouvait relier directement au corpus de départ, puisque, dans les deux cas, l'approche est journalière. Pour un même corpus d'archives, nous avons donc été conduits à augmenter le corpus de don-

2. *PRCF1* a bénéficié d'un financement majeur de la part de l'ANR (2013-2016), porté par l'université Paris-Nanterre, mais aussi du Programme national de numérisation du ministère de la Culture et de la Communication, le Labex de Paris Ouest-Nanterre (Les Passés dans le présent), le Labex de Paris 8 et Paris Ouest-Nanterre (Arts-H2H), l'Institut universitaire de France, le laboratoire Idefi CréaTIC, aux États-Unis par le Massachusetts Institute of Technology (MIT, Boston), l'université Harvard, la Florence Gould Foundation, Gladys Krieble Delmas Foundation et par le financement franco-américain Partner University Fund – Face.
3. Cf. https://www.cfregisters.org/fr/
4. Projet porté par l'université Paris-Nanterre.

nées dans un second temps, au cours du deuxième programme (*PRCF2*[5]).

9. La réflexion sur l'augmentation du corpus de données a amené à considérer que d'autres corpus de documents devaient être intégrés, comme étant indispensables à l'interprétation des premiers, notamment les registres de feux[6] (figure 2) donnant les distributions journalières des acteurs – le « feu » étant l'indemnité versée au comédien quand il joue pour chauffer et éclairer sa loge. Donnée comptable que l'on retrouve au titre des dépenses, les feux permettent en outre d'étudier les carrières d'acteurs, les effets de vedettariat et l'incidence de la distribution sur la fréquentation du public. Les dépenses et les feux étaient donc en parfaite complémentarité avec les recettes : la notation est quotidienne et ces informations complètent et augmentent les premières.

10. Toutes ces données étaient intégrables dans une base de données classique, les catégories étant régulières et normalisables. Cependant, des annotations manuscrites apparaissaient au fil des pages, donnant des informations contextuelles de première importance, difficiles à intégrer dans cette structure rigide.

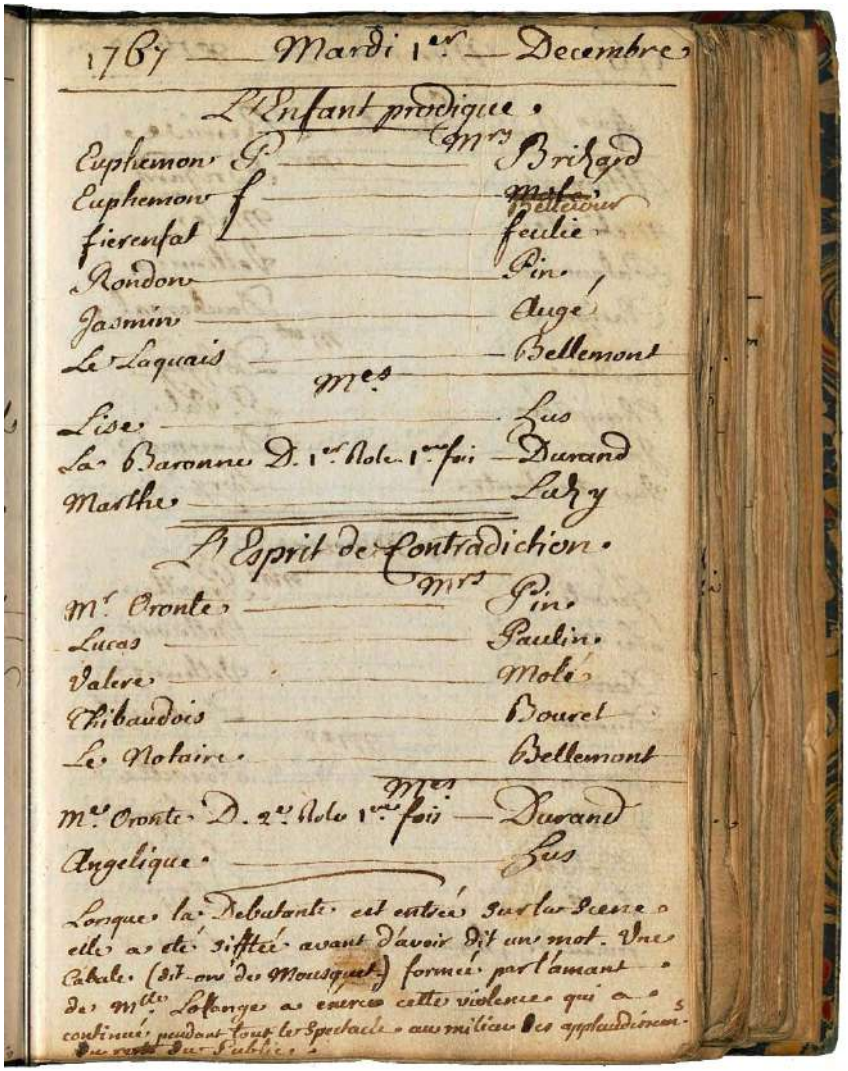


Figure 2. Registre de feux, 1er décembre 1767

© Comédie-Française https://flipbooks.cfregisters.org/R130_2/index.html#page/123/mode/1up.

5. Ce deuxième projet bénéficie actuellement d'un financement de la part de l'ANR (2019-2022) porté par l'université Paris-Nanterre, ainsi que le laboratoire Idefi Créa-TIC, l'université de Rouen, au Canada par le CRSH, aux États-Unis par le MIT et par le financement franco-américain Partner University Fund – Face.

6. Projet porté par l'université de Victoria et notamment Sara Harvey.

11. Le registre donne les distributions des deux pièces jouées, *L'Enfant prodigue* de Voltaire et *L'Esprit de contradiction* de Du Fresny, ainsi qu'une information contextuelle : « Lorsque la Débutante [M^lle^ Durand] est entrée sur la scène, elle a été sifflée avant d'avoir dit un mot. Une cabale (dit-on de Mousquet) formée par l'amant de M^lle^ Lalange a exercé cette violence qui a continué pendant tout le spectacle au milieu des applaudissements du reste du Public. »

12. Nous avons donc été amenés à considérer que le mode d'accès à l'information par la date de l'évènement pouvait être insuffisant. Pour ces notes marginales, non exploitables au moyen d'une base de données, deux possibilités s'offrent à nous actuellement. Les notes de nature régulière bien que non systématiques peuvent être intégrées à la base de données – par exemple la mention des débuts d'acteurs[7] ou encore certains achats concernant une typologie définie qui revient régulièrement. En revanche, les notations d'un contenu original et unique – telle la cabale occasionnée par le début de M^lle^ Durand – sont difficilement intégrables dans notre structure. La réflexion sur l'intégration d'un corpus textuel – telles ces notes marginales – nous a amenée à considérer encore d'autres corpus de documents en lien avec les premiers.

13. L'intégration au programme de la critique théâtrale[8] s'est alors imposée. Publiée dans des périodiques proposant des orientations politiques partisanes variées, relatant les premières représentations des pièces et les débuts d'acteurs, ce corpus étend les documents à des fonds extérieurs à ceux de la bibliothèque-musée de la Comédie-Française qui ne conserve pas ces collections dans leur entièreté. Ce corpus textuel sera indexé suivant les tables existantes (noms d'auteurs, d'acteurs, titres des pièces, dates) ainsi que suivant un traitement automatisé du langage (TAL*) orienté vers l'analyse de l'opinion et des sentiments, prisme majeur de la réception théâtrale par la critique et le public au XVIII^e^ siècle, et centre d'intérêt particulier de l'équipe en charge de ce corpus.

14. Ces derniers développements permettent d'imaginer que ce corpus de données sera amené à augmenter dans les années à venir car la critique des spectacles de la Comédie-Française peut aussi se lire dans les correspondances privées ou publiées (par exemple la *Correspondance littéraire* de Grimm), ou en regard de celle des autres salles parisiennes. De la même manière, nous envisageons d'augmenter le corpus des registres d'assemblées et de comités[9] qui donnent les comptes-rendus de réunions et les décisions prises par les instances de direction du théâtre. Ils apportent des informations abondantes sur le fonctionnement du théâtre, les relations avec les auteurs, les fournisseurs, la tutelle royale.

7. Le début d'un acteur est ordonné par l'administration royale et permet à l'impétrant d'interpréter les rôles principaux de son emploi pendant une dizaine de représentations.

8. Projet porté par l'université de Victoria et Sara Harvey.

9. Ce projet est en cours à l'université de Victoria pour la période 1765-1793.

Troisième programme (*PRCF3*) : recettes, dépenses, feux, assemblées, critique théâtrale (1799-1914)

15. Nos réflexions nous ont amenés naturellement à augmenter la période chronologique en abordant le XIXe siècle[10]. Le corpus est de même nature, mais les documents plus abondants et parfois plus complexes. La récurrence des catégories et leur régularité nous permettent d'envisager pour de larges pans de ces ensembles l'utilisation d'un logiciel de reconnaissance automatique de caractères manuscrits, les techniques s'étant perfectionnées et l'écriture étant plus régulière sur cette période[11].

16. Le programme des registres journaliers s'est donc ramifié au cours du temps en intégrant des institutions et des chercheurs, en multipliant les corpus documentaires autour du premier ensemble et en allant chercher dans les documents des données toujours plus précises et plus profondes.

Ce que le numérique fait au corpus

17. Les mutations du corpus par le biais du numérique changent radicalement l'approche scientifique. Le corpus n'est plus délimité comme autrefois dans sa dimension matérielle – la série des registres journaliers – mais s'envisage sous forme de corpus de données, exploitables via des outils existants ou des outils à construire[12]. Ce changement fondamental va de pair avec des techniques d'analyse renouvelées.

Calcul, macro-analyse

18. L'outil numérique influe sur l'appréhension globale du corpus en apportant une puissance de calcul nouvelle : auparavant, les chercheurs mettaient parfois des années à dépouiller les registres sur un sujet particulier (les succès de Voltaire, les cabales contre Marivaux, etc.). Désormais, quelques manipulations dans les outils de recherche suffisent à obtenir des résultats autrefois longs et fastidieux à établir. La recherche peut donc se reporter sur des sujets plus subtils que la stricte compilation de données, qui constitue en général le point de départ d'une réflexion plus qualitative. La recherche sur ces fonds a changé de nature et se concentre sur l'analyse fine du corpus, mettant en valeur, paradoxalement, les variations, anomalies, exceptions, alors que la macro-analyse est facilitée.

10. Le projet *PRCF3* est mené à la Sorbonne Université par Florence Naugrette, et financé en France par l'Institut universitaire de France, Sorbonne Université, l'université de Rouen, le Labex OBVIL (Sorbonne Université), l'Idex SUPER, la Fondation pour la Comédie-Française, et au Canada par le CRSH.

11. Des liens sont d'ores et déjà établis avec le projet *LECTAUREP* des Archives nationales. Cf. http://www.archives-nationales.culture.gouv.fr/l-intelligence-artificielle-et-le-patrimoine.

12. Les données brutes sont également disponibles.

Définition des données, analyse fine

19. L'outil numérique permet aussi d'analyser des catégories de données qui ne pouvaient l'être auparavant, par leur complexité. La catégorie des dépenses permet de l'expérimenter, au niveau même de la saisie des données, et donc de la construction du futur ensemble à exploiter. Certaines catégories d'information changent de nom, ne recouvrent pas tout à fait la même réalité au cours du temps et seules les comparaisons massives, sur la longue durée, ainsi que le recours à des sources extérieures permettent de l'identifier. Les frais courants du théâtre sont successivement nommés « frais ordinaires » et « frais du jour ». Certaines catégories sont autonomisées au cours du temps (frais d'affiches, frais pour la garde) alors qu'ils sont intégrés aux frais ordinaires à d'autres périodes. Dans le cadre de cette analyse nécessaire pour la saisie ainsi que pour l'exploitation future, les données vont donc être mieux définies et seront à même d'être interprétées avec plus de pertinence.

Extension à des corpus proches ou éloignés

20. Les corpus connexes à un domaine de recherche propre sont plus facilement accessibles dans le contexte des humanités numériques. Les données disponibles sont donc explorées, et parfois exploitées, par des chercheurs de domaines connexes, qui, a priori, ne se seraient pas tournés vers ces fonds. Pour notre projet, des historiens de l'économie (Velde 2017), du climat[13], se sont penchés sur des données qu'ils n'auraient pas eu idée de regarder auparavant, mais qui, mises à disposition, devenaient un terrain d'étude possible.

21. À la suite de notre projet, d'autres initiatives se sont montées et sont en cours concernant les autres théâtres parisiens, notamment le Théâtre-Italien[14]. L'interopérabilité* entre ces bases est un des objectifs, à terme. En attendant, le corpus de la Comédie-Française a été lié à l'iconographie et à la documentation que l'on trouve dans Gallica et dans la base La Grange, le catalogue de la bibliothèque-musée de la Comédie-Française, par le biais de moissonnages de ces catalogues dont certaines images sont intégrées à l'outil « découverte » du programme de la Comédie-Française[15]. Dans le contexte des humanités numériques, le corpus de données parait donc comme une entité mouvante et extensible.

Édition des corpus numériques

22. Les chercheurs impliqués dans le programme des registres de la Comédie-Française étaient représentatifs de générations et de cultures scientifiques hétérogènes. Face aux bouleversements apportés par l'approche numérique des

13. Travaux d'Emmanuel Garnier, non publiés à notre connaissance.
14. Projet *Recital* de l'université de Nantes, sous la forme d'une base de transcription participative : http://recital.univ-nantes.fr/#/.
15. Outil élaboré par Adrien di Mascio (alors employé par l'entreprise Logilab) : https://www.cfregisters.org/fr/nos-donn%C3%Ages/outil-de-base?q=en/the-data/basic-tool.

sources, la plupart d'entre eux ont éprouvé la nécessité de développer des outils propres au programme, à même de faciliter l'accès aux données.

Le choix des outils, un compromis nécessaire

23. Après l'alimentation de notre base de données, l'équipe du projet a donc travaillé à la conception d'outils de visualisation des données, qui peuvent être considérés comme des modes d'édition du corpus. Les données brutes restaient accessibles et téléchargeables, mais nous voulions aussi proposer un accès plus convivial. Les outils obtenus procèdent de compromis ménageant les intérêts des différents chercheurs et permettent d'explorer le corpus assez largement.

24. Deux outils ont été développés à destination des chercheurs : un « outil par facettes » (figure 3) qui sélectionne les représentations à l'aide de filtres (conception Hyperstudio, Jamie Folsom revu par Christopher York), et un « outil par graphe et calendrier dynamique » (figure 4 et 5) qui s'est avéré assez complexe à utiliser (Christopher York).

Figure 3. Outil à facettes
https://www.cfregisters.org/fr/nos-donn%C3%A9es/faceted-browser

Figure 4. Calendrier dynamique, vue 1
https://www.cfregisters.org/app

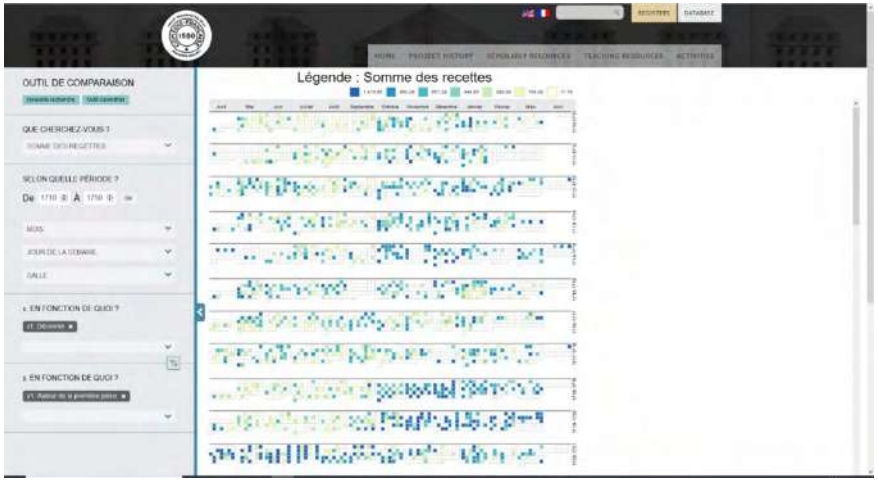


Figure 5. Calendrier dynamique, vue 2
https://www.cfregisters.org/app

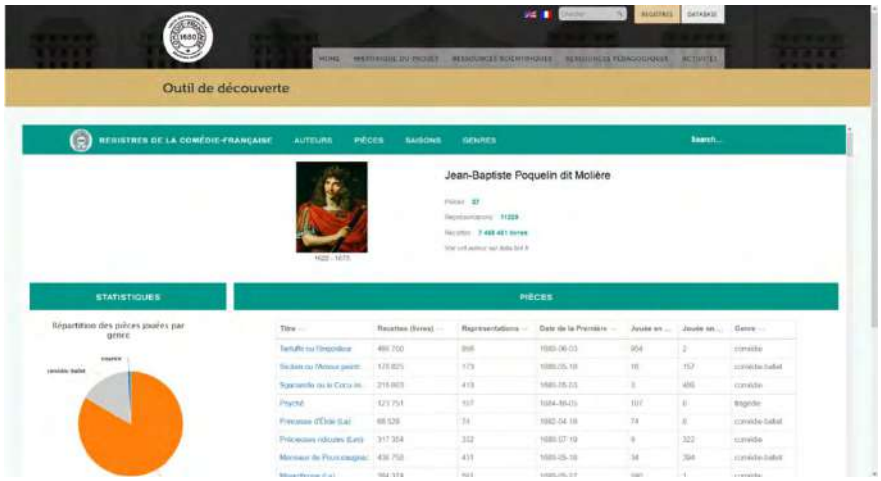


Figure 6. Outil de découverte
https://www.cfregisters.org/fr/nos-donn%C3%A9es/outil-de-base?q=en/the-data/basic-tool

25. Un outil à destination d'un public plus large (figure 6) a été proposé par un informaticien au cours d'un *hackathon* (Adrien Di Mascio de Logilab). Il est aujourd'hui le plus utilisé pour une première approche de la base de données et permet d'obtenir des statistiques globales sur le répertoire, les recettes et quelques visualisations simples.

26. L'apport des outils de visualisation est majeur : ils permettent de confirmer certaines hypothèses qui ne peuvent apparaître en observant simplement les données quantitatives. De ce fait, un effort particulier a été apporté au design de ces outils afin d'en améliorer la performance et l'efficacité visuelle[16].

Créer un environnement de savoirs : dictionnaire de données, pédagogie et accompagnement des corpus

27. Le programme développe actuellement un outil à destination du grand public, permettant d'avoir accès à des définitions, familières aux chercheurs mais peu connues du grand public. Un premier essai en ligne décrit les différentes salles successives occupées par la Comédie-Française sous l'Ancien Régime : ces courts essais permettent de resituer la physionomie des salles et les enjeux concernant les publics. D'autres entrées de dictionnaire sont actuellement en cours d'élaboration, par exemple la notion de « début d'acteur » : qui l'ordonne (la tutelle), quels enjeux (l'entrée d'un acteur dans la

16. Travail mené avec une école de webdesign de Montréal.

troupe, le renouveau ou non de l'interprétation, les rivalités et cabales, les jeux de protection de la cour), selon quel fonctionnement (dix représentations, pendant un an, pendant lesquelles l'acteur débutant a la priorité des grands rôles de son emploi), quelle influence sur la programmation (choix de certaines pièces, positionnement du public face au débutant). Depuis l'origine du projet, la dimension pédagogique et expérimentale a été intégrée au site lui-même comme pouvant permettre une approche du corpus. Tout travail pédagogique sur la base de données doit constituer une clé d'interprétation répétable sur tout ou partie du corpus.

Figure 7. Atelier Idefi CréaTIC 2019

Contextualisation d'une pièce de théâtre polémique, *Charles IX* de Marie-Joseph Chénier, 1789. Le travail des étudiants est une introduction à la pièce, sous une forme inventive, originale : ici une présentation des personnages (qui s'opposent en deux clans) et des acteurs qui les interprètent.

28. Le projet a bénéficié pendant plusieurs années du dispositif d'atelier Idefi CréaTIC – par le biais de Tiphaine Karsenti et Charlotte Bouteille-Meister (université Paris-Nanterre), Sylvaine Guyot (université Harvard) et Sara Harvey : des étudiants en arts du spectacle (Nanterre) et en littérature française (Harvard) ont travaillé de concert sur des dispositifs innovants mettant en œuvre un processus de recherche basé sur le projet (figure 7).

Vulgarisation et spectacle vivant : une forme d'édition éphémère

29. Ce processus de contextualisation du répertoire – par le biais de l'analyse historique possible grâce au programme des registres – fait partie intégrante de la programmation de la Comédie-Française depuis 5 saisons, à raison de 4 spectacles par an donnés une fois chacun (quelques-uns ont été repris) : le cycle des « Journées particulières[17] ». Il s'agit de rejouer un évènement théâtral, un moment important du répertoire, dont les enjeux dépassent la simple représentation (figure 8). C'est aussi l'occasion d'expliquer les pratiques du public : comment se rendait-on au théâtre, comment la salle était-elle composée, etc. Jouer ces évènements sous forme de spectacle, pour un public amateur, est une manière de valoriser le programme, de faire connaître l'histoire théâtrale, mais aussi pour la troupe elle-même de s'approprier son histoire.

17. Le cycle est coordonné par Agathe Sanjuan ; chaque séance est dirigée par un comédien de la troupe de la Comédie-Française : https://www.comedie-francaise.fr/fr/saison/2019-2020?type=.eventtype_6.

Figure 8. Journée particulière du 9 décembre 2017 dirigée par Julie Sicard, consacrée au 20 octobre 1695

Le Mari sans femme de Montfleury et *Les Vendanges de Suresnes* de Dancourt, spectacle en partenariat avec le Département de musique ancienne du CRR de Paris

© Vincent Pontet – Comédie-Française

Un nouveau modèle organisationnel collaboratif

30. Un programme tel que celui des registres de la Comédie-Française procède d'une étroite collaboration entre informaticiens, chercheurs, et personnels des bibliothèques et des archives. Il convoque aussi les praticiens qui peuvent se saisir de ce matériau pour leurs créations. Il accentue l'interdisciplinarité. Au cours de la soixantaine d'interventions (*hackathons*, communications, spectacles, tables rondes) du premier projet (*PRCF1*), les praticiens se sont faits historiens, les chercheurs en littérature ont systématiquement intégré les conditions matérielles de représentations à leurs études, les chercheurs en histoire des spectacles ont joint les données économiques à leurs analyses esthétiques.

31. Les intérêts des différentes équipes engagées dans le programme ont permis d'intégrer d'autres corpus de données et de documents au projet de départ : les données sur les dépenses (à l'instigation de l'université Paris-Nanterre, de l'université Paris-Sorbonne, de l'université de Rouen et de l'université Harvard), les corpus des feux et de la critique théâtrale (par l'université de Victoria), le corpus des assemblées et comités (par l'université de Victoria et la Comédie-Française).

Conclusion

32. Le programme, tel qu'il se présente actuellement, semble laisser penser que le corpus peut s'étendre à l'infini par agrégation de corpus complémentaires mais aussi par des entrées et analyses de plus en plus fines dans l'ensemble de documents. Les techniques de traitement automatisé du langage laissent même penser qu'une analyse presque immanente des textes est désormais possible.

33. L'une des conclusions que l'on peut tirer aujourd'hui de plus d'une décennie de travail est que les projets d'hu-

manités numériques sont des projets longs : en effet, ils nécessitent des financements importants et la conjonction de compétences qui, jusqu'aux années 2000, avaient parfois du mal à communiquer, entre la science historique et la science informatique. Ce projet a donc été aussi un apprentissage à la communication et au dialogue entre professions.

34. L'ambition du programme est de faire du site internet un « environnement de savoirs » autour de la Comédie-Française historique, de manière à faire converger les intérêts des chercheurs qui pourront y trouver la matière de leurs travaux et du grand public qui pourra obtenir des informations précises et contextualisées sur la Comédie-Française d'autrefois, si ce n'est s'initier à la recherche.

Le projet *eBalzac* : construire une bibliothèque hypertextuelle des sources intellectuelles

Andrea Del Lungo et Karolina Suchecka

Introduction

1. Le projet *Phœbus-eBalzac*[1] consiste à mettre en résonance l'ensemble de l'œuvre balzacienne (*La Comédie humaine*, les romans de jeunesse, les contes drolatiques, le théâtre et les œuvres diverses) avec un vaste corpus d'écrits contemporains, à aires culturelles multiples et de nature variée, qui ont pu la nourrir : œuvres romanesques d'autres auteurs de l'époque ; recueils collectifs de littérature panoramique (auxquels Balzac apporta des contributions) ; ouvrages scientifiques susceptibles d'avoir influencé la création balzacienne, notamment dans les domaines de la médecine, de la physiologie et des sciences naturelles. L'objectif du projet est de permettre des recherches et des comparaisons intertextuelles élaborées, à l'intérieur de l'œuvre de Balzac, et dans le corpus plus vaste de textes littéraires et scientifiques de l'époque, entre 1800 et 1850 afin de faire émerger des correspondances, de repérer des emprunts, des citations, des reprises, des plagiats éventuels, et de constituer ainsi une cartographie de l'univers intellectuel de Balzac à partir des traces que d'autres textes ont laissées dans l'œuvre.

2. Ce projet vise à fournir aux chercheurs en littérature et sciences humaines de nouveaux outils d'interrogation de vastes corpus textuels, susceptibles de permettre une connaissance approfondie de phénomènes génétiques, poétiques, stylistiques, ainsi que leurs implications en termes idéologiques et de mieux comprendre les processus qui régissent l'apparition d'une réutilisation, qu'elle soit avérée ou issue inconsciemment des traces de lecture. Du point de vue de la méthodologie littéraire, il se situe dans le domaine de l'intertextualité, de l'inter-discursivité et de la génétique de l'imprimé, auquel le corpus balzacien offre un champ d'application particulièrement fécond : Balzac a la spécificité de multiplier les supports d'écriture (livres, volumes collectifs, feuilletons, articles de journal), en réutilisant ses textes antérieurs ; en même temps, son œuvre, qui définit par son ampleur l'état socio-historique contemporain, se nourrit de l'apport d'autres textes (notamment ceux de la littérature panoramique), et intègre diverses formes de savoir qui dépassent le champ littéraire. Au terme de ce projet, on sera en mesure de cartographier les sources que Balzac a

1. Le projet a été financé par l'Agence nationale de la recherche pour la période 2015-2019 et porté par Andrea Del Lungo (ALITHILA, université de Lille), Pierre Glaudes (CELLF, Centre d'études de la langue et de la littérature françaises, Sorbonne Université) et par Jean-Gabriel Ganascia (LIP6, Laboratoire d'informatique de Paris 6).

utilisées au cours de l'écriture des différents romans de *La Comédie humaine* et de livrer un développement permettant d'effectuer le même type de recherche sur tout autre corpus textuel.

3. Dans le cadre de ce chapitre, nous présenterons le projet ANR *Phœbus-eBalzac* à partir de sa première réalisation, le site ebalzac.com[2], ouvert en avril 2017, pour se focaliser ensuite sur son dernier axe qui consiste en l'édition hypertextuelle de l'œuvre de Balzac, encore en phase de préparation. Il s'agira alors d'exposer la chaîne de traitement avec les logiciels TextPAIR et Galaxies, de souligner les problèmes que nous avons dû affronter notamment en ce qui concerne le tri et la visualisation des résultats, et de montrer enfin le prototype de visualisation des homologies détectées entre les textes de notre corpus, en commentant quelques résultats.

ebalzac.com : édition numérique, génétique et hypertextuelle

4. Pour réaliser les objectifs du projet, il fallait naturellement disposer d'une édition numérisée fiable de *La Comédie humaine*, ce qui n'était pas une mince affaire : 95 textes, 25 millions de signes ! C'est donc par là que nous avons commencé, en rendant accessible cette édition, jusqu'alors inédite en ligne, grâce à l'ouverture du site ebalzac.com. Sa création a constitué l'occasion de définir un élargissement considérable du périmètre du projet *Phœbus*. En effet, dans le but de créer une édition exhaustive de l'œuvre, il a été décidé d'intégrer à ce site un volet sur l'histoire du texte balzacien, qui consiste à numériser et à rendre accessibles en ligne tous les états imprimés des textes, publiés du vivant de l'auteur, afin de permettre leur comparaison avec le texte de référence que constitue l'édition dite « Furne corrigé ». Le site *eBalzac*, outre l'accueil et la description du projet, comporte quatre grandes rubriques.

5. La première propose une édition électronique de l'œuvre de Balzac, à commencer par *La Comédie humaine*, dans une version inédite en ligne et philologiquement exacte, qui intègre les corrections apportées par Balzac sur son exemplaire personnel et qui corrige de nombreuses éditions antérieures se basant sur ce dernier état du texte. L'entrée dans les textes se fait suivant trois critères au choix de l'utilisateur, via des menus déroulants : plan de l'œuvre, ordre chronologique, ordre alphabétique. L'ensemble du site a été conçu graphiquement avec une séparation verticale au centre qui mime la page du livre et qui permet surtout d'articuler deux espaces en vis-à-vis : dans l'édition, la colonne de droite est consacrée au texte numérisé, et la colonne de gauche à l'ouverture (en cliquant sur le numéro de la page, figure 1) en mode image de la page du support d'origine (dans ce cas, l'exemplaire personnel de l'édition Furne, présentant les corrections de la main de Balzac). L'édition est multi-format, et donne notamment la possibilité de télécharger les textes en EPUB.

2. Cf. http://ebalzac.com/

Figure 1. *La Maison du chat-qui-pelote* : édition Furne corrigé

Crédit : Andrea Del Lungo et Karolina Suchecka

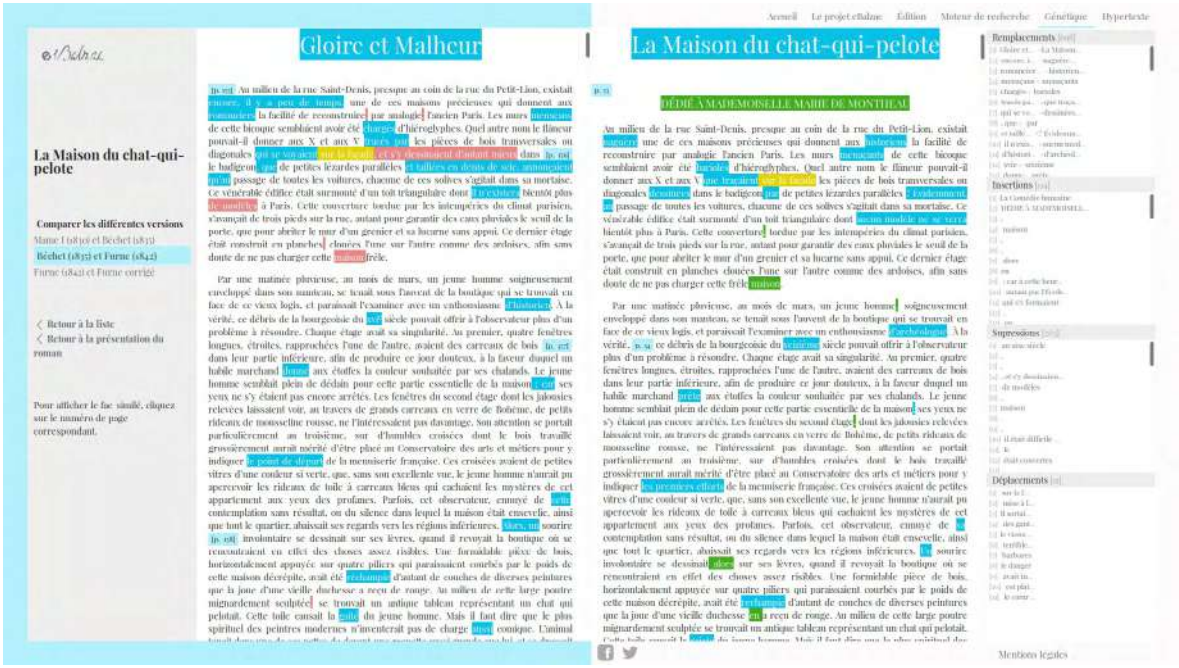


Figure 2. Comparaison des éditions Béchet (1835) et Furne (1842) de *La Maison du chat-qui-pelote*

Crédit : Andrea Del Lungo et Karolina Suchecka

6. Le deuxième axe propose l'édition des multiples états imprimés des textes de *La Comédie humaine* et offre la possibilité d'une comparaison informatique des différents états du texte (exploitant toujours le système graphique de la double colonne pour mettre les textes en vis-à-vis) via le logiciel Medite, permettant ainsi une étude génétique de l'œuvre de Balzac[3].

7. Medite est un outil de comparaison des versions d'une œuvre qui puise entre autres dans l'algorithme d'alignement par fragments grâce à la détection des homologies, une méthode de détection des séquences de macromolécules (ADN ou protéines) afin de faire ressortir leurs régions homologues (Ganascia et Bourdaillet 2006 ; Fenoglio et Ganascia 2008 ; Ganascia 2011). Actuellement, il propose une comparaison de deux textes en format brut : les blocs communs sont analysés et les différentes variantes sont signalées grâce à des codes-couleur. Les remplacements sont marqués en bleu, les insertions en vert, les suppressions en rouge et les déplacements en jaune.

8. Pour cet affichage génétique, nous avons développé une chaîne de traitement permettant d'exploiter la structure XML* afin d'introduire la mise en page éditoriale (alors que l'affichage initial n'admettait que le format de texte brut). Cela implique la facilitation de lecture par la prise en compte des éléments structurels (styles de paragraphe et de caractères, mais aussi images, etc.), l'affichage des fac-similés pour les deux textes comparés et l'alignement automatique au niveau des variantes et des blocs communs (figure 2).

9. Actuellement, la comparaison entre Furne et Furne corrigé est possible pour tous les textes ; la comparaison avec des états antérieurs, déjà disponible pour quelques textes (dont *Le Cousin Pons*), sera à court terme généralisée à l'ensemble, afin de montrer l'intégralité du parcours génétique balzacien à partir de la publication en feuilleton.

10. Le troisième espace du site est consacré à un moteur de recherche lexical, grâce au partenariat avec le projet *ARTFL* de l'université de Chicago, qui met à disposition les fonctionnalités d'un logiciel déjà développé (Philologic4[4]). Trois modes d'exploration sont possibles :

- la concordance (c'est-à-dire, l'extrait du texte avec le mot recherche)
- le KWIC (*Key word in context**) (qui permet un triage par le contexte droit et gauche)
- et enfin, la collocation qui montre le nuage des mots en cooccurrence

11. Enfin, le dernier axe du projet *eBalzac* est orienté vers une édition hypertextuelle afin de recomposer une bibliothèque virtuelle qui comprend l'ensemble des textes lit-

3. Sur la réécriture éditoriale chez Balzac et les enjeux de sa présentation numérique, cf. (Del Lungo 2017).

4. Cf. https://artfl-project.uchicago.edu/philologic4 et (Allen, Gladstone et Whaling 2013 ; Whaling 2010 ; Olsen 2008).

téraires et non littéraires dont on a repéré la trace dans l'œuvre de Balzac. Par son ampleur, mais aussi par le caractère hétérogène de ses sources, *La Comédie humaine* constitue un objet idéal pour ce type d'édition expérimentale qui pourra prendre valeur de paradigme. En effet, le modèle ainsi constitué vise à être opératoire pour d'autres auteurs chez qui l'usage d'une intertextualité abondante et éclectique est avéré.

Détecter et visualiser les correspondances : prototype de l'édition hypertextuelle

12. Cette partie du projet, qui est la plus expérimentale et aussi la plus ambitieuse, reste encore à développer, mais des premiers résultats probants ont pu être obtenus grâce à la collaboration interdisciplinaire et internationale menée dès le début du projet.

13. Les travaux des ingénieurs et de chercheurs en informatique du laboratoire ACASA de LIP6 (Sorbonne Université, sous la direction de Jean-Gabriel Ganascia) et du projet *ARTFL* de l'université de Chicago (sous la direction de Clovis Gladstone) ont abouti au développement de deux prototypes de logiciels qui, ensemble, permettent de détecter les reprises assez subtiles entre les différents textes d'un grand corpus textuel et de visualiser les résultats à l'aide de graphes.

14. Notre rôle principal a été d'enrichir et d'adapter ces prototypes pour qu'ils répondent le mieux aux besoins des utilisateurs visés, notamment des chercheurs en littérature et en linguistique. Cela inclut, principalement, le développement des visualisations modulables, la préparation du corpus XML des textes analysés (optimisation des métadonnées*, calcul des fréquences, etc.). La détection des communautés pour les graphes de taille supérieure à 300 nœuds a été prise en charge par Fleur Gaudfernau (master d'analyse de données et d'intelligence artificielle à AgroParis Tech). Communément, nous avons également mis en place des fonctionnalités permettant de cibler les résultats les plus pertinents (scores, listes des lemmes communs, requêtes de filtrage, statistiques générales, etc.).

La chaîne de traitement : collecter et exploiter le corpus des sources

15. La chaîne du traitement comporte actuellement trois étapes, dont la première est l'établissement d'un corpus des textes structurés à l'aide du standard XML-TEI*[5]. Ce corpus a été progressivement constitué dès le début du projet. À terme, il regroupera au total presque 500 œuvres de 56 auteurs différents, dont la totalité de *La Comédie humaine* de Balzac, qui constitue notre corpus principal (le corpus cible). Le corpus associé (source) est partagé en trois sous-ensembles :

5. Le lecteur pourra également se référer à l'entrée « OCR : *Optionnal character recognition* » pour plus d'informations sur le procédé de numérisation.

- le corpus romanesque d'autres auteurs contemporains ou antérieurs (George Sand, Chateaubriand, Gautier, Sue, etc.), qui est pour le moment le plus riche[6]
- les ouvrages de la littérature panoramique (depuis Mercier jusqu'aux recueils collectifs des années 1840)
- les ouvrages scientifiques contemporains, notamment dans les domaines des sciences naturelles, de la médecine et de la physiologie[7]

16. En ce qui concerne la détection des correspondances, un prototype de logiciel, nommé TextPAIR, a été conçu par le projet *ARTFL*[8]. Il permet de procéder, à partir d'un corpus XML, à la détection des correspondances selon des paramètres assez modulables. On peut, entre autres, choisir le nombre minimal des mots communs détectés, soumettre une liste des mots à ne pas prendre en compte et définir si l'on veut effectuer la recherche sur les mots pleins, les lemmes ou les stemmes (mots racinisés). C'est un logiciel d'alignement de séquences conçu pour identifier des passages similaires dans de grands corpus de textes en s'appuyant sur les techniques d'analyse de séquences employées dans les sciences dures, comme la bio-informatique, avec des applications allant du séquençage du génome à la détection du plagiat. Dans un premier temps, il génère un ensemble des séquences de mots qui se chevauchent pour chaque texte du corpus, puis stocke et indexe les informations à analyser par rapport aux séquences des autres textes. Par exemple, la déclaration liminaire du *Contrat social* de Rousseau, « L'homme est né libre, et partout il est dans les fers. Tel se croit le maître des autres, qui ne laisse pas d'être plus esclave qu'eux », sera traduite en séquences tri-grammes (avec lemmatisation, accents aplatis et mots faibles supprimés), comme : homme_naitre_libre, naitre_libre_partout, libre_partout_fer, partout_fer_croire, fer_croire_maitre, croire_maitre_laisser et maitre_laisser_esclave. Les séquences communes entre les textes indiquent de nombreux types d'emprunts textuels, des citations directes aux utilisations les plus ambiguës et non attribuées.

17. Deux problèmes principaux liés à TextPAIR ont mené à la création d'un logiciel de visualisation nommé Galaxies. Premièrement, le nombre important des résultats, présentés sous forme d'une base de données, où chaque couple de correspondances est inscrit sur une ligne, est très

6. 175 textes du sous-corpus romanesque ont été numérisés et structurés en XML-TEI de manière semi-automatique. Pour ce faire, des transformations XSL ont été conçues à partir du format adaptable DAISY DTBook, disponible sur Gallica pour certaines œuvres les plus connues, et à partir du format EPUB pour les œuvres numérisées par les bibliothèques numériques, dont une partie nous a été mise à disposition par le projet ANR *Chapitre*. Malgré quelques erreurs d'OCR qui persistent, ces numérisations sont à taux d'erreur beaucoup plus faible que par exemple celles disponibles sur Gallica en format texte, ce qui allège de manière significative la question de la gestion des bruits au sein du logiciel.

7. Ce corpus regroupe pour le moment six ouvrages : Nacquart, J-B. 1808. *Traité sur la nouvelle physiologie du cerveau* ; Gall, F. J. et J. G. Spurzheim. 1810. *Anatomie et physiologie du système nerveux en général et du cerveau en particulier*. 4 vol. ; Lavater, J. G. 1820. *L'art de connaître les hommes par la physionomie*. Vol. 1 ; Gall, F. J. 1832. *Sur les fonctions du cerveau et sur celles de chacune de ses parties*. 6 vol. ; Spurzheim, J. G. 1832. *Manuel de phrénologie* ; et Bourdon, l. 1842. *La physiognomonie et la phrénologie*.

8. Cf. http://artfl-project.uchicago.edu/text-pair et, par exemple, (Abdul-Rahman *et al.* 2017 ; Horton, Olsen et Roe 2011). Initialement, les expérimentations sur la détection des homologies ont été menées, dans le cadre du projet *eBalzac*, avec le logiciel Phœbus (http://obvil-dev.paris-sorbonne.fr/phoebus/). Cf. par exemple (Boukhaled, Sellami et Ganascia 2015 ; Ganascia, Glaudes et Del Lungo 2014).

difficile à explorer. Le format de sortie n'est pas adapté aux utilisateurs non-spécialisés et reste très peu lisible. Ensuite, cette détection binaire rend difficile la détection des correspondances croisées (où le même extrait d'un texte correspond à plusieurs autres extraits). Enfin, le nombre des banalités détectées est resté trop important pour que les résultats puissent être présentés au public des chercheurs. Un des buts de Galaxies a donc été de trouver une manière, en combinant les deux logiciels, de limiter le nombre des homologies non-pertinentes et de cibler les recherches selon les besoins spécifiques des chercheurs. Cela inclut, d'un côté, le développement des visualisations modulables à l'aide des graphes et des fonctionnalités comme le calcul de scores, la liste des mots communs, les requêtes de filtrage et les statistiques générales et, de l'autre, l'optimisation du traitement TextPAIR, notamment en écartant les mots les plus fréquents du corpus traité et en conformant la structuration du corpus à l'arborescence des métadonnées prise en compte par le logiciel.

18. Les résultats du logiciel sont ensuite analysés et enrichis par Galaxies afin de construire des graphes de correspondances (figure 3). Les couples de correspondances sont aussi comparés pour retrouver les éléments communs, que nous présentons sous forme de lemmes (pour rendre compte également des correspondances établies sur les différentes formes d'un même mot) et pour calculer le score de chaque couple. Ce score a été conçu au sein du projet, en se basant sur les calculs déjà existants et en expérimentant plusieurs méthodes afin de trouver la plus performante. Les résultats ont ensuite été soumis aux chercheurs littéraires afin de juger de leur pertinence. Le calcul retenu pour le moment prend en compte les proportions de la longueur de deux correspondances (le nombre des chaînes de caractères qui les composent), la fréquence inversée de chaque mot commun par rapport au corpus traité (moins le mot est fréquent, plus il aura de poids pour le score) et le ratio des chaînes de caractères appartenant aux mots communs par rapport à la totalité des chaînes de l'extrait (sur dix caractères de l'extrait, combien, en moyenne, appartiennent à des mots communs ?). Les recherches continuent afin d'améliorer ces résultats, mais la première version du logiciel permet déjà d'écarter la plupart des banalités et d'identifier immédiatement les correspondances les plus proches.

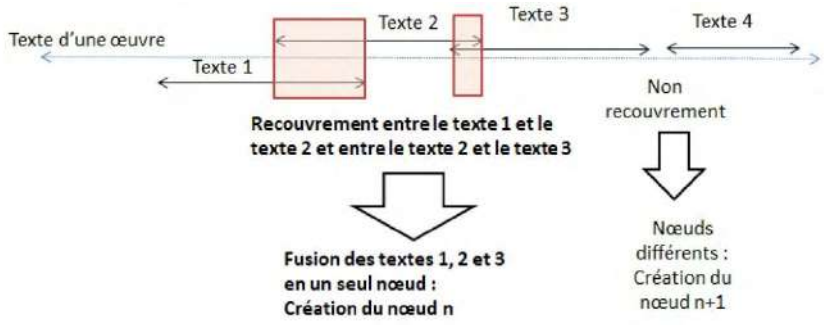


Figure 3. Schéma du traitement des résultats TextPAIR par Galaxies
© Fleur Gaudfernau, AgroParis Tech

19. La liste des galaxies affichée dans l'interface peut être triée selon plusieurs facteurs (score moyen, nombre de nœuds, etc.). L'utilisateur peut ensuite choisir la galaxie qui l'intéresse et l'afficher dans le navigateur, formuler

une requête de filtrage ou accéder aux statistiques générales concernant la totalité du corpus.

Galaxies des relations : visualisation modulable des résultats

20. L'entrée dans les statistiques générales s'effectue par un graphe des auteurs dans lequel le nœud représentant Balzac est situé au centre. En cliquant sur un nœud représentant un des auteurs du corpus romanesque, l'utilisateur peut accéder aux informations relatives au nombre de correspondances détectées avec Balzac (figure 4). Le clic sur l'arête qui lie cet auteur avec Balzac transfère l'utilisateur à la page relative, qui présente à gauche le tableau des correspondances et, à droite, les cinq résultats les plus importants (où les textes sont jugés les plus proches).

21. Nous proposons également un graphique statistique qui illustre le nombre des correspondances détectées dans toutes les œuvres de Balzac (figure 5). Il est partagé en trois diagrammes, selon les différentes sous-sections de *La Comédie humaine*. En pointant la souris sur un des bâtons, les informations détaillées concernant les auteurs avec lesquels les correspondances ont été détectées s'affichent en infobulle.

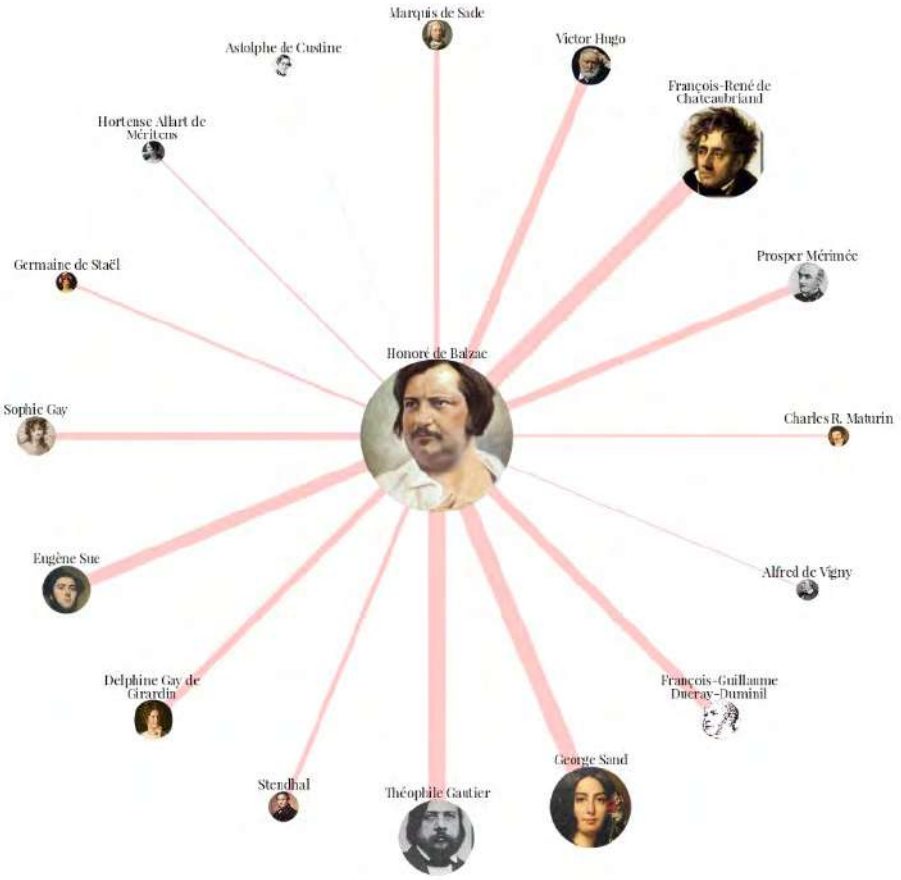


Figure 4. Statistiques générales : graphe des auteurs
Crédit : Andrea Del Lungo et Karolina Suchecka

22. En ce qui concerne les différents graphes qui ont été constitués pour les correspondances croisées, deux modes de visualisation sont disponibles : une concentrée sur les auteurs et les titres (figure 6) et une qui focalise les mots qui ont permis d'établir une correspondance

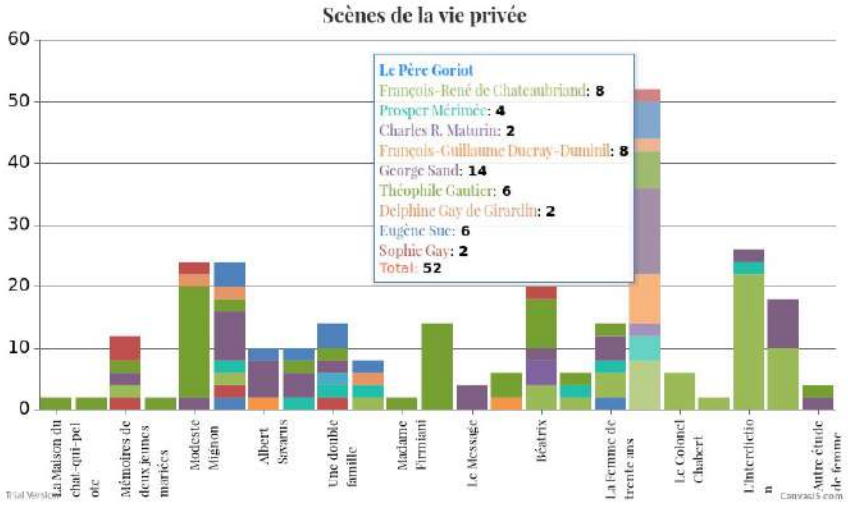


Figure 5. Statistiques générales : graphique des correspondances
Crédit : Andrea Del Lungo et Karolina Suchecka

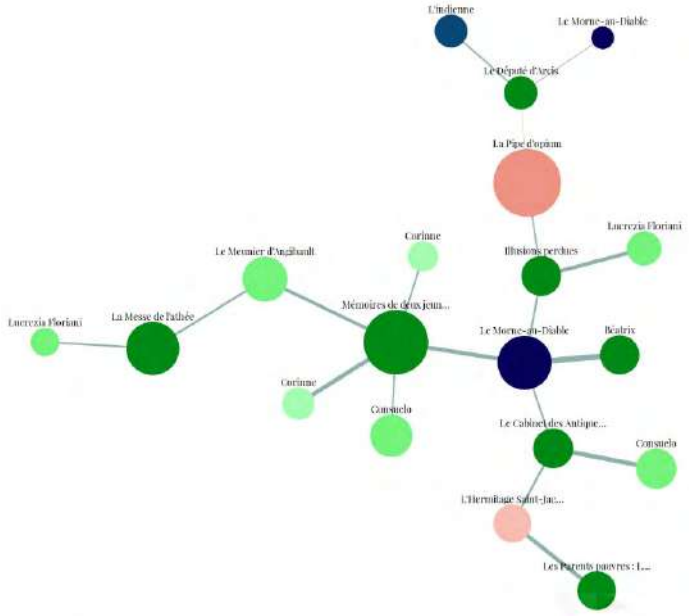


Figure 6. Visualisation focalisée sur les auteurs
Crédit : Andrea Del Lungo et Karolina Suchecka

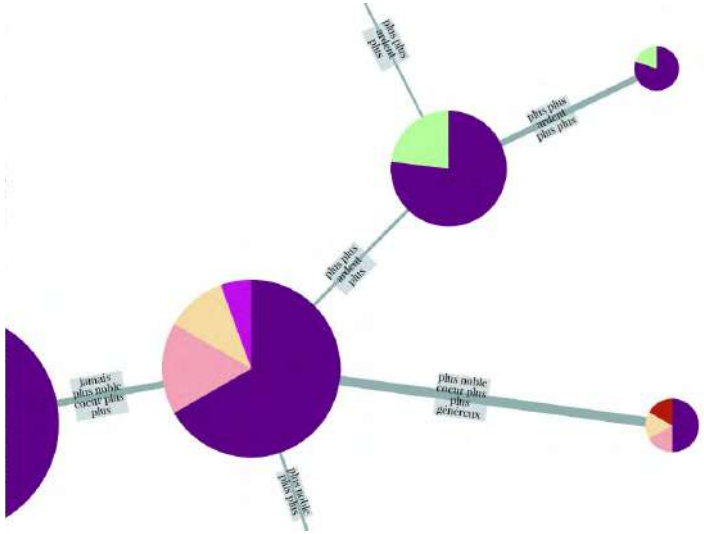


Figure 7. Visualisation concentrée sur les mots communs
Crédit : Andrea Del Lungo et Karolina Suchecka

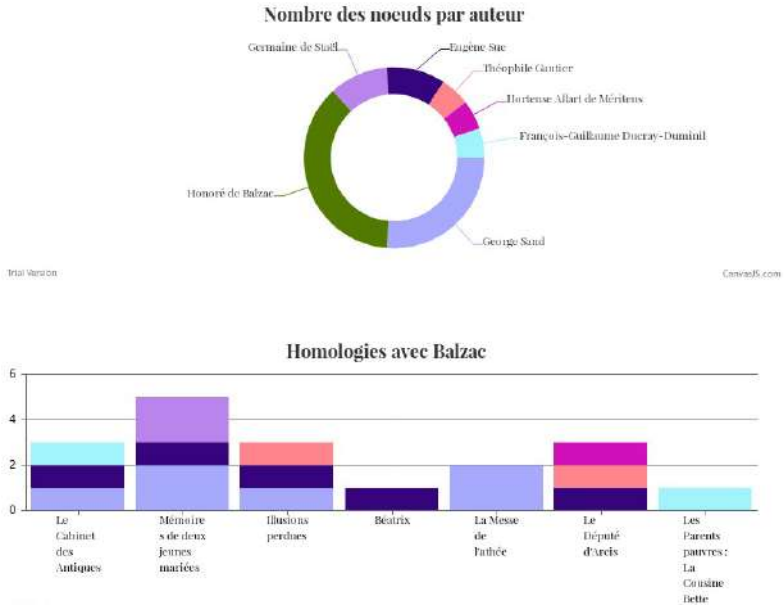
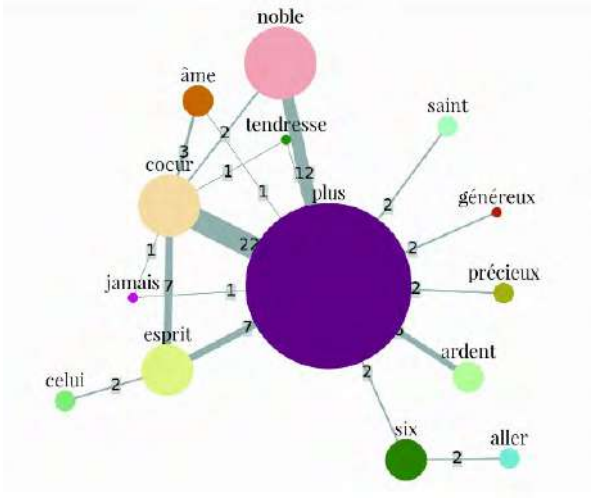


Figure 8. Graphiques statistiques
Crédit : Andrea Del Lungo et Karolina Suchecka

(figure 7). Il est également possible d'accéder aux diverses statistiques concernant la galaxie affichée : deux graphiques qui visualisent les proportions de la présence des différents auteurs et la répartition des correspondances dans les différentes œuvres de Balzac (figure 8), le graphe illustrant les cooccurrences des mots communs détectés, la liste des lemmes communs et leur nombre d'occurrences et enfin, le score maximal, minimal et moyen de la galaxie (figure 9).



Lemmes communs

Lemme	Occ.	%
plus	16	29.09
noble	7	12.73
coeur	6	10.91

Scores de la galaxie

Score maximal :	10.945
Score moyen :	5.07
Score minimal :	0.648

Figure 9. Graphe de cooccurrences, liste des mots communs et scores
Crédit : Andrea Del Lungo et Karolina Suchecka

Exemple de l'exploration des résultats : Chateaubriand et *Sur Catherine de Médicis*

23. Plusieurs types de traitements ont été opérés : nous avons, entre autres, confronté le corpus balzacien aux autres textes du corpus romanesque en écartant les mille mots les plus fréquents ou en ignorant uniquement les mots faibles (les articles, les auxiliaires, etc.). En comparant ces deux approches, nous avons constaté que cette première restriction permet d'écarter un grand nombre de banalités, au risque d'omettre les correspondances pertinentes. Elle résulte également en un nombre restreint de galaxies complexes. Pour Balzac *vs* corpus romanesque sans les mots faibles, la plus grande galaxie compte 3 835 nœuds, dont les mots les plus courants sont « vingt », « sept », « an », « trois » et « bien ». En écartant les mille mots les plus fréquents, la galaxie la plus grande n'en compte que 18, mais les mots communs sont indubitablement plus pertinents : « Henri », « Charles », « François », « Marguerite », « reine », « roi » (figure 10 et figure 11).

Identifiant	nombre de noeuds	score moyen	termes réutilisés les plus courants
1	3835	6.6	vingt sept an trois bien
1-124	270	7.4	somme neuf cent mille franc
1-84	137	7.8	quatre vingt mille livre rente
1-91	121	6.6	cent mille franc cinq cinquante
1-69	100	5.6	vingt cinq an lieu aller
1-18	100	4.7	eh bien partir dire oui
1-11	100	4.7	bien reprendre eh oui monsieur

Figure 10. Balzac *vs* corpus romanesque sans les mots faibles
Crédit : Andrea Del Lungo et Karolina Suchecka

Identifiant	nombre de noeuds	score moyen	termes réutilisés les plus courants	galaxie marquée
59	18	12.3	François II Charles IX Henri	
25	16	7.5	pièce servir cuisine salle manger	
8	12	2.5	salle manger rez donner	
77	10	11.3	cent rente viager foi payer	
35	9	8.2	habit bleu bouton ciseler porter	
85	8	12.4	tribunal premier instance département Seine	
13	8	5.4	tel	

Figure 11. Balzac *vs* corpus romanesque sans les mille mots les plus fréquents
Crédit : Andrea Del Lungo et Karolina Suchecka

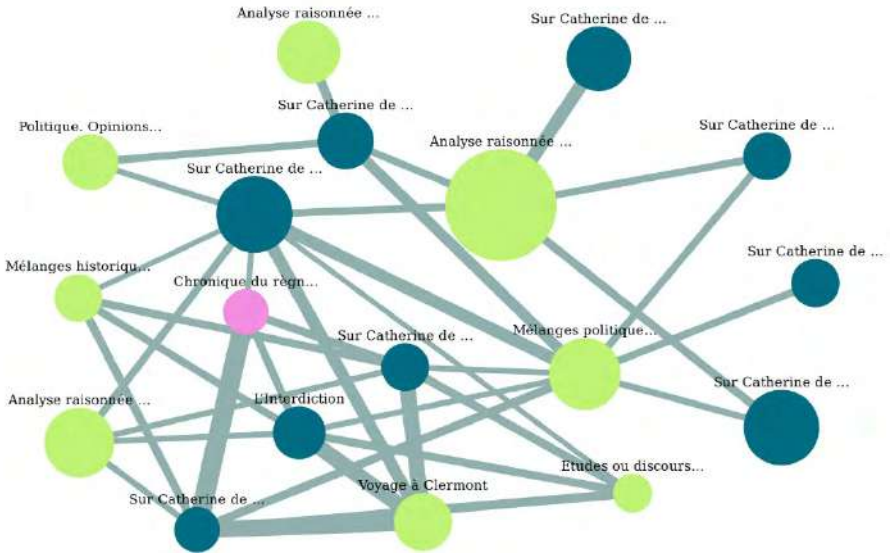


Figure 12. Galaxie n° 59 (18 nœuds)
Crédit : Andrea Del Lungo et Karolina Suchecka

24. Parmi les 18 nœuds textuels de cette galaxie (figure 12), neuf sont issus des textes de Balzac (en bleu), dont huit de *Sur Catherine de Médicis* et un de *L'Interdiction*, huit viennent des textes divers de Chateaubriand (en vert). Nous recensons également une occurrence de *Chronique du règne de Charles* IX de Prosper Mérimée (en rose). Si nous regardons les différents extraits (figure 13), nous pouvons nous rendre compte qu'il ne s'agit pas ici de plagiats ou réécritures, mais plutôt de proximités sémantiques, principalement des énumérations de rois et de reines de France.

25. Grâce à la visualisation focalisée sur les mots communs (figure 14), on peut se rendre compte des glissements thématiques au sein de la galaxie. Pour toute la partie gauche, les extraits traitent principalement d'Henri IV et d'Henri III (nœuds violets, roses et jaunes), alors qu'à droite, les entités nommées* communes sont plutôt Charles IX et François II.

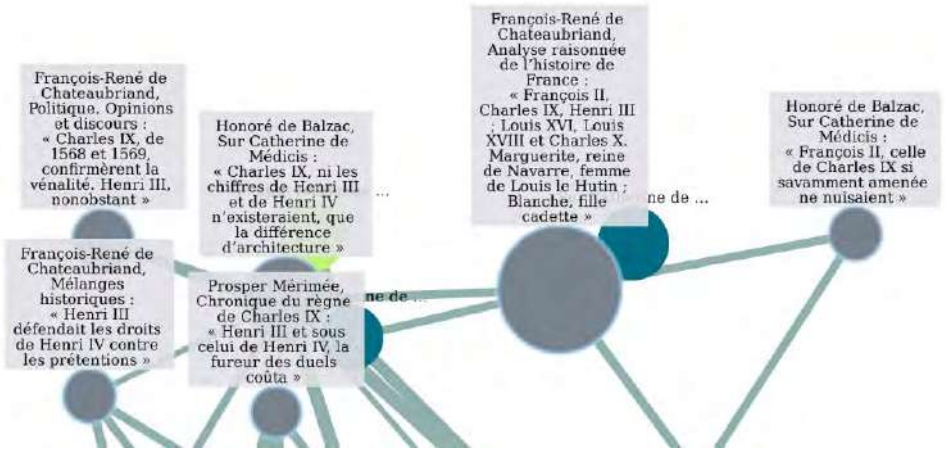


Figure 13. Galaxie n° 59 (18 nœuds) : extraits textuels
Crédit : Andrea Del Lungo et Karolina Suchecka

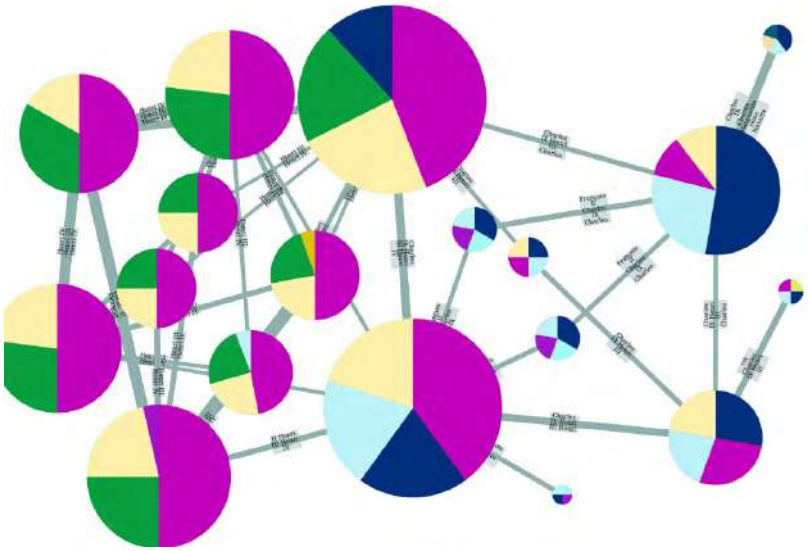


Figure 14. Galaxie n° 59 (18 nœuds) : glissements thématiques

Crédit : Andrea Del Lungo et Karolina Suchecka

Si nous filtrons tous les résultats pour ne garder que les galaxies contenant les textes de Chateaubriand et *Sur Catherine de Médicis*, nous obtenons 18 galaxies au total (dont le n° 59). Presque toutes les correspondances ont été constituées sur les noms propres de personnages historiques (Henri II, Diane de Poitiers, Marie Stuart, Catherine de Médicis, etc.) ou sur les noms de lieux constitutifs des noms de rois et reines (Poitiers, Navarre, etc.). Mais quelques résultats s'écartent de ce schéma.

26. La galaxie n° 53 (figure 15) constituée de trois nœuds, présente des extraits textuels qui sont quasiment les mêmes entre *Analyse raisonnée de l'histoire de France* de Chateaubriand, *Sur Catherine de Médicis* de Balzac et *L'Hermitage Saint-Jacques* de Ducray-Duminil : « Reine, n'ayant de femme que le sexe, l'âme entière aux choses viriles, l'esprit puissant aux affaires, le cœur invincible aux adversités ». Grâce au référencement direct de Chateaubriand qui fait partie des résultats, nous pouvons donc identifier automatiquement cette citation qu'Agrippa d'Aubigné formule à propos de Jeanne d'Albret, et qui reste inavoué chez Balzac, faisant partie d'une réplique du chancelier de Navarre.

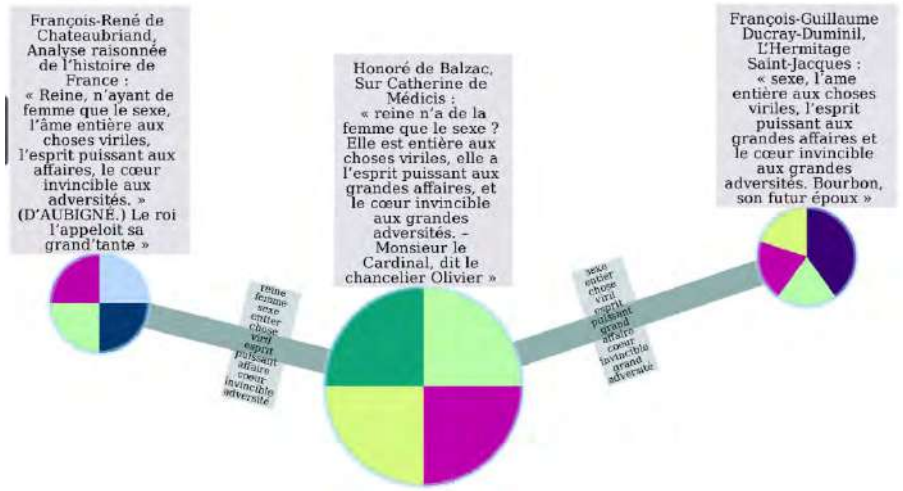


Figure 15. Galaxie n° 53 (trois nœuds), score : 30,9 – citation d'Agrippa d'Aubigné

Crédit : Andrea Del Lungo et Karolina Suchecka

27. Les deux correspondances suivantes (figure 16 et figure 17) restent dans l'ordre de l'énumération, mais la spécificité des entités nommées cooccurrentes (« Inde », « Perse », « Égypte », « Grèce »), les mots savants employés communément (« connétable ») et les adjectifs qualificatifs

(« fameux ») nous permettent, nous semble-t-il, de formuler la thèse suivante : sans que nous puissions parler de réécriture, Balzac documente les aspects historiques de ses œuvres, et notamment *Sur Catherine de Médicis*, en s'appuyant sur les œuvres de Chateaubriand.

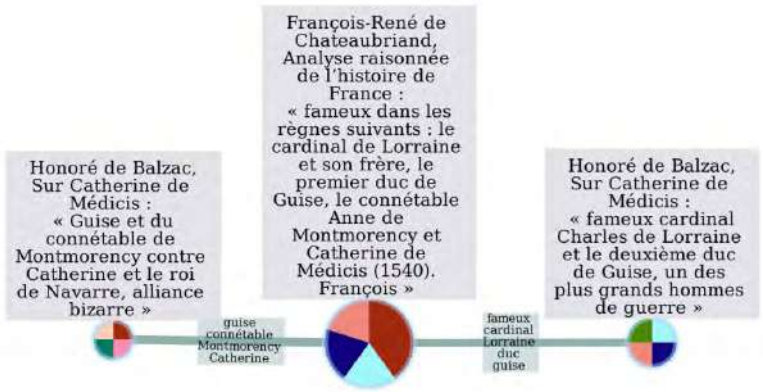


Figure 16. Galaxie n° 123 (trois nœuds), score : 12,1
Crédit : Andrea Del Lungo et Karolina Suchecka

Figure 17. Galaxie n° 65 (trois nœuds), score : 8,4
Crédit : Andrea Del Lungo et Karolina Suchecka

Quelques autres découvertes : des écritures à quatre mains aux proximités sémantiques

28. L'expérimentation que nous avons présentée ici n'a montré qu'une parmi les nombreuses découvertes qui ont été permises par le logiciel et qui sont observables dans les premiers résultats.

29. Nous retrouvons également des réécritures à quatre mains (figure 18), notamment entre Balzac (*Une fille d'Ève*, *Madame de Firmiani*) et Théophile Gautier (*Mademoiselle de Maupin*) qui ont déjà été identifiées de manière « traditionnelle » par (Duclos 2013), et dont les extraits communs sont quasiment identiques. Toutefois, elles ne pourraient pas être retrouvées dans leur totalité par un logiciel de détection de plagiat classique, notamment à cause de quelques ajouts de Gautier (« avec leurs mille têtes chevelues ») et quelques reformulations (« sans faire saigner le cœur, sans que de ta tige brisée suintent des gouttes rouges » ou « sans faire saigner les cœurs à tous ses recoins, et de la tige brisée suintent des gouttes rouges »).

30. Des proximités sémantiques assez surprenantes ont également été relevées par exemple en ce qui concerne les descriptions des lieux et des personnages balzaciens très proches de celles d'Eugène Sue (figure 19). Si nous nous penchons sur les termes communs retrouvés dans toutes les galaxies contenant les textes de ces deux auteurs, nous nous rendons compte qu'ils appartiennent principalement à deux champs lexicaux : le vestimentaire (« redin-

gote », « bouton », « habit », « veste », « chapeau », etc.) et le mobilier (« rez-de-chaussée », « chambre », « cuisine », « salle à manger », etc.). Là encore, les résultats issus du traitement automatique permettent d'appuyer les analyses des chercheurs littéraires : les échanges entre les deux auteurs ont été, entre autres, analysés par (Lascar 2010). Mais l'approche critique que Balzac manifeste envers l'écriture de Sue, surtout à partir de 1836, rend cette proximité assez inattendue.

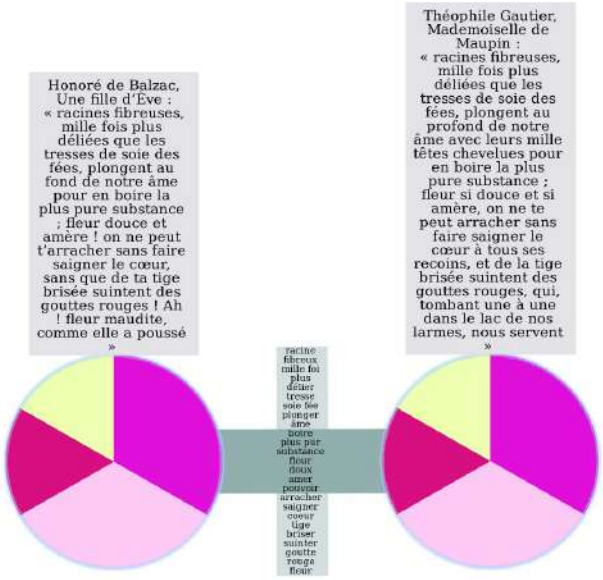


Figure 18. Galaxie n° 316 (deux nœuds), score : 84,3
Crédit : Andrea Del Lungo et Karolina Suchecka

31. Nous recensons également quelques expressions figées qui ne sont plus utilisées de nos jours et dont les occurrences sont assez importantes dans plusieurs œuvres du corpus (« méchant comme un âne rouge », « vouloir manger » ou « faire manger de la vache enragée », « conduite est un chef-d'œuvre de politique », etc.). Une expression empruntée au duc d'Albe « une tête de saumon vaut mieux que dix mille grenouilles » a été, par exemple, utilisée une fois par Prosper Mérimée (*Chronique du règne de Charles IX*) qui la cite mot à mot. Au contraire, deux réutilisations de Balzac, ont été légèrement reformulées : dans *La Paix du ménage*, il écrit qu'« un saumon vaut mieux que mille grenouilles », dans *Les Secrets de la princesse de Cadignan*, « une tête d'un seul saumon vaut celle de toutes les grenouilles ». Outre l'analyse littéraire, ce type de résultats a donc un certain intérêt également pour les linguistes et les historiens de la langue.

Identifiant	nombre de noeuds	score moyen	termes réutilisés les plus courants	galaxie
21	3	16.2	rez chaussee premier etage chambre	
34	3	18.3	redingote bleu boutonner jusque cou	
12	5	10.8	chausser Antin faubourg saint Germain	
41	2	11.7	croisée dont carreau remplacer papier	
25	16	7.5	pièce servir cuisine salle manger	
35	9	8.2	habit bleu bouton ciseler porter	
39	2	9.2	expression cheveu noir ressortir oeil	
30	2	9	manoeuvre couronner plein succès	
40	2	8.8	longue perche charger linge	
37	2	8.7	environ soixante mille livre rente	
31	2	8.5	jambe droit gauche	
43	3	8	escalier conduire étage supérieur	
33	2	8	taille moyen svelte physionomie	
23	3	7.3	cravate noir nouer négligemment pantalon	
36	3	6.4	pantalon noir bas soie soulier	
28	2	7.7	partir éclat rire interdire	
32	2	7.2	chapeau forme rond bord	
26	2	6.9	profond régner troubler coup	
22	4	6	veste gros drap bleu chapeau	
27	2	6.3	nez bec oiseau proie	
24	2	5.9	paraître âgé trente quarante an	
20	2	4.3	sourire satisfaction lèvre commencer	
38	2	4.1	expression peindre trait	
29	2	3.5	perdre qualité mauvais	
42	2	0		

Figure 19. Liste de correspondances entre Balzac et Sue
Crédit : Andrea Del Lungo et Karolina Suchecka

32. Enfin, le logiciel permet de formuler ou de confirmer plusieurs hypothèses, et notamment celle de l'influence des théories scientifiques de l'époque – par exemple celles de la physiognomonie et la phrénologie – sur l'écriture balzacienne, surtout dans la description physique des personnages. Ainsi par exemple, une homologie significative a été retrouvée entre le traité de Gall *Anatomie et physiologie du système nerveux* (publié en français entre 1810 et 1819, en collaboration avec son disciple Johann Gaspar Spurzheim) et un passage de *La Grenadière* (figure 20). Autant Balzac s'inspire des hypothèses de la phrénologie pour décrire l'aspect extérieur du personnage comme indice de son caractère, autant il adapte ces théories et les modifie dans le contexte de la fiction. Chez Gall, le front haut et bombé est un signe d'intelligence, alors que chez Balzac il renvoie à l'énergie de la vigueur. On observe ainsi un déplacement du paradigme herméneutique du domaine intellectuel au domaine physique.

33. Cet exemple de Balzac s'inspirant d'un ouvrage de phrénologie de Gall ne serait plus alors à considérer comme une simple source, mais deviendrait le nœud d'un réseau conceptuel susceptible de montrer comment les modèles scientifiques sont investis, mais aussi déformés dans une œuvre de fiction qui prend la valeur d'une forme de connaissance, et qui contribue à établir de nouveaux paradigmes. Il serait alors possible de repenser la relation de l'auteur avec un ensemble de disciplines positivistes qui ont pu fonder son œuvre, en dépassant la traditionnelle étude des sources afin de situer l'auteur dans un réseau : celui d'un savoir partagé de la culture d'une époque.

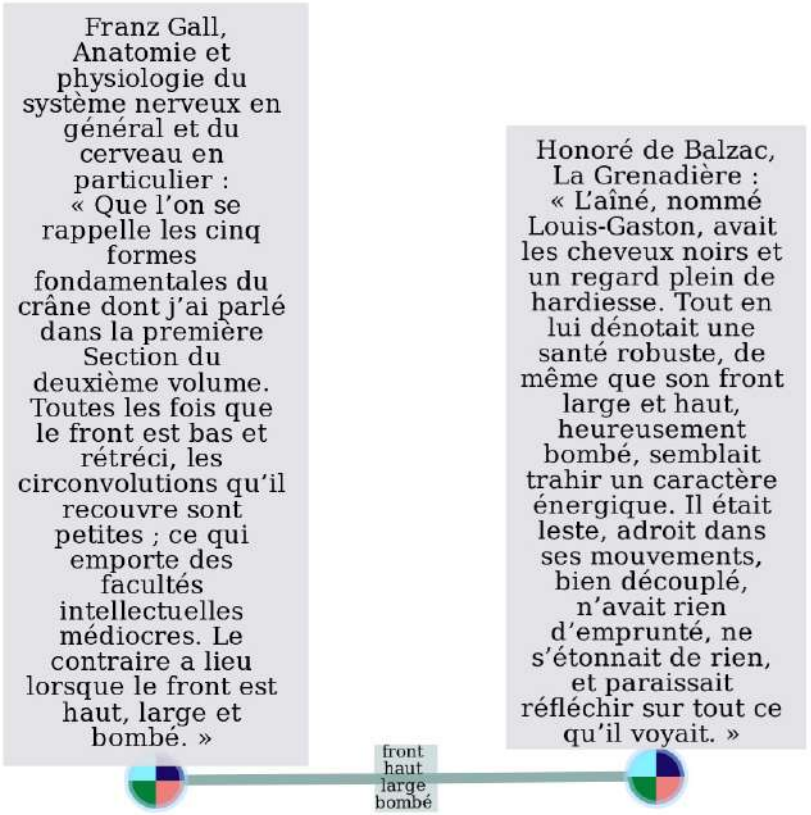


Figure 20. Galaxie n° 4 (deux nœuds), score : 7
Crédit : Andrea Del Lungo et Karolina Suchecka

34. Ces relations croisées sont par ailleurs visibles également dans les premiers résultats de la confrontation du corpus romanesque à lui-même (figure 21) : en approfondissant l'analyse de ce traitement, plus problématique du point de vue du traitement informatique puisque les mêmes

textes constituent le corpus source et le corpus cible, nous pensons qu'il est possible de prouver que les procédés de réécriture ne sont pas caractéristiques uniquement à Balzac, mais marquent l'ensemble de l'histoire littéraire, au moins au XIX^e^ siècle.

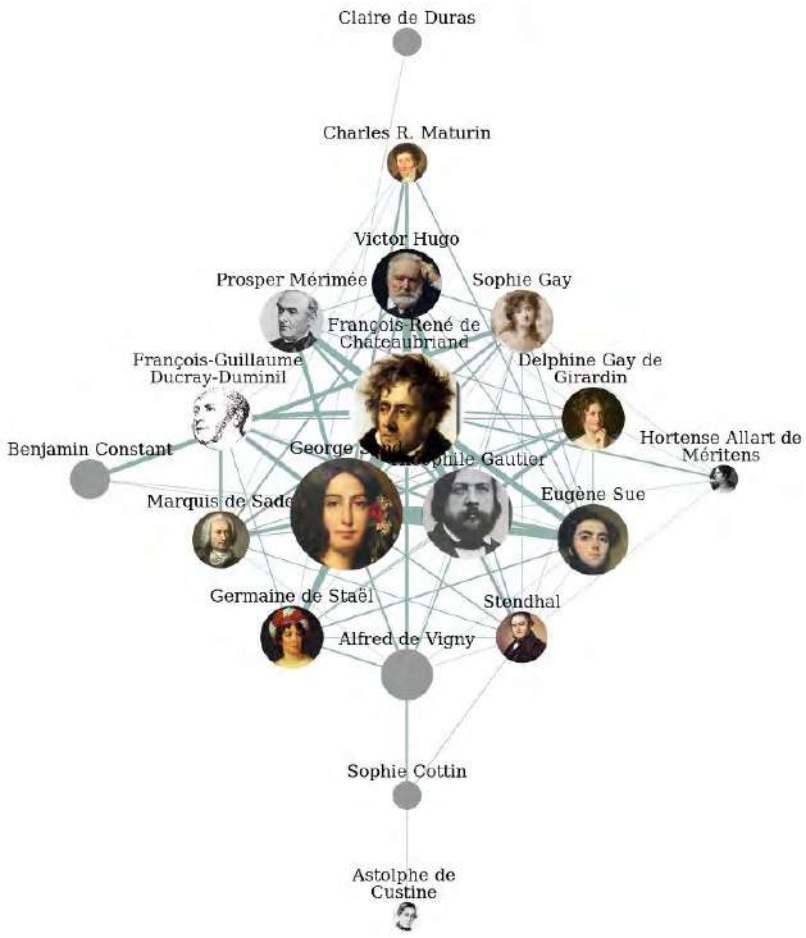


Figure 21. Graphe des auteurs pour le corpus romanesque *vs* lui-même
Crédit : Andrea Del Lungo et Karolina Suchecka

Conclusion

35. Soucieux d'inscrire notre projet dans l'esprit des humanités numériques ouvertes*, nous tenons à ce que tous les développements et numérisations produits dans le cadre de l'édition et du projet en général soient généralisés et mis à disposition de la communauté des chercheurs. Nous souhaitons qu'ils puissent être réexploités au-delà de notre recherche particulière afin de contribuer au passage, désormais indispensable tant dans le domaine des humanités numériques littéraires que de l'édition numérique savante, du quantitatif au qualitatif. Il ne s'agit donc pas seulement d'exploiter les possibilités offertes par les outils informatiques, mais aussi de réfléchir aux modalités d'adaptation des contenus enrichis à la lecture numérique, qui diffère de la lecture papier de manière significative, notamment par sa non-linéarité.

36. Le projet *eBalzac* ne se restreint donc pas à l'exploitation, à la mise en valeur ou à la numérisation d'un corpus. L'originalité de son approche numérique tient au caractère systématique d'une investigation opérée sur des quantités considérables de textes qu'il eût été impossible d'exploiter manuellement.

Pour consulter les données mobilisées dans le chapitre, voir **https://hnso-corpus.nakala.fr/**

L'*Édition numérique collaborative et critique de l'Encyclopédie de Diderot et D'Alembert* (*ENCCRE*), comme prototype d'un laboratoire virtuel de recherches sur l'*Encyclopédie* et les Lumières

Alexandre Guilbaud

Introduction

1. L'*ENCCRE* est la première édition critique de l'*Encyclopédie, ou Dictionnaire raisonné des arts, des sciences et des métiers*, œuvre phare du siècle des Lumières publiée entre 1751 et 1772 sous la direction de Diderot, de D'Alembert et de Jaucourt. Malgré la place importante que l'*Encyclopédie* occupe dans l'histoire, aucune équipe de recherche n'avait tenté d'en réaliser une édition annotée et commentée, qui permette d'y naviguer aisément, d'en expliquer le contenu, l'histoire, le contexte, les enjeux, de mettre à la disposition de tous les fruits de plusieurs décennies de recherches initiées par les travaux fondateurs de Jacques Proust (1995), John Lough (1968), Richard Schwab (1971 ; 1984), ou encore d'accueillir le fruit des recherches en cours sur l'ouvrage. Telle est la raison d'être de l'*ENCCRE*.

2. L'*Encyclopédie*, qui renferme plus de 74 000 articles et près de 2 600 planches sur tous les domaines de savoir, est néanmoins une œuvre trop riche, trop vaste et trop complexe pour espérer en réaliser une édition exhaustive et achevable à un instant t. Dès sa conception, l'*ENCCRE* a donc fondamentalement été pensée comme un projet de long terme, donnant à lire une édition progressivement et collaborativement corrigée, enrichie, annotée et commentée par une équipe internationale et pluridisciplinaire qui compte aujourd'hui près de 160 membres, parmi lesquels plus de 130 spécialistes du XVIII^e^ siècle et de l'*Encyclopédie*. Pour ce faire, l'infrastructure numérique mise au point repose sur deux entités : l'interface numérique de consultation[1], et la plateforme collaborative d'édition, réservée à l'équipe, qui permet à chaque membre de travailler où qu'il se trouve, de s'y authentifier, d'y entrer ses corrections, ses annotations et présentations, le tout étant ensuite publié sur l'interface de consultation au terme d'un processus de relecture et de validation éditorial et scientifique supervisé par un comité de lecture.

3. Au-delà de l'édition critique proprement dite, cette méthode de travail, fondée sur un processus d'enrichissement progressif et collaboratif, permet d'atteindre un

1. Librement accessible depuis le 19 octobre 2017 à l'adresse http://enccre.academie-sciences.fr.

autre objectif essentiel pour l'*ENCCRE* : parvenir à initier une nouvelle dynamique de recherche sur l'*Encyclopédie* et les Lumières. C'est sur cette dimension particulière du projet que nous souhaiterions ici nous arrêter, en nous demandant dans quelle mesure la plateforme collaborative d'édition, qui sous-tend fonctionnellement ce processus, peut être considérée comme un laboratoire virtuel de recherche, ou pour reprendre le terme le plus couramment utilisé dans le domaine des humanités numériques, comme un environnement virtuel de recherche (*virtual research environment*) sur l'*Encyclopédie* et les Lumières. Ce concept d'environnement virtuel de recherche recouvre, comme on sait, une grande variété de situations et de finalités (Carusi et Reimer 2010), au travers desquelles s'imposent néanmoins deux grands principes, fondamentaux dans le cadre de l'*ENCCRE* : l'aspect collaboratif des environnements et leur capacité à contribuer à un processus de recherche.

4. Dans quelle mesure la plateforme collaborative d'édition constitue-t-elle donc un environnement numérique propice à la conduite de travaux de recherche novateurs sur l'ouvrage et son contexte historique, intellectuel et scientifique ? Nous nous pencherons, pour répondre à cette question, sur les principaux modes et fonctionnalités de production et d'enrichissement des données de la plateforme et leur articulation avec les problématiques de recherches actuelles sur l'ouvrage, afin d'évaluer leur capacité à répondre aux besoins intellectuels des spécialistes. Nous identifierons, pour finir, leurs lacunes et les solutions que nous envisageons pour les pallier[2].

L'*Encyclopédie*

5. Il ne devait initialement s'agir, en 1745, que d'une simple traduction d'un ouvrage anglais à succès, la *Cyclopædia* d'Éphraïm Chambers, parue en 1728. Entre les mains de Diderot et D'Alembert, qui se voient confier les rênes de l'entreprise en 1747, le projet prend une tout autre envergure. Si l'ouvrage anglais tenait en 2 volumes, l'ouvrage français, à travers un processus éditorial complexe, en comptera finalement 28 infolios : 17 volumes d'articles et 11 de planches, assorties de leurs explications. Diffusée à 4 000 exemplaires, l'*Encyclopédie* est la plus grande entreprise éditoriale du XVIII^e^ siècle, tant en volume, en capital investi, qu'en force humaine employée, ainsi qu'un vif succès dont témoignent ses multiples contrefaçons et rééditions pirates en France et en Europe. Elle est le résultat d'une histoire éditoriale particulièrement mouvementée, notamment marquée par les polémiques qu'elle a soulevées, et par deux interdictions.

6. L'*Encyclopédie* est à la fois le fruit d'héritages et d'innovations. Des héritages d'abord dans la mesure où elle

2. L'*ENCCRE* est le fruit d'un travail collectif coordonné avec Marie Leca-Tsiomis, Irène Passeron, Alain Cernuschi, et le renfort, depuis janvier 2018, de Malou Haine, Alain Sandrier et Christine Le Sueur. Cet article résulte pour une large part du travail de ce comité de pilotage ainsi que de celui de toute l'équipe de l'*ENCCRE*. Bien entendu, les erreurs et imprécisions commises n'en restent pas moins de mon seul fait. Pour une présentation plus large de l'édition, voir (Guilbaud 2017).

s'inscrit dans une tradition déjà ancienne de recueils de savoirs qui remonte à l'Antiquité. Elle hérite également des traités techniques réalisés sous Louis XIV à l'instigation de Colbert, des recueils de mémoires académiques qui voient le jour à la même époque, des dictionnaires universels dont l'âge d'or s'ouvre à la fin du XVII^e siècle, ou encore de la pensée du chancelier Bacon, fondateur des sciences expérimentales modernes. L'ouvrage innove cependant aussi de bien des façons.

7. Jusqu'alors, les auteurs d'ouvrages à visée encyclopédique, comme Furetière, Chambers, ou le père Souciet, principal auteur du *Dictionnaire de Trévoux*, avaient été essentiellement des solitaires compilant des savoirs livresques produits de seconde main. L'*Encyclopédie*, elle, recourt directement aux savants, et réunit des collaborateurs qualifiés, dont plusieurs comptent parmi les plus illustres de leur temps (Rousseau, Voltaire, Montesquieu, Daubenton, Bourgelat, Berthoud..., et bien sûr Diderot et D'Alembert eux-mêmes, pour ne citer que les plus connus). Les deux-cents collaborateurs, techniciens ou praticiens que les recherches ont permis d'identifier jusqu'ici viennent en outre d'horizons très différents, ce qui fait de l'*Encyclopédie* une œuvre à la fois collective et polyphonique. L'*Encyclopédie* innove aussi en intégrant ce qu'on appelait les « arts mécaniques » dans le cercle des connaissances. Ses onze volumes de planches gravées, assorties d'explications, comptent parmi les plus belles réalisations du dessin et de la gravure au XVIII^e siècle, et témoignent de la grande importance qui est accordée à l'illustration.

8. L'*Encyclopédie* articule enfin deux logiques dans un même ensemble : celle alphabétique du dictionnaire avec celle, raisonnée, qui permet de faire apparaître les liaisons entres les connaissances. Cette autre innovation repose sur l'utilisation de trois moyens : le « Système figuré des connaissances humaines », résultat d'une tentative d'organisation des savoirs inspirée de celle de Bacon et présentée sous la forme d'un arbre ramifié ; l'indication, appelée désignant, qui apparaît généralement à la suite du titre de l'article et qui mentionne le domaine du savoir dont il relève ; des renvois enfin, placés dans le contenu d'un grand nombre d'articles, afin d'indiquer la liaison des matières. À cela s'ajoute la dimension critique de l'*Encyclopédie* : critique des savoirs, dans leur élaboration, leur transmission et leur représentation, critique des préjugés du langage et des interdits de pensée, critique aussi, bien sûr, de l'autorité et du dogme.

9. Ces héritages et innovations, dont nous avons pris le temps de rappeler l'essentiel, définissent plusieurs des principaux axes d'étude passés et actuels sur l'*Encyclopédie* : recherches sur les encyclopédistes et leurs contributions à l'ouvrage, recherches sur l'organisation, les relations et la dynamique des savoirs à l'intérieur de l'œuvre, recherches sur les planches et leurs explications, sur le traitement des arts et métiers, sur l'inscription intellectuelle, philosophique et scientifique de l'ouvrage dans son époque, sur sa postérité et ses reprises dans les dictionnaires et encyclopédies ultérieurs. Des recherches auxquelles s'ajoutent naturellement nombre d'études thématiques et d'études de cas, ainsi qu'une autre dimen-

sion essentielle, celle de la manufacture de l'œuvre. Cette dernière perspective, qui recouvre les modalités de fabrication intellectuelle et éditoriale du contenu de l'ouvrage (qu'il s'agisse de sa nomenclature, de ses articles ou de ses planches), va de pair avec une autre spécificité de l'*Encyclopédie* : la pratique de l'emprunt, indissociable de la manufacture de tout dictionnaire. Les sources utilisées, nombreuses et variées, pour composer le contenu de l'ouvrage, incluent les dictionnaires (dont les universels) et mémoires académiques déjà évoqués, ainsi que de nombreux autres ouvrages, recueils, imprimés ou manuscrits, constituant tout autant de liens, explicites et implicites, entre l'*Encyclopédie* et les savoirs de l'époque. L'analyse de la façon dont ces sources sont citées, empruntées, modifiées et complétées par chaque contributeur à l'*Encyclopédie* constitue un autre axe de recherche incontournable sur l'ouvrage.

Architecture générale de l'*ENCCRE*

10. Diderot, évoquant la nécessaire collaboration des savants à l'Encyclopédie, écrivait : « Quand on en vient à considérer la matière immense d'une encyclopédie, la seule chose qu'on aperçoive distinctement, c'est qu'elle ne peut être l'ouvrage d'un seul homme [...]. Qui est-ce qui définira exactement le mot conjugué, si ce n'est un géomètre ? le mot conjugaison si ce n'est un grammairien ? le mot azimuth si ce n'est un astronome ? le mot épopée si ce n'est un littérateur ? » (article « Encyclopédie », 1755, vol. 5, p. 635). De la même façon, l'*ENCCRE* fait appel à l'historien des mathématiques pour annoter les articles de mathématiques, à l'historien de la grammaire pour les articles de grammaire, etc. Elle s'appuie pour ce faire sur une plateforme en ligne partagée par l'équipe et sur un processus d'édition critique collaboratif fondé sur une répartition des articles et de dossiers thématiques entre ses membres (les *éditeurs*) ainsi que la mise à disposition de fonctionnalités de collation, d'enrichissement, d'annotation et de saisie. Suivant cette logique, toutes les données sont donc produites par le biais de cette plateforme collaborative d'édition, avant d'être vérifiées puis publiées sur l'interface de consultation accessible au public (figure 1).

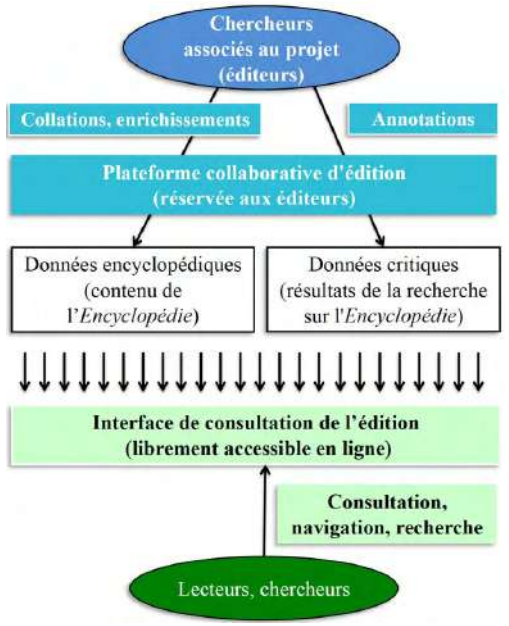


Figure 1. Architecture générale de l'*ENCCRE*
Crédit : Alexandre Guilbaud

11. Les données sont de deux types. D'un côté, les « données encyclopédiques » correspondent à la version transcrite et spécifiée du contenu de l'*Encyclopédie*. De l'autre, les données critiques sont partiellement ou totalement le fruit du travail de l'éditeur : elles se présentent sous forme d'annotations, de commentaires, de présentations ou de notices, et supposent donc un certain degré d'interprétation. L'ensemble de ces données répond à un schéma de structuration – ce que nous pourrions plus classiquement appeler la politique éditoriale et scientifique de l'édition – commune à tous les articles et à tous les éditeurs chargés de leur annotation. Permettre à chaque éditeur de réaliser un travail critique pertinent nécessitait donc d'adopter un schéma de structuration des données construit à partir de l'historiographie disponible et prenant en compte la diversité des problématiques de recherche actuelles sur l'ouvrage.

12. Tel est le principe fondamental qui a guidé l'établissement de la structure de données de l'*ENCCRE* : l'articulation forte avec les acquis et les enjeux de la recherche passée et en cours sur l'*Encyclopédie*. Ce travail a été conduit préalablement à tout développement fonctionnel, et réalisé sous la forme de plusieurs dizaines d'ateliers collaboratifs, afin d'intégrer la diversité des approches. Il se structure suivant trois dimensions complémentaires :

- les modalités de description des données encyclopédiques, propres au corpus
- les différents types et niveaux d'informations critiques prévues dans l'édition
- les modalités d'articulation possibles entre les typologies de données encyclopédiques et critiques précédemment définies

13. Nous les passerons ici successivement en revue, afin de mesurer la capacité de cette structure de données à assurer une articulation effective entre la perspective éditoriale du projet et les enjeux de recherche du corpus ainsi édité.

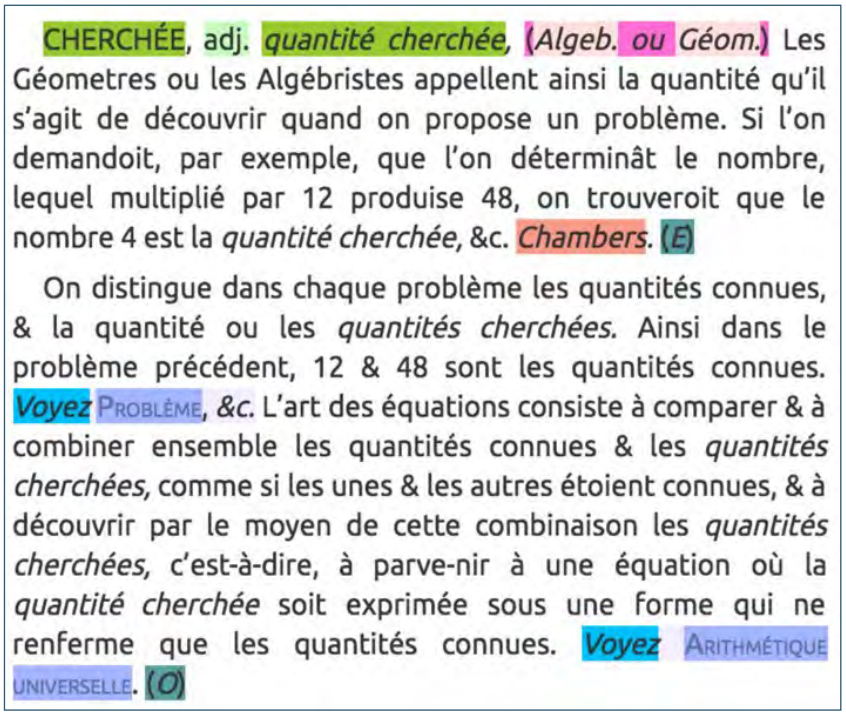
Les données encyclopédiques

14. La structure de données de l'édition, telle que nous l'avons documentée et mise au point à partir des recherches disponibles, s'appuie sur un premier principe fondamental : le repérage, dans la transcription, d'éléments possédant une fonction éditoriale particulière dans l'*Encyclopédie*. C'est ce que nous avons appelé les « constituants encyclopédiques ». Nous avons défini plusieurs types de constituants pour décrire les articles des 17 premiers volumes de l'ouvrage. Citons ici les principaux :

1. La vedette, qui correspond au mot ou au groupe de mots en tête d'article, mis en évidence par des capitales (ex. : « CHERCHÉE », complétée par le complément « quantité cherchée » dans l'exemple de la figure 2). C'est à la fois le titre et le sujet de l'article, comme il est de cou-

tume dans un dictionnaire. Le constituant « vedette » est inextricablement lié à la définition éditoriale du concept d'article dans la mesure où l'*ENCCRE* définit précisément un article comme une portion de texte introduite par une vedette

2. L'indication grammaticale, qui complète généralement le mot en position d'adresse et le suit immédiatement : il s'agit d'une information, généralement abrégée, portant sur la nature grammaticale du mot vedette (ex. : « adj. » pour adjectif)
3. Le désignant, qui correspond au mot ou à l'ensemble des mots explicitant le domaine d'emploi du mot-vedette et donc le champ de connaissance dont relève l'article (ex. : « Algeb. » et « Géom. », pour algèbre et géométrie)
4. La signature : c'est l'indication, souvent située à la fin d'un article, ou d'une partie d'article, qui attribue explicitement le morceau qu'elle ponctue à un (ou plusieurs) contributeur(s) (ex. : « [O] » et « [E] », correspondant aux signatures de D'Alembert et La Chapelle)
5. Les mentions bibliographiques désignent des références explicites à des ouvrages imprimés, qu'ils soient mentionnés, cités, commentés ou compilés dans le texte de l'article (ex. : « Chambers », qui indique ici un emprunt à la *Cyclopædia*)
6. Les renvois encyclopédiques correspondent aux liens vers d'autres articles, vers les planches, ou d'autres parties de l'ouvrage. Leur repérage repose sur plusieurs constituants, dont l'indication de renvoi (ex. : « Voyez ») et l'article visé (« Problème », « Arithmétique universelle »). Chaque renvoi implique logiquement l'identification de sa cible dans le dictionnaire (afin d'assurer la navigation numérique de l'un à l'autre dans l'édition)


CHERCHÉE, adj. *quantité cherchée*, (*Algeb. ou Géom.*) Les Géometres ou les Algébristes appellent ainsi la quantité qu'il s'agit de découvrir quand on propose un problème. Si l'on demandoit, par exemple, que l'on déterminât le nombre, lequel multiplié par 12 produise 48, on trouveroit que le nombre 4 est la *quantité cherchée*, &c. *Chambers*. (*E*)

On distingue dans chaque problème les quantités connues, & la quantité ou les *quantités cherchées*. Ainsi dans le problème précédent, 12 & 48 sont les quantités connues. *Voyez* PROBLÈME, *&c.* L'art des équations consiste à comparer & à combiner ensemble les quantités connues & les *quantités cherchées*, comme si les unes & les autres étoient connues, & à découvrir par le moyen de cette combinaison les *quantités cherchées*, c'est-à-dire, à parve-nir à une équation où la *quantité cherchée* soit exprimée sous une forme qui ne renferme que les quantités connues. *Voyez* ARITHMÉTIQUE UNIVERSELLE. (*O*)

Figure 2. Exemple de repérage des constituants encyclopédiques dans un article

Crédit : Alexandre Guilbaud

15. Plusieurs de ces constituants ont pu être repérés de façon exhaustive grâce à des algorithmes de balisage automatique puis des campagnes de vérification manuelles menées par l'équipe sur la plateforme colla-

borative de l'*ENCCRE* : 74 125 vedettes, 67 421 désignants, 41 150 signatures, 26 602 indications grammaticales et 59 600 renvois encyclopédiques entre articles sont ainsi désormais repérés dans la transcription XML*-TEI* de l'édition. Le repérage systématique des vedettes, des désignants et des signatures a permis de construire des modes d'accès et de recherche privilégiés dans l'*Encyclopédie*, reposant sur un balisage vérifié avec le plus grand soin. Le premier mode, basé sur les vedettes, permet d'effectuer des recherches dans la nomenclature du dictionnaire. Les deux autres, dont nous détaillerons plus loin les principes de construction, donnent respectivement accès à la contribution de tel ou tel auteur d'articles, dessinateur ou graveur de planches, ainsi qu'à l'ensemble des articles et des planches associés à chaque domaine de connaissances. Le repérage des mentions bibliographiques est, quant à lui, effectué au fur et à mesure de l'avancée des travaux des membres de l'équipe, grâce aux fonctionnalités de la plateforme collaborative d'édition.

Les données critiques

16. Venons-en à la structure de l'apparat critique. Celui-ci s'appuie sur le principe d'un éclairage à plusieurs échelles, fondé sur cinq niveaux d'explication entre lesquels une circulation directe est possible : deux niveaux de notes et trois étages de commentaires plus généraux. Il est complété par deux bases de données bibliographiques collectivement enrichies par l'ensemble de l'équipe.

17. Le premier niveau explicatif de la structure de données critiques permet d'effectuer des commentaires ponctuels éclairant le sens d'un terme, identifiant une personne mentionnée, dévoilant la portée d'une allusion, signalant une variante, traduisant un passage en grec, en latin, etc. Si ce niveau d'annotation se rapproche le plus des notes classiques de l'édition papier, le format numérique élargit toutefois sensiblement les capacités d'expressivité de l'éditeur, qui peut numériquement faire apparaître le caractère, le mot, la phrase, le (ou les) paragraphe(s) sur lesquels il a souhaité faire porter son explication ou son commentaire. Le deuxième s'appuie sur la possibilité de faire porter des notes sur les constituants encyclopédiques repérés dans la transcription de l'ouvrage. Ces notes, qui ont des visées explicatives plus spécifiques sur lesquelles nous reviendrons, forment l'un des principaux moyens d'articuler les données encyclopédiques et les données critiques de l'*ENCCRE*.

18. L'étage suivant, plus général, consiste en la présentation de l'article dans son ensemble, sous la forme d'un dossier. Ce dossier inclut notamment un exposé des enjeux de l'article, ainsi qu'une bibliographie des études dont il a fait l'objet. Il synthétise le résultat d'études conduites ou synthétisées par l'éditeur (par exemple sur la manufacture de l'article, ou sa réception), selon un protocole de recherche partagé par l'ensemble de l'équipe. L'édition permet aussi de construire des dossiers transversaux permettant d'ap-

porter des informations ou des commentaires pertinents pour une série plus ou moins grande d'articles. De nombreuses questions, liées à des thèmes de recherche particuliers, éditorialement attachés à des corpus d'articles spécifiques, peuvent être traitées par ce biais. L'*ENCCRE* est enfin dotée d'une riche documentation sur l'ouvrage, constituée de dossiers généraux sur ce qu'est l'*Encyclopédie*, ses héritages et ses innovations, son histoire éditoriale, ses acteurs, sa manufacture, ses sources ou encore sa réception, ses suites et diverses métamorphoses au XVIII^e^ siècle.

19. L'apparat critique intègre aussi deux bases de données bibliographiques, l'une dédiée aux sources primaires, l'autre aux sources secondaires. Les deux bases, intégralement partagées et accessibles par tous les éditeurs, sont progressivement enrichies grâce à la plateforme collaborative d'édition qui intègre des fonctionnalités permettant à chaque membre de créer de nouvelles entrées, de compléter les entrées existantes et d'y faire référence à tous les niveaux de l'apparat critique.

20. La base bibliographique secondaire renferme l'ensemble des études citées par les éditeurs dans leurs annotations, leurs dossiers et leurs présentations, avec les liens vers leurs versions en ligne lorsqu'elles existent. Les éditeurs ont la possibilité de lier chaque référence à chacun des articles et chacun des contributeurs concernés, ce qui permet d'extraire des bibliographies ciblées, et constamment mises à jour, à l'échelle de l'ensemble de l'ouvrage, d'un article ou d'un contributeur particulier. L'ensemble constitue ainsi une base historiographique interrogeable des travaux sur l'*Encyclopédie* et les Lumières, finement articulée avec plusieurs dimensions encyclopédiques et critiques de l'édition.

21. La base bibliographique primaire héberge quant à elle l'ensemble des données relatives aux œuvres mentionnées dans l'ouvrage ou citées par les éditeurs dans leurs notes, dossiers et présentations. Sa structure permet de renseigner l'ensemble des éditions d'une même œuvre, la liste complète de ses volumes, les informations bibliographiques correspondantes, ainsi que l'ensemble des liens vers les versions en ligne de ces éditions et de chacun de leurs volumes dans les bibliothèques (Gallica, Europeana, Biodiversity, Archive, Google Books, etc.) ou tout autre centre d'archives numériques accessibles sur le Web.

22. Après plus de deux ans de travail sur la plateforme collaborative, chacune de ces deux bases contient plusieurs milliers de références, mobilisables à volonté par les éditeurs à chaque étage de l'apparat critique. Penchons-nous à présent sur les connexions de cette structure de données critiques avec les données encyclopédiques, et plus particulièrement, avec les principaux constituants que nous avons fait le choix d'y repérer.

Données encyclopédies et données critiques : les potentialités de recherche de l'édition

Recherches sur les attributions et les encyclopédistes

23. Nous nous intéresserons en premier lieu aux signatures que notre politique de description des données encyclopédiques nous a permis de baliser dans les 74 125 articles de l'ouvrage. Le schéma de structuration entre données encyclopédiques et données critiques prévu par l'*ENCCRE* permet de lier chacune des signatures repérées dans le texte original de l'*Encyclopédie* avec le contributeur correspondant et d'annoter le lien ainsi tissé entre les deux types de données : la marque « (K) » visible dans l'article « Jet d'eau » (vol. 5, p. 521) peut ainsi être associée au contributeur D'Argenville et à la notice qui lui est consacrée, avec une note permettant de renvoyer le lecteur vers les tables souvent bien cachées dans l'*Encyclopédie* qui explicitent la correspondance entre cette marque et le nom de l'encyclopédiste qu'elle désigne.

24. Le même principe de structuration permet de traiter les cas plus complexes, comme celui de l'article « Allées de jardin », dont une étude (Ferlin 2008) a pu montrer que sa seconde partie, quoiqu'introduite par la signature « * » de Diderot, est en fait l'œuvre de D'Alembert. La signature repérée est dans ce cas liée au contributeur D'Alembert et à sa notice, avec une note justifiant cette attribution. Nous gérons de la même façon tous les articles non signés – et ils sont de fait extrêmement nombreux, puisque l'*Encyclopédie* ne contient que 41 150 signatures pour un total de 74 125 articles ! – en tissant un lien, systématiquement justifié par une note, entre l'article et le contributeur identifié.

25. La structure des données définie pour traiter les questions des signatures et des attributions d'article implique donc la mise en relation de toutes les signatures repérées avec leurs contributeurs respectifs ainsi que l'implémentation des 41 150 notes correspondantes, où ces associations sont justifiées et où l'éditeur reste libre d'apporter toute information pertinente relativement à la question de l'auctorialité. Cette structure permet à l'équipe de l'*ENCCRE* de rendre compte des résultats de recherche déjà disponibles dans diverses études sur les attributions d'articles non signés ou dont les signatures peuvent être trompeuses, tout en conduisant un travail de recherche inédit (et progressivement publié dans l'édition) sur cette question via la plateforme collaborative : des travaux inédits sont actuellement en cours sur plusieurs contributeurs, notamment Diderot, D'Holbach, Saint-Lambert, La Chapelle ou D'Alembert.

26. Les liens établis entre données encyclopédiques (les signatures et les articles non signés) et données critiques (les contributeurs et leurs notices) permettent aussi de générer d'autres précieux résultats : l'inventaire des formes distinctes des signatures du contributeur repérées dans la transcription, l'inventaire des articles signés (c'est-à-dire contenant une ou plusieurs signa-

tures repérées) par ce contributeur et l'inventaire des articles attribués à ce contributeur grâce aux résultats de la recherche. Ces trois listes sont présentées dans des fiches et complétées par d'autres informations résultant de recherches en cours sur chaque contributeur, que ce soit sous la forme de données sur les dates, lieux de naissance et de mort du contributeur (et leurs sources), ou d'une notice bio-bibliographique inédite, rédigée par un ou plusieurs spécialistes de l'équipe[3].

Recherches sur la cartographie et la dynamique des savoirs dans l'*Encyclopédie*

27. Le même principe d'articulation entre données encyclopédiques et données critiques a été appliqué pour concevoir un accès par domaine de savoir dans l'*Encyclopédie*. Construite de façon critique, la notion de domaine résulte d'une analyse préalable de l'ensemble des désignants repérés dans l'ouvrage ainsi que des titres d'explications de planches, cette analyse ayant progressivement permis, par regroupements successifs de ces constituants, de dresser une liste de 312 domaines[4]. Suivant une logique similaire à celle appliquée aux signatures, chacun des 67 421 désignants a été lié à un domaine avec une note associée, visant à informer la mise en relation des deux éléments. De même, 10 243 des 12 635 articles ne possédant pas de désignants ont directement pu être associés aux domaines par le biais de notes justificatives.

28. Le résultat représente un ensemble de plus de 77 000 notes que les éditeurs ont la possibilité de compléter afin d'éclairer les questions d'adéquation entre le désignant et les domaines abordés dans le contenu de l'article, ou entre le désignant et le « Système figuré des connaissances humaines ». Ces 77 000 notes sont, bien sûr, autant de liens tissés entre données encyclopédiques (les désignants et les articles sans désignants) et données critiques (les domaines), grâce auxquels nous sommes en mesure de générer une fiche par domaine contenant les listes d'articles et de planches associées, la liste des formes distinctes de désignants sur lesquelles ces inventaires s'appuient, ainsi qu'une notice complète visant à expliciter la façon dont le domaine a été construit. Il faut encore y ajouter les 59 600 renvois entre articles que nous avons pu repérer, et dont l'équipe a récemment fini de vérifier et de corriger les cibles dans le dictionnaire, pour se faire une idée complète du jeu de données actuellement annotable et à disposition des éditeurs pour conduire de nouveaux travaux sur l'épineuse et importante question des frontières entre domaines de savoirs dans l'*Encyclopédie*.

Les mentions d'œuvres, témoins des héritages et des innovations de l'*Encyclopédie*

29. Au contraire des signatures, des désignants et des renvois entre articles, qui ont déjà fait l'objet d'un repérage sys-

3. Voir, par exemple, la fiche dédiée à la contribution de Jean-Jacques Rousseau : http://enccre.academie-sciences.fr/encyclopedie/contributeur/rousseau.

4. Pour plus de détails sur la construction de la notion de domaine, voir : http://enccre.academie-sciences.fr/encyclopedie/politique-editoriale/?s=24.

tématique, la description des mentions bibliographiques constitue, nous l'avons dit, un processus envisagé sur le long cours, au fur et à mesure du travail d'analyse critique des articles par les éditeurs de l'*ENCCRE*. La structure de données fondant cette démarche de spécification du contenu dans l'*Encyclopédie* n'en est pas moins similaire à celle des constituants précédents puisque le repérage d'une mention, tel que prévu sur la plateforme, va obligatoirement de pair avec l'identification (et si nécessaire l'ajout) de l'œuvre correspondante dans la base de données bibliographique primaire.

30. Cette modalité d'implémentation de chaque mention possède de multiples avantages. Elle permet, d'une part, de fournir les références bibliographiques précises des œuvres mentionnées, et donc de désambiguïser les mentions approximatives, ou d'identifier les allusions. Elle permet, d'autre part, d'y associer une note contenant les informations sur l'œuvre dans la base de données bibliographique – nous donnons ainsi accès, lorsque les liens ont été renseignés, à la version numérisée du document original disponible en ligne. Elle fournit enfin, grâce à cette même note, un autre espace d'expression à l'éditeur, ici orienté sur le rôle de cette mention (s'agit-il d'une simple référence, d'une mention d'emprunt, d'une citation ?), sur les modalités d'identification de l'œuvre qui y a été associée, ou sur les éventuelles incertitudes demeurant par exemple sur l'édition de l'œuvre utilisée par les encyclopédistes.

31. Les éditeurs ont en outre la possibilité de déclarer de nouvelles mentions bibliographiques dans la transcription de chaque article, de lier ces mentions aux œuvres renseignées dans la base bibliographique primaire, ainsi que de les annoter librement en fonction des questionnements rencontrés. Ceci constitue un autre moyen de conduire un véritable travail de recherche sur l'*Encyclopédie*, spécifiquement orienté dans le cas des mentions bibliographiques, sur la question des savoirs mobilisés par les contributeurs dans le processus de manufacture éditoriale et intellectuelle des articles de l'ouvrage.

Capacités, limites et perspectives de la plateforme collaborative d'édition

32. Les fonctionnalités de la plateforme collaborative d'édition permettent donc de conjuguer les compétences d'une équipe pluridisciplinaire et internationale de chercheurs pour procéder, d'une part, à la vérification et à l'enrichissement progressif des données de l'œuvre originale (au format XML-TEI) et, d'autre part, à leur annotation et leur connexion aux bases de données critiques de l'*ENCCRE*. La structure de données qui gouverne ces modalités de connexion possibles entre données encyclopédiques et données critiques et les fonctionnalités d'annotation associées sur la plateforme permettent aux éditeurs de contribuer à plusieurs axes de recherches privilégiés sur l'ouvrage (attribution des articles, cartographie des savoirs, manufacture de l'*Encyclopédie*), que ce soit sous la forme de notes (justifiant les liens établis

entre données encyclopédiques et données critiques) ou d'informations critiques de plus grande ampleur, dont les fiches contributeurs et les fiches domaines constituent l'une des manifestations les plus significatives. En synthétisant l'intégralité des informations agrégées par l'équipe d'éditeurs, ces fiches ne donnent rien de moins qu'un état de l'art, sans cesse mis à jour, des travaux de recherche passés et actuels sur les deux axes d'étude correspondants. La même plateforme permet à l'équipe d'alimenter une bibliothèque virtuelle de sources primaires et secondaires connectées aux grandes bibliothèques numériques actuellement accessibles sur le Web ainsi qu'aux principales bases ouvertes des données de la recherche[5]. Ces sources sont continuellement utilisées et enrichies par les éditeurs dans le cadre des opérations de rédaction des contenus critiques (références bibliographiques dans les notes, les dossiers et les notices) et de spécification du texte de l'*Encyclopédie* (repérage et annotation possible des mentions bibliographiques).

33. L'*ENCCRE* intègre aussi une chaîne de validation scientifique et éditoriale doublée de fonctionnalités de publication numérique. La première repose sur de strictes modalités de relecture, et plus généralement d'expertise des apports critiques de chaque éditeur, avant leur validation pour publication sur l'interface de consultation de l'édition. Les fonctions de publication qui gouvernent cette dernière étape vont de pair avec le dépôt de chaque jeu de données critiques sur les entrepôts OAI-PMH de la TGIR Huma-Num et la récupération de permaliens, ce qui permet de garantir non seulement leur accessibilité et leur visibilité sur les bases de recherche nationales et internationales, mais aussi la citabilité des apports scientifiques de chaque membre de l'équipe. Nous disposons en outre d'une fonctionnalité de publication étendue permettant une mise à jour complète du jeu de données (et des index de recherche) de l'interface de consultation à partir de celui de l'interface d'édition, de façon à pouvoir répercuter l'ensemble des fruits du travail de correction et d'enrichissement conduit par l'équipe. Grâce à ce système, l'interface de consultation de l'*ENCCRE* donne à lire une version régulièrement actualisée des 74 125 articles de l'ouvrage enrichis de plusieurs centaines de milliers de constituants repérés et de plus de 110 000 notes qui permettent de connecter ces constituants avec les bases bibliographiques primaires et secondaires, les bases de contributeurs, de domaines et les notices détaillées mises à jour au gré des dernières découvertes. Les articles annotés et commentés paraissent progressivement, quant à eux, au fur et à mesure de l'avancée des travaux d'édition critique et de relecture.

34. L'*ENCCRE* s'apprête enfin à se doter d'une revue numérique, l'*ENCCRiER, la gazette de l'Encyclopédie*, dans laquelle nous ferons semestriellement paraître l'ensemble des articles annotés et des notices publiées sur l'interface de consultation de l'édition au cours des six mois précédents. Cette republication dans une revue des contributions de l'équipe à l'édition critique permettra

5. Pour une brève présentation générale de l'Open access, le lecteur pourra se reporter au glossaire à l'entrée correspondante.

d'accroître la valorisation et la visibilité des travaux de recherche conduits par chaque éditeur. Elle constitue la dernière pierre d'un écosystème de production et de diffusion des données de recherche de l'*ENCCRE* que nous avons schématisé sur la figure 3.

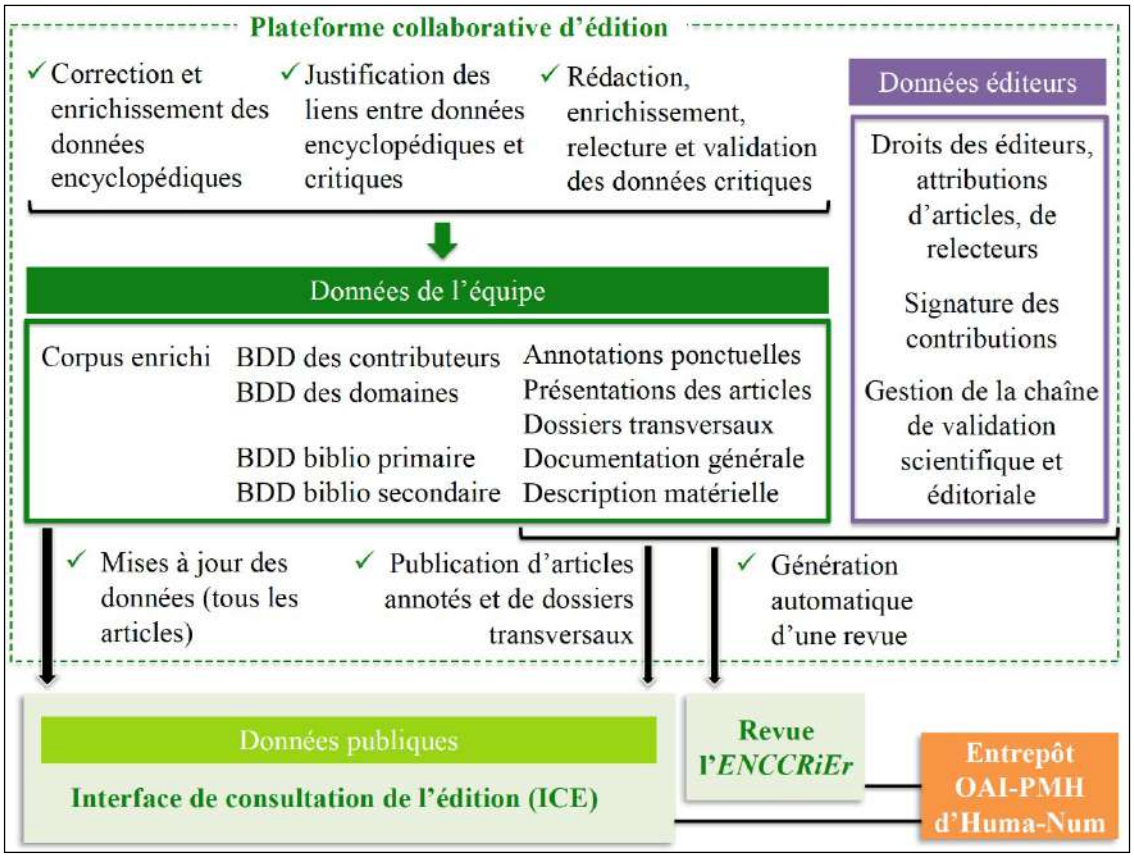


Figure 3. Écosystème de production et de diffusion des données de recherche de l'*ENCCRE*

Crédit : Alexandre Guilbaud

35. Si cet écosystème permet à l'équipe de conduire une activité de recherche soutenue, de faire émerger et de publier de nouveaux résultats, de nombreux moyens manquent encore afin de fournir un environnement de recherche complet, capable non seulement d'explorer la multitude des axes d'analyse envisageables sur l'*Encyclopédie* et les Lumières, mais aussi la richesse des données déjà réunies et scrupuleusement vérifiées.

36. Malgré la disponibilité d'un moteur de recherche permettant d'exploiter l'ensemble des subtilités de la structure de données mise en place, l'absence de plusieurs outils d'exploration et de visualisation des données adaptés constitue ainsi une lacune évidente. Nous pensons à des outils d'identification de réutilisations textuelles qui permettraient à l'équipe de procéder à des recherches de sources parmi les dictionnaires, encyclopédies ou autres traités océrisés accessibles sur le Web, ainsi qu'à des outils d'alignement textuels pour effectuer des comparaisons fines (potentiellement éditables) entre les articles de l'*Encyclopédie* et leurs sources. Nous pensons aussi à des outils d'analyse et de visualisation des réseaux de données déjà exploitables dans l'édition (réseaux de renvois entre articles, entre les articles et les planches, liens entre les articles et leurs

auteurs, entre les articles et les domaines, etc.) pour valider, invalider ou faire émerger de nouveaux résultats ou de nouvelles hypothèses de recherche sur le corpus. Des outils de repérage d'entités nommées permettant d'effectuer des tests de spécification automatique de certains constituants pourraient de même constituer un moyen d'étude complémentaire, capable de contribuer à tester certaines hypothèses et, en cas de validation, d'aider à leur implémentation concrète dans le jeu de données de l'édition. De nombreux environnements de recherche virtuels intègrent d'ores et déjà ce type d'outils[6] et certaines fonctionnalités d'analyse testées sur l'*Encyclopédie* sont également disponibles. Les difficultés à surmonter dans ce domaine tiennent donc clairement moins aux solutions aujourd'hui à notre disposition qu'au travail de réflexion et de préparation méthodologiques et ergonomiques indispensables pour réussir l'intégration de ces nouveaux outils de travail dans l'infrastructure et les routines de pratiques actuelles du projet.

37. L'expérience accumulée dans l'*ENCCRE* ces dernières années montre cependant que la mise en place d'un environnement numérique compatible avec la conduite d'une activité de recherche intellectuellement satisfaisante pour les membres du projet passe par la prise en considération de deux autres questions essentielles : la capacité de l'environnement à fournir des moyens d'expressivité individuels sur ce corpus, et sa capacité à s'adapter à de nouveaux besoins d'expressivité de ses utilisateurs. Une première réponse possible consisterait classiquement en la mise en place de comptes utilisateurs permettant à chaque éditeur d'enregistrer les fruits de ses requêtes, d'archiver ses propres commentaires ou toute autre annotation non destinée à la publication, sur n'importe quelle partie du corpus. Ces espaces permettraient aussi de stocker, d'indexer et d'exploiter les résultats des outils d'exploration de données que nous projetons de déployer sur la plateforme, afin de tester de façon individuelle les hypothèses de recherches qu'ils permettent d'échafauder.

38. Au-delà de comptes permettant un usage personnalisé des fonctionnalités de recherche et d'analyse existantes, une réponse plus ambitieuse au double problème posé consiste à concevoir un environnement dont la structure de données, qui modélise une politique scientifique et éditoriale préétablie, puisse être capable d'évoluer en fonction des besoins des éditeurs et de l'état de leurs connaissances sur le corpus. Cette approche a fait l'objet d'une thèse (Barrellon 2017) réunissant les acteurs de l'*ENCCRE* et de trois autres projets d'éditions critiques numériques (le projet *Desanti*[7], le projet des *Manuscrits de Stendhal*[8] et le projet d'édition des dossiers de *Bouvard et Pécuchet* de Flaubert[9]). La réponse apportée dans ce travail propose de considérer qu'une structure de données peut être à la fois constituée d'un cœur, comportant des

6. Voir, par exemple, l'environnement de recherche proposé par TextGrid (Neuroth, Lohmeier et Smith 2011 ; Hedges *et al.* 2013) ou les outils développés par l'équipe *ARTFL* (https://artfl-project.uchicago.edu/).

7. Cf. http://archive.desanti.huma-num.fr/desanti

8. Cf. http://stendhal.demarre-shs.fr/

9. Cf. http://www.dossiers-flaubert.fr

règles générales partagées par l'ensemble des éditeurs, et de structures spécialisées satellites, dynamiquement formulables et instanciables par un éditeur ou un groupe d'éditeurs en fonction de leurs besoins d'expressivité respectifs sur le corpus. Cette proposition conceptuelle, dont V. Barrellon définit toutes les modalités théoriques, y compris le modèle d'annotation adéquat (sensiblement plus expressif que XML, tout en restant compatible avec les schémas de validation associés), permettrait à un éditeur de définir une structure de données alternative à la structure de cœur et de la tester sur le corpus, tout en garantissant la traduction des données instanciées d'une structure vers l'autre. L'environnement numérique ainsi conçu serait capable de faire coexister simultanément plusieurs façons de représenter, décrire, annoter et commenter un corpus, et donc plusieurs hypothèses de recherche sur ce corpus, formulables, discutables et opposables par et entre différents éditeurs.

39. Il s'agit là, nous semble-t-il, de la piste à privilégier afin de nous rapprocher, dans le cadre d'une édition critique telle que l'*ENCCRE*, d'un processus virtuel où la pluralité des structures de données envisageables pour étudier et comprendre l'*Encyclopédie* garantirait l'expressivité et l'évolutivité nécessaires d'une recherche en cours, en train de se faire, par une équipe de chercheurs dont les travaux incarnent nécessairement une pluralité d'approches intellectuelles possibles et en perpétuelle mutation sur une œuvre et la période au sein de laquelle elle s'inscrit.

Pour un regard à 360 degrés sur les corpus visuels : pratiques de mise à disposition et de réutilisation

Antoine Courtin

1. À l'heure où tout est corpus, cette notion débattue différemment selon les disciplines, n'a encore été que peu conceptualisée dans sa dimension numérique en histoire de l'art, même si la journée d'études « Qu'est-ce qu'un corpus ? » organisée par l'équipe des CBMA (Chartae Burgundiae Medii Aevi) le 7 novembre 2016 à Paris 1, accueillie par l'IRHT avec le soutien du LAMOP et du Consortium Sources médiévales (COSME), en a précisé les grandes lignes (Magnani 2017). Le terme est de plus en plus utilisé dans le domaine de la recherche en histoire de l'art et de manière générale en sciences humaines. Une simple recherche sur Calenda permet de se rendre compte de la vitalité de la notion depuis 2010, de la prolifération de manifestations scientifiques sur le sujet jusqu'à l'appel à résidence en 2019 intitulé « Exploration de corpus visuels » au sein du laboratoire Invisu, unité de service et de recherche de l'INHA (Institut national d'histoire de l'art) et du CNRS. Pour paraphraser Christian Jacob, la profusion des corpus visuels, les multiples outils d'exploitation, d'analyse, de compilation, « peuvent générer une certaine ivresse » et il apparaît ainsi qu'une mise en perspective critique soit de mise afin de « garder un pied sur la terre ferme » (Jacob 2019). Il semble ainsi qu'il soit nécessaire, malgré sa nature mouvante et polysémique, de revenir sur la notion de corpus visuel. Il importe notamment d'analyser les principales options et problématiques devant lesquelles tout acteur (des GLAM[1] jusqu'à l'amateur en passant par les laboratoires de recherche) est confronté lors d'un projet de mise en ligne de corpus visuel numérisé à des fins de diffusion, de valorisation d'un travail de recherche ou même par obligation (de plus en plus de financements de numérisation de corpus exigent maintenant dans un temps imparti, la diffusion de ces derniers). Ce *passage* est un moment crucial car il cristallise de nombreux choix (et donc des priorités) concernant toutes les facettes d'un projet à forte valeur technologique, qu'il s'agisse de choix fonctionnels, de gouvernance, de politique de diffusion (à la fois sur les enjeux juridiques mais aussi éthiques), de moyens techniques et financiers, d'utilisabilité et d'usagers finaux (réutilisations fortuites ou impulsées par les acteurs institutionnels notamment), ou encore de modèles et de méthodes d'exposition des données. Ainsi, pour publier un corpus visuel, faut-il développer une plateforme *ad hoc*, utiliser des systèmes *open source* sur ses propres serveurs

1. L'acronyme anglais GLAM, apparu dans les années 2000, signifie *Galleries, Libraries, Archives and Museums*.

(Dsapce[2], Omeka[3], Scalar[4]), des solutions liées à des prestataires, des services d'institutions culturelles (Gallica marque blanche[5]), des services tiers, qu'ils soient liés au monde du libre (Wikimedia Commons[6], Internet Archives[7], MediaHAL[8]) ou plus opaques (flickr[9], Google Art & Culture[10]), sans oublier des solutions plus légères abandonnant les CMS* au profit de générateurs de sites statiques qui s'inscrivent dans une logique de *minimal computing*[11] (par exemple le modèle Wax[12], qui repose sur Jekyll) ? La question n'est-elle pas, au-delà de l'outil utilisé, de promouvoir un cadre permettant la fourniture de corpus numérisés et de métadonnées* associées, riches et structurées, comme le promeut IIIF (International Image Interoperability Framework) ?

2. Nous évoquerons les principaux enjeux actuels de la constitution des corpus visuels numériques en histoire de l'art ainsi que le champ des possibles ouvert par leur mise à disposition en particulier grâce aux avancées du *deep learning**. Non seulement les corpus visuels ouvrent un réel terrain de collaboration pour développer les connaissances sur les objets culturels mais ils permettent aussi d'imaginer de nouveaux modes de recherche. L'objectif de ce chapitre est de montrer comment les corpus visuels permettent de constituer un « millefeuille informationnel » sur les artefacts culturels par les différents acteurs qui constituent, manipulent, enrichissent, publient, ou encore analysent ces corpus visuels. Ces problématiques seront abordées à la lumière d'une expérience double, à la fois en tant qu'acteur institutionnel et en tant que réutilisateur de contenus. Elles seront illustrées par des travaux qui nous semblent exemplaires des pratiques actuelles, à l'intersection des GLAM et de la recherche en histoire de l'art.

Constituer des corpus visuels en histoire de l'art : enjeux actuels

3. La mise à disposition de corpus numériques visuels pose pour toute institution culturelle la question préalable de l'étape de numérisation. La course à la qualité est l'un des premiers éléments qui vient à l'esprit lorsque l'on parle de numérisation. Le site web de la Bibliothèque nationale de France est une ressource de référence pour connaître les bonnes pratiques actuelles[13]. Des initiatives, comme le projet NumaHOP[14], pour l'élaboration de ces fameux trains de numérisation montrent bien qu'il

2. Cf. https://github.com/DSpace/DSpace
3. Cf. https://omeka.org/, disponible en version Omeka Classic ou Omeka S
4. Cf. https://scalar.me/anvc/
5. Cf. https://www.bnf.fr/fr/gallica-marque-blanche/
6. Cf. https://commons.wikimedia.org/wiki/Accueil
7. Cf. https://archive.org/
8. Cf. https://medihal.archives-ouvertes.fr/
9. Cf. https://www.flickr.com/
10. Cf. https://artsandculture.google.com/
11. Cf. http://go-dh.github.io/mincomp/
12. Cf. https://minicomp.github.io/wax/
13. Cf. https://www.bnf.fr/fr/outils-de-la-numerisation
14. Cf. https://www.numahop.fr/

s'agit d'une étape cruciale. Cette plateforme mutualisée permet de gérer toute la chaîne de numérisation, de la préparation de lots de documents jusqu'à leur dépôt pour l'archivage pérenne, en passant par la publication sur une bibliothèque numérique après avoir subi différents traitements de contrôle qualité. Bien que maintenant relativement bien maîtrisée, de nouveaux challenges frappent à la porte de cette étape de numérisation. À titre d'exemple, la numérisation en très haute définition dite en *gigapixel*, portée et promue par Google Art & Culture, est dorénavant une fonctionnalité de plus en plus attendue des usagers finaux. C'est ainsi que, à des fins de médiation, l'application Second Canvas de Paris Musées met à disposition du public une centaine d'œuvres numérisées en très haute définition issues des collections de ses 14 musées, permettant d'en apprécier les moindres détails. La *reflectance transformation imaging* (RTI), une technique qui permet de montrer les variations de relief d'une surface (par exemple les reliefs créés par des touches de pinceaux sur une toile) est parfois associée à la numérisation en très haute définition. La société ArtMyn[15] propose ainsi un visualiseur interactif en « 5D » qui permet de « jouer » avec la source de lumière afin d'appréhender des œuvres perçues en 2D (par exemple un tableau) comme s'il s'agissait d'objets en 3D. Symptomatique de ce nouvel engouement, cette jeune start-up issue du monde de la recherche (en l'occurrence de l'EPFL, pour École polytechnique fédérale de Lausanne) a réussi à lever auprès d'investisseurs près de 3,6 millions d'euros[16].

4. Les corpus visuels, et plus généralement les images, ont été les « oubliés » ou, pour reprendre l'expression de Régis Robineau (2016), les « parents pauvres » des différentes initiatives de partage et de diffusion des métadonnées patrimoniales telles que le protocole OAI-PMH* (pour Open Archives Initiative – Protocol for Metadata Harvesting) adopté avec succès par les institutions patrimoniales[17]. C'est pour poursuivre cet effort que l'initiative IIIF a vu le jour. IIIF désigne à la fois une communauté de pratique et surtout un cadre technique. Amorcée en 2011, la première des 4 API*, intitulée API Image 1.0, fut publiée en août 2012. Elle a été suivie de l'API Presentation 1.0 en 2013. Ces deux premières API ont été mises à jour en 2014 avec la version 2.0 avant d'être rejointes par l'API Search en 2016 puis l'API Authentification en 2017. L'objectif de IIIF est de créer un cadre technique commun grâce auquel tous les acteurs engagés dans l'élaboration de corpus visuels numériques peuvent délivrer leurs contenus de manière standardisée sur le Web afin de les rendre consultables, manipulables et annotables par *n'importe quelle application ou logiciel compatible*. Ce

15. Cf. https://artmyn.com/

16. Musgrove, Annie. 2019. « Swiss startup Artmyn raises €3.6 million to revolutionize art market with advanced scanning technology ». *Tech.eu*, 30 octobre 2019. https://tech.eu/brief/swiss-startup-artmyn-raises-e3-6-million-to-revolutionize-art-market-with-advanced-scanning-technology/.

17. Le succès d'OAI-PMH est largement dû à sa facilité d'usage et de mise en place. Il est néanmoins confronté aujourd'hui à sa trop grande simplicité dans l'expression des informations.

dernier point est essentiel car il permet d'envisager le corpus visuel comme « matière », sans qu'il soit assujetti aux « contenants » et de privilégier ainsi l'utilisabilité, ou encore les tendances à un moment t, que ces tendances soient techniques ou bien même esthétiques. Après seulement quelques années d'existence, IIIF a remporté un large succès et a été adopté par de nombreuses institutions culturelles et patrimoniales pour diffuser leurs collections d'images numériques. Parmi les nombreuses applications issues de IIIF, Mirador, développée par Stanford et Harvard, est un « visualiseur web qui offre des fonctionnalités avancées de zoom, de comparaison et d'annotation d'images en haute résolution, indépendamment du type de document ou de la bibliothèque numérique qui les héberge. Il permet d'afficher dans une interface commune des documents numériques provenant d'entrepôts d'images différents[18] ». Il devient ainsi possible de constituer un corpus visuel à partir d'items provenant de différentes collections, de citer des portions d'image, ou encore de les comparer facilement grâce à un système de multi-fenêtrage. Le programme intitulé « France et Angleterre, 700-1200 : manuscrits médiévaux de la BNF et de la British Library[19] », est exemplaire des nouvelles possibilités offertes par IIIF. Huit cents manuscrits conservés par la BNF et la British Library sont réunis dans un même espace numérique grâce au protocole IIIF et au visualiseur Mirador. Tout un chacun peut comparer les manuscrits, les annoter, sélectionner des détails, etc., au sein d'une même interface[20].

5. Malgré ces avancées indéniables, les visualiseurs posent de vrais soucis trop souvent occultés. En effet, ces interfaces qui s'adaptent à toutes les tailles d'écran inhibent l'échelle des œuvres originales, gommant la matérialité de l'objet : sur un écran, l'image numérisée d'une carte de plusieurs mètres occupe le même espace que celui d'une miniature de quelques centimètres. Ce phénomène est préjudiciable car il tend à effacer l'objet matériel au profit de son double numérique. Ainsi, si des efforts ne sont pas faits pour mettre en garde ou donner quelques éléments de « mise à l'échelle », des surinterprétations, voire des erreurs d'interprétation pourraient être commises. Apparaissent parfois dans les interfaces de visualisation des images des repères permettant de percevoir le rapport d'échelle, par exemple entre l'objet original et le corps humain ou des objets de la vie quotidienne[21].

6. La mise à disposition de corpus numérisés engage ceux qui les diffusent pour deux raisons principales. La première est liée à la continuité de service, soit la possibilité d'accéder de manière pérenne au corpus. Pour l'anecdote, une région avait publié en 2015 un inventaire des objets religieux au format CSV*. Il s'agissait alors de l'une des

18. Cf. https://doc.biblissima.fr/visualiseur-mirador
19. Cf. https://manuscrits-france-angleterre.org/
20. À noter que la BNF a mis en place en 2021 ou plutôt a généralisé l'intégration du visualiseur Mirador en le rendant accessible directement depuis son interface classique. L'icône IIIF est affichée dans la section « synthèse ».
21. Voir les expériences du V&V (http://waddesdon-bequest.herokuapp.com/) ou du Brooklyn Museum (https://www.brooklynmuseum.org/opencollection/).

premières réelles initiatives d'Open data* culturel en France. L'inventaire était composé d'une vingtaine de champs de description et d'une URL pour accéder aux photographies des objets, URL d'un serveur avec comme nom de domaine, la région. Avec la fusion des régions, effective en 2016, l'ensemble des noms de domaines des anciennes régions a été rendu obsolète, ne permettant plus l'accès au corpus numérique visuel des objets religieux. Il a fallu près de 2 ans pour que le jeu de données soit mis à jour avec de nouvelles URL. La seconde raison est que la diffusion de corpus visuels peut être considérée comme une prolongation de la responsabilité de l'institution vis-à-vis des œuvres qu'elle conserve, et dont elle propose une représentation numérique, en particulier en terme de droit moral des artistes. Une simple recherche sur Google donne un panorama de toutes les reproductions numériques d'une même œuvre en différentes tailles, résolutions et combinaisons de couleurs, soit dans des qualités extrêmement diverses. La difficulté réside alors pour l'utilisateur dans l'identification de la version numérique la plus fidèle à l'original. C'est le *Yellow Milkmaid Syndrome*[22], du nom de l'œuvre de Vermeer conservée au Rijskmuseum, dont de multiples versions de qualité pour le moins très variables étaient disponibles en ligne, avant que le musée ne mette en *open content* sa propre version en haute résolution, devenue depuis la version de référence pour toutes les personnes souhaitant la présenter sous format numérique.

Des métadonnées au *deep learning*

7. Si nous venons d'évoquer rapidement les enjeux actuels de constitution des corpus visuels[23], il importe de revenir sur la notion de corpus visuel en tant que telle. Depuis les années 1980, et plus fortement depuis le début des années 2000, les recherches en linguistique[24] s'intéressent à définir l'objet corpus comme « un recueil, formé d'un ensemble de données sélectionnées et rassemblées pour intéresser une même discipline » (Mellet 2002). Pourtant, la conceptualisation de la notion de corpus du point de vue des recherches en histoire est bien plus récente, comme en témoigne la journée d'études intitulée « Qu'est-ce qu'un corpus ? » organisée le 7 novembre 2016. En s'appuyant sur les travaux d'Elena Tognini-Bonelli (2001), cette journée d'études a notamment posé la question suivante : « le corpus est-il un objet ou un moyen, comme on pourrait l'envisager dans la distinction entre *corpus-based* et *corpus-driven* – le premier correspondant à une manière traditionnelle de constituer un corpus puis de le piller, le deuxième engageant à penser la manière d'exploiter un corpus » (Magnani 2017) ?

22. Sarah Stierch, experte en muséologie et défenseur de la culture ouverte, a lancé le blog *Yellow Milkmaid Syndrome* où elle collectionne et présente la grande variété de versions d'une œuvre d'art que l'on peut trouver en ligne, https://yellowmilkmaidsyndrome.tumblr.com/. Voir également : https://pro.europeana.eu/post/the-yellow-milkmaid-syndrome-paintings-with-identity-problems.

23. Beaucoup d'autres aspects brièvement évoqués en introduction ne peuvent être développés dans le cadre de ce chapitre. Pour l'enjeu fondamental de l'Open Data en histoire de l'art nous renvoyons à (Denoyelle et al. 2018).

24. Voir le chapitre « Les corpus textuels numériques (re)spécifiés » de Damon Mayaffre dans cet ouvrage.

Cette réflexion nous semble tout à fait pertinente pour analyser la notion de corpus visuel en histoire de l'art où les problématiques de constitution et d'utilisation du corpus sont intimement liées. Il est en effet difficile dans ce domaine d'appréhender une recherche sur corpus au travers de la seule approche *corpus-based* même si cette dernière peut être rapprochée du travail effectué par les institutions culturelles lors de la création de catalogues numérisés de leurs collections. À partir de ces corpus de collections, une seconde approche – plus proche du *corpus-driven* – menée par le chercheur, vise à constituer, à partir d'œuvres numérisées émanant de différents catalogues d'institutions, un corpus d'étude autour d'une thématique de recherche. Une telle approche permet de ne pas présager de ce que le corpus va apporter, plaçant ainsi le corpus comme réel objet de recherche[25]. À l'heure actuelle, les bases de données en ligne sur l'art ne peuvent être consultées qu'à l'aide de mots clés ou de balises, à savoir les métadonnées textuelles de l'objet original. Pour constituer son corpus d'étude, le chercheur doit passer par la consultation des métadonnées descriptives des œuvres qu'il envisage d'intégrer à son corpus. Or ces métadonnées peuvent être incomplètes par rapport à son objet d'étude – et ne comporter par exemple que des informations sur le titre ou l'origine de l'œuvre mais aucune description iconographique – ce qui risque d'entraîner la non-intégration de l'œuvre au corpus. Les requêtes textuelles ne sont pas conçues pour rechercher des informations visuelles qui n'ont pas été indexées : elles ne peuvent par exemple pas trouver de zones, de formes ou de motifs similaires. Pour pallier cette difficulté, il faut envisager un enrichissement le plus exhaustif possible de la description des œuvres dans les métadonnées grâce à l'exploration du corpus par l'image elle-même. L'idée d'explorer un corpus visuel issu des institutions culturelles par ses propres composants commence à émerger. L'objectif est d'augmenter les capacités de recherche en se soustrayant aux métadonnées descriptives, bien que celles-ci soient indispensables pour les corpus d'entraînements, au cœur de l'écosystème des méthodes dites de *machine learning**. Ce fantasme est au centre de projets de recherche dont le nombre s'est envolé au tournant des années 2010 à partir des progrès majeurs de l'intelligence artificielle et des méthodes de reconnaissance visuelle issues de l'apprentissage profond. Ils s'inscrivent dans la discipline plus ancienne qu'est la vision par ordinateur (*computer vision**). De nombreux colloques, journées d'études et workshops ont ainsi vu le jour, comme par exemple le workshop « Computer Vision in Digital Humanities » lors du colloque DH2017 à Montréal[26] ou encore les workshops Visart (pour « Computer Vision for Art Analysis ») qui se déroulent tous les 2 ans depuis 2012[27]. Le développement de ce type de techniques rend possible des recherches automatisées qui font apparaître des motifs par similarité dans un corpus d'œuvres numé-

25. Pour donner un seul exemple, les recherches de Fabienne Gallaire sur les « représentations occidentales de chouette ou hibou entouré/picoré/attaqué par des oiseaux » illustrent parfaitement cette méthodologie : https://twitter.com/Lignedescience/status/718350304199196672.

26. Cf. https://avindhsig.wordpress.com/workshop-2017-montreal/

27. Voir la page officielle de la dernière édition en 2018 à Munich : https://visarts.eu/.

risées. Les recherches de John Resig autour des archives photographiques de la Frick Collection sont pionnières et ont porté leurs fruits au travers du projet européen *Pharos* (2013). L'objectif de ce projet est ainsi présenté : « *Pharos is an international consortium of fourteen European and North American art historical photo archives committed to creating a digital research platform allowing for comprehensive consolidated access to photo archive images and their associated scholarly documentation*[28]. » En 2016 commence à l'EPFL le projet *Replica* (Seguin 2018b), porté par les nouvelles perspectives offertes par les technologies de *machine learning*. Réalisée notamment dans le cadre de la thèse de doctorat de Benoit Seguin (2018a) l'objectif de cette recherche est de proposer, à partir de la photothèque de la Fondation Cini, l'un des premiers moteur de recherche conçu spécifiquement pour la recherche et l'exploration des collections artistiques. Intégrée à Diamond et au projet plus vaste *Time Machine*[29], cette preuve de concept permet à l'usager de « rogner » la zone d'une image de la photothèque pour isoler un motif ou encore d'importer une image à partir de son propre ordinateur et de lancer une recherche pour identifier toutes les autres images, ou portions d'images, similaires au sein de la collection.

8. Aujourd'hui, les corpus visuels sont tiraillés entre l'exploration par l'image elle-même et la nécessaire mise en données de différentes natures (données des institutions culturelles, des résultats de la recherche, des instruments d'analyse, issues du *crowdsourcing**, etc.) apposées sur le corpus visuel et permettant ainsi son exploitation. La base Art UK[30] réconcilie ces deux approches qu'il faut se garder de considérer comme étant antinomiques. Le Visual Geometry Group de l'université d'Oxford[31] (Crowley and Zisserman) a été invité à y expérimenter de nouvelles méthodes d'exploration visuelle. Il a créé un logiciel de reconnaissance d'images permettant d'identifier les sujets dans les peintures, à partir d'un corpus d'entraînement de plus de 3,5 millions d'étiquettes issues du projet de *crowdsourcing Tagger*[32]. Chacun était invité à repérer les différents éléments reconnaissables dans un tableau, comme une fleur, un enfant, une rivière, et à les annoter sur l'image numérique. Sans sous-estimer l'ampleur de ce « *digital labor* », le travail effectué par les usagers permet d'améliorer l'usage final du corpus visuel d'Art UK en proposant des centaines de nouvelles étiquettes et ainsi de « dépasser » les index limités créés par les professionnels des institutions culturelles. Cet exemple montre un possible cercle vertueux des données où l'alliance entre *machine learning* et corpus visuels numérisés vise à une consolidation ou à un enrichissement des métadonnées existantes, enrichissement qui permet d'ouvrir vers de nouveaux modes d'accès, vers de nouveaux usages des corpus visuels.

28. Cf. http://pharosartresearch.org/about
29. Cf. https://diamond.timemachine.eu/
30. Cf. https://artuk.org/
31. Cf. https://www.robots.ox.ac.uk/~vgg/people.html
32. Cf. https://artuk.org/participate/tag-artworks

Fournisseur de données, fournisseur de service et usagers finaux : des frontières qui s'estompent

9. Une galaxie d'outils de mise en ligne de corpus visuels est aujourd'hui disponible. Comme nous l'avons rapidement évoqué en introduction, exploiter visuellement les corpus visuels, c'est aussi penser à de nouvelles interfaces, au-delà des CMS, car d'autres moyens de diffusion existent. À titre d'exemple, des laboratoires de recherche sur la visualisation de l'information s'intéressent plus particulièrement aux corpus visuels issus des institutions culturelles et proposent en *open source* des interfaces. Ainsi, le Urban Complexity Lab de l'université des sciences appliquées de Postdam propose l'application web Vikus[33]. Celle-ci permet de visualiser un corpus visuel (qu'il s'agisse de dessins, de photographies anciennes numérisées, etc.) selon différents critères comme le temps, les thèmes représentés ou certaines caractéristiques visuelles (dénommées « texture »), via les métadonnées structurées au format JSON*. Cette prolifération pose question : faut-il dissocier les outils de « constitution de corpus », des outils de publication et les outils d'exploitation de ces corpus ou faut-il, à l'instar des électroménagers « 4 en 1 », penser de manière intégrée ? En effet, à l'heure de la multiplication des outils développés dans x projets à travers le monde, et ce à toutes les étapes du *cycle de vie d'un corpus* comme l'attestent de nombreux annuaires ou articles en ligne[34], ne faut-il pas réaffirmer la nécessaire décorrélation entre le corpus lui-même, que l'on nommera « matière » et les outils et services qui ont permis de le constituer, de le façonner, de l'enrichir, de l'exposer, de le diffuser et de le médier ? À l'instar du monde de l'entreprise où, à partir des années 2005, la notion de BYOD (pour *bring your own device*) devenait une priorité croissante, il faudrait inventer le UYOS (*use your own software*).

10. Avec la profusion de logiciels, de services web ou encore d'applications mobiles, mettant au centre de leur développement les notions d'expérience et d'interface utilisateur pour faciliter accès et manipulation des corpus visuels, le rôle de chacun des protagonistes classiques du secteur s'est peu à peu transformé. Les frontières entre ceux qui fournissent de la « matière » (traditionnellement les GLAM) et ceux qui construisent des corpus, en regroupant des ensembles isolés, en les documentant et en les analysant avec des outils *ad hoc* (traditionnellement le monde de la recherche), s'estompent de plus en plus. À titre d'exemple, la numérisation peut ainsi être déportée de plus en plus directement vers ceux qui traditionnellement en faisaient la demande. Le chercheur peut recourir à des applications telles que PhotoScan, ou encore à la ScanTent, un dispositif portatif composé d'une tente de

33. Cf. https://uclab.fh-potsdam.de/projects/vikus-viewer/

34. Voir par exemple le Digital Research Tools (DIRT) Directory sur le site internet Digital Humanities at Berkeley : https://digitalhumanities.berkeley.edu/resources/digital-research-tools-dirt-directory ; ou encore un « Call for Reviews : Tools and Environments for Digital Scholarly Editing » de l'Institut für Dokumentologie und Editorik : https://www.i-d-e.de/cfr-tools/.

numérisation mobile et d'une application dédiée[35]. Certaines institutions culturelles ont par l'intermédiaire de laboratoires *in situ* engendré des collaborations de plus en plus étroites avec la recherche. C'est par exemple le cas dans plusieurs bibliothèques nationales qui intègrent un « laboratoire » en propre (KB Lab, Library of Congres Lab) ou dans des bibliothèques universitaires qui portent un *Center for Digital Scholarship*, en particulier dans le monde anglo-saxon. Un outil comme Tropy[36] rebat également les cartes car il permet aux chercheurs, ou du moins à celui qui crée, consolide un corpus à la suite de différents dépouillements de fonds d'archives, de les décrire en y associant des métadonnées. Finalement, le chercheur devient peu à peu expert des questions de métadonnées et construit déjà un ensemble possiblement diffusable, articulant métadonnées et numérisation (voir à ce sujet le plugin permettant de « relier » un projet Tropy et la plateforme de publication Omeka S).

11. Par ailleurs, de nouveaux acteurs sont apparus dans cet écosystème, qu'il s'agisse des usagers finaux ou des fournisseurs de contenus. Du côté des premiers, le désir – ou plutôt la nécessité – d'éditorialisation de contenus de plus en plus massifs, dans les bibliothèques numériques, peut être vu comme une opportunité afin d'ébaucher des corpus d'images appréhendables par différents publics, non plus seulement les chercheurs mais aussi par exemple les scolaires ou les amateurs[37]. Du côté des seconds, les fournisseurs de contenus, on trouve par exemple Wikimedia Commons[38], sorte de banque d'images en ligne dont le contenu est librement réutilisable et à laquelle tout un chacun peut contribuer. Libre à l'utilisateur de créer ensuite à partir des millions d'item proposés un corpus plus délimité en fonction de ses intérêts. Il existe également des projets à une échelle plus réduite. Ainsi, l'association Proscitec[39], consacrée à la valorisation du patrimoine des métiers et des industries dans la région Hauts-de-France, ouvre sa plateforme d'inventaire en ligne à des particuliers, comme le prouve la collection des couvertures de cahiers, véritable corpus visuel à nos yeux. Bien que l'inventaire soit assez rudimentaire dans l'approche documentaire, il permet néanmoins de rendre compte de l'articulation entre corpus visuels et métadonnées aujourd'hui disponibles à tous. Sans l'évoquer en détail, l'ensemble des démarches liées au *crowdsourcing* a fait entrer dans la phase d'élaboration d'un corpus (notamment pour les opérations de mises en données afin de les rendre plus exploitables) un public varié, à la fois éloigné des pratiques dites « métiers » des institutions culturelles, mais également des objectifs des projets de recherche. On peut citer la plateforme de *crowdsourcing* Zooniverse, dont l'objectif est de proposer une solution qui permette à toute personne bénévole de contribuer

35. Projet *Read*, Computer Vision Lab, université technique de Vienne et université d'Innsbruck. 2018. « The ScanTent ». https://scantent.cvl.tuwien.ac.at/en/.

36. Cf. http://tropy.com/

37. Voir par exemple les applications collaboratives pour étudier les images développées dans le cadre de l'ANR *Visuall* et de la plateforme « Sciences et cultures du visuel » : http://www.visuall-tek.org/.

38. Cf. https://commons.wikimedia.org/wiki/Accueil

39. Cf. https://www.inventaire.proscitec.asso.fr/musee/collection-p-m-courtin/

à des projets de recherches conduits par des chercheurs professionnels ayant besoin d'aide[40] : a priori orientée vers les sciences dites dures dans le monde anglo-saxon, elle a été utilisée par un service d'archives municipales pour un projet de transcription des registres paroissiaux d'Ancien Régime[41] ainsi que par l'INHA pour aider à identifier les œuvres antiques dessinées par Jean-Baptiste Muret au XIXe siècle[42]. Quant à la BNF, elle a mené une expérience de *crowdsourcing* nativement intégrée dans sa bibliothèque numérique (et donc non pas via un service tierce) en invitant tout un chacun à géolocaliser des documents numérisés, ce qui permet ensuite d'enrichir l'expérience utilisateur en proposant à ce dernier un affichage cartographique des résultats de ses requêtes[43].

Conclusion : le millefeuille informationnel de l'artefact culturel

12. Dans ces différents projets où les frontières entre acteurs s'estompent, il s'agit alors de créer un cercle vertueux de circulation et d'enrichissements mutuels des données sur un même objet d'étude. C'est précisément l'un des éléments majeurs pour une conception à 360° des corpus numériques que nous pouvons formaliser dans ce que nous appelons le « millefeuille informationnel de l'artefact culturel[44] », présenté dans la figure ci-dessous (figure 1). Sont représentées les différentes strates d'information que peut porter un objet culturel et ce au prisme de sa numérisation, de sa mise en données et de son exploitation en ligne.

13. Le graphique se lit de bas en haut. La première étape consiste à créer le pendant numérique d'un objet patrimonial en le numérisant en 2D ou en 3D (strate grise). La numérisation s'accompagne de données descriptives (strate jaune) effectuées par l'institution qui possède l'artefact en fonction de normes métiers (inventaire réglementaire dit à 18 colonnes pour les musées, catalogue dans un système de gestion pour une bibliothèque, description en XML*-EAD* pour un centre d'archives, etc.). Il est possible d'y adjoindre (strate gris foncé) des données issues d'instruments de mesures physico-chimique (par exemple effectuées par le C2RMF) qui peuvent compléter des éléments consolidés ou non dans les notices d'œuvres de l'inventaire informatisé. Le monde universitaire produit dans le cadre de programmes de recherche des contenus sur les artefacts. Ces contenus peuvent prendre la forme de données structurées (strate bleu foncé) comme une base de données thématique intégrant l'artefact dont les éléments d'identification, l'iconographie ou encore l'historique de l'attribution sont détaillés. À cela peuvent s'ajouter des données non structurées (strate blanche entourée de bleu) correspondant à des contenus éditoriaux (articles scientifiques, catalogues d'exposition…)

40. Cf. https://www.zooniverse.org/
41. Cf. https://www.zooniverse.org/projects/archivesdesaint-brieuc/family-certificates
42. Cf. https://www.zooniverse.org/projects/inha/digital-muret
43. Cf. https://arpenteur.bnf.fr/#/about
44. Nous utilisons ici le terme d'artefact au sens du CIDOC (E22 « Man-Made Object »).

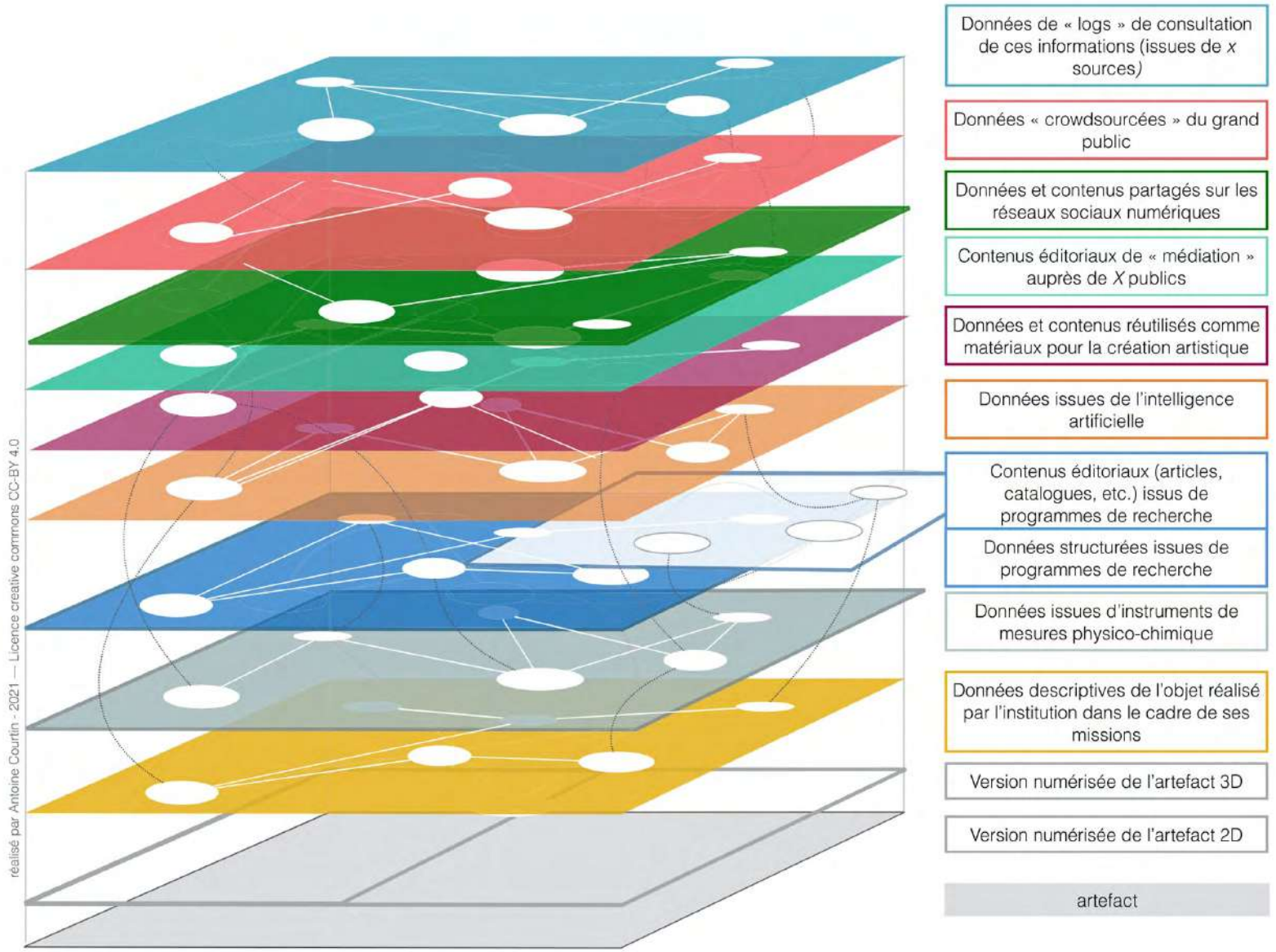


Figure 1. Millefeuille informationnel de l'artefact culturel
Crédit : Antoine Courtin

mobilisant cet artefact dans leur argumentaire. Les avancées spectaculaires de la vision assistée par ordinateur et les technologies dite de *deep learning* permettent de proposer des indexations alternatives par zone de l'image et d'adjoindre des données réalisées par les machines (strate orange). Les données peuvent être également réutilisées comme des matériaux pour la création artistique (strate violette). De nombreux musées produisent des contenus dits de médiation afin de préparer, accompagner ou poursuivre la visite (le fameux « avant/pendant/après ») des différents publics (strate vert clair), voire de les mettre en forme pour répondre aux usages spécifiques des réseaux sociaux afin d'en accroître leur dissémination (strate vert foncé). Au côté des données produites par les GLAM détentrices des objets et par le monde académique, de plus en plus de projets font appel à l'intelligence collective en invitant à participer à l'amélioration des données de description grâce au *crowdsourcing* (strate rose). La dernière strate (bleu clair) contient toutes les traces d'utilisation ou de consultation des différentes couches informationnelles mentionnées ci-dessus, soit les données d'usage.

14. Cette figure complexe est un modeste outil de réflexion mais surtout une proposition invitant toutes critiques, afin de mieux comprendre – et visualiser – les différentes natures des données attachées aux éléments d'un corpus visuel basé sur la numérisation d'artefacts culturels. La question qui reste à soulever réside dans l'articulation de ces différentes strates informationnelles : faut-il les rassembler dans une unique et même interface où l'usager pourra, à la manière de calques à cocher dans un logiciel de PAO, afficher à sa guise telle ou telle strate (si, par exemple, je souhaite voir les données du musée et de la recherche universitaire mais pas le podcast sur cet objet qui a été réalisé pour l'audio-guide scolaire), et si oui, qui peut, ou qui doit être porteur de ce millefeuille ?

Le musée comme service d'information. Pour une politique des interfaces muséales

Emmanuel Château-Dutier

1. Au cours de la dernière décennie, nombre de musées d'art européens, rapidement rejoints par plusieurs grands musées américains, se sont fortement engagés dans un processus d'« ouverture de leurs collections » qui s'est traduit par la mise à disposition des informations documentaires sur leurs artefacts et la publication des images des œuvres qu'ils conservent, en ayant de plus en plus largement recours à des licences libres[1]. Il ne s'agit donc pas tant d'évoquer ici les catalogues de musées en ligne, les applications pour terminaux mobiles, ou les déploiements d'expositions virtuelles, mais plutôt des formes plus spécifiques à l'informatique comme la mise à disposition de jeux de données, ou la création d'interfaces programmables (API*) qui peuvent directement intéresser les historiens d'art. L'aménagement de ce type d'accès oblige en quelque sorte à penser aujourd'hui le musée comme un service d'information. Cette mise à disposition des ressources des collections ne va pas sans soulever des enjeux politiques pour les institutions patrimoniales. Elle suppose que les musées envisagent, tant dans leur gestion que dans leur organisation, les nouvelles modalités qu'implique leur publicisation[2]. Elle soulève surtout des questions politiques relatives au positionnement de ces établissements qui m'amènent à vouloir proposer l'esquisse de ce que l'on pourrait désigner comme une politique des interfaces.

2. Ce faisant, il s'agit d'essayer d'aller au-delà de la seule discussion militante portant sur la question des contenus ouverts (*open content*) ou des données ouvertes (*open data*). Non pas que je ne sois pas convaincu, en tant que chercheur et historien de l'art, qu'une telle approche ne soit plus nécessaire ou encore particulièrement utile, mais à défaut de pouvoir entièrement adopter le point de vue du musée, il paraît important d'essayer de réinscrire l'action des institutions culturelles dans un contexte décisionnel politique plus large pour mieux comprendre comment elles se trouvent parfois confrontées à des injonctions contradictoires. Aborder le sujet de cette façon implique d'essayer de faire l'histoire des politiques de numérisation, mais aussi sans doute d'aborder la numérisation de manière critique[3]. Pour ce faire, il est possible de se référer à la recherche

1. Le glossaire de l'ouvrage offrira au lecteur plus d'explications sur l'Open access ainsi que sur les licences Creative Commons aux entrées correspondantes.

2. Voir sur ce sujet le chapitre « Pour un regard à 360 degrés sur les corpus visuels : pratiques de mise à disposition et de réutilisation » d'Antoine Courtin dans cet ouvrage.

3. Ce travail s'inscrit dans une recherche en cours plus large menée en tant que spécialiste de la muséologie numérique à l'Université de Montréal. Elle est évidemment informée de manière directe par ma pratique d'historien de l'art.

pionnière de Ross Parry (2007) sur l'informatisation de la sphère muséale et à de nombreux travaux récents dans le domaine de la muséologie numérique (Marty et Jones 2007 ; Cameron et Kenderdine 2010 ; Parry 2009 ; Hamilton et Saunderson 2017 ; Drotner *et al.* 2018 ; Lewi *et al.* 2019), mais peut-être aussi de tirer parti des travaux développés ces dernières années dans le cadre des *critical data studies* (Winner 1980 ; Iliadis et Russo 2016 ; Dalton, Taylor et Thatcher 2016).

3. Aborder les conditions politiques de l'ouverture des collections, c'est prendre acte du fait que cette question n'est pas seulement technique ou juridique, mais avant tout sociale et politique. En tant qu'artefacts numériques, on peut dire avec Langdon Winner (1980) que les interfaces constituent des arrangements techniques. Cette notion renvoie à celle de dispositif chez Michel Foucault (Agamben 2007). Il s'agit tout autant de considérer les objets techniques comme les produits d'un ensemble d'arrangements de pouvoirs et d'autorités survenant dans les organisations humaines, que les activités qui interviennent au sein même de ces arrangements. La production des objets techniques est une manière de construire de l'ordre dans notre monde (Iliadis et Russo 2016). On peut ainsi envisager l'ouverture des collections patrimoniales comme un processus pour comprendre, s'engager et expérimenter avec la manière dont les biens culturels sont présentés et disséminés. Aussi, aborder de manière généalogique l'histoire de cette numérisation patrimoniale et de la revendication d'ouverture peut permettre de déterminer des discours séparés et parfois conflictuels qui président à l'ouverture des données culturelles.

4. La situation européenne à l'égard de ce qu'on appelle couramment l'Open GLAM fait depuis de nombreuses années l'objet d'une veille soutenue et d'une forte mobilisation de la société civile[4]. On peut dorénavant affirmer que l'ouverture des données des collections commence à être bien installée dans le secteur muséal. Même si la France accuse beaucoup de retard sur la question des droits photographiques, l'accès ouvert aux collections s'étend rapidement, qu'il s'agisse des données ou des contenus. L'intérêt de ces politiques ne fait aucun doute pour les historiens d'art et les chercheurs. En ménageant un large accès aux données et aux reproductions, elles offrent aujourd'hui la possibilité d'agréger des informations et des contenus issus de multiples collections à travers le monde. Comprendre la mise en place de ces politiques implique de les replacer dans le contexte plus large des politiques européennes. Cependant, il est intéressant de se demander ce que cette ouverture fait aux collections proprement dites. Que devient la collection dans un contexte distribué ? Comment le musée doit-il reconsidérer ses missions en regard à cette ouverture des collections ?

4. GLAM pour *Gallery, Libraries, Archives and Museums*.

Une décennie d'ouverture du patrimoine culturel

5. Depuis les années 2000, la production d'état des lieux a typiquement accompagné le mouvement d'ouverture des collections. Alors que plusieurs grands musées internationaux ont fait événement à partir du début de la décennie 2010 en s'engageant à rendre largement accessibles les données et les contenus de leurs collections, divers *census* se sont depuis intéressés, non seulement aux licences, mais aussi aux formats et aux modalités de publication.

6. Les premiers grands recensements sur les droits d'auteur servirent en particulier aux historiens d'art à adresser la question – cruciale pour la discipline – de la disponibilité des images d'art en ligne et des droits de réutilisation appliqués par les musées sur les œuvres du domaine public. Parmi ceux-ci, il faut évoquer la grande étude souvent citée de Kenneth Crews (2010), ou encore la publication de rapports récurrents par la College Art Association sur les droits d'auteur et l'histoire de l'art. Plus récemment, le recensement collaboratif réuni par Douglas McCarty et Andrea Wallace, « *Survey of GLAM open access policy and practice* », a permis de se faire une meilleure image des progrès réalisés sur le front de la mise à disposition de jeux de données muséaux[5]. Dernier avatar de cette démarche, l'excellent rapport *Images et Usages* commandé par la Fondation de France, dont la rédaction a été dirigée par Martine Denoyelle *et al.* (2018) à l'INHA, dresse un bilan critique de la disponibilité des contenus muséaux en France.

7. Dans son acception la plus simple, ouvrir signifie rendre disponibles les données ou les contenus en vue de leur réutilisation en dehors de l'institution qui les a créés. En tant que chercheur, nous sommes évidemment particulièrement intéressés par les possibilités et les collaborations nouvelles que cette ouverture introduit et notamment les perspectives qui dépassent l'échelle stricte de la collection. Depuis les années 2010, les choix volontaristes opérés par un certain nombre de grandes institutions culturelles ont profondément changé la donne du point de vue du partage de l'information patrimoniale. Prolongeant la politique d'ouverture des métadonnées* entamée autour du tournant de la décennie, plusieurs établissements ont également mis à disposition du public, avec plus ou moins de restriction, les images de leurs collections.

8. De fait, les concepts liés à l'accès ouvert naquirent à la croisée des domaines informatiques et universitaires avec les déclarations de chercheurs sur l'Open Access de Budapest en 2002, le *Bethesda Statement on Open Access Publishing* de 2003 puis la Déclaration de Berlin sur l'Accès ouvert du 22 octobre 2003 et toutes les déclarations subséquentes. Mais là où « une vieille tradition et une nouvelle technologie avaient convergé pour apporter un bénéfice public sans précédent. » (« Budapest Open Access Initiative » 2002), les musées ne pouvaient pas vraiment se réclamer de la même tradition du partage.

5. McCarthy, Douglas et Andrea Wallace. 2018. « Survey of GLAM open access policy and practice ». Consulté le 10 avril 2020. http://bit.ly/OpenGLAMsurvey.

Au contraire, c'est plutôt le *connoisseurship* qui caractérise le domaine artistique et culturel. En dépit du renouveau de la muséologie des années 1970, les habitudes avaient la vie dure et les musées ont légitimement pu être identifiés comme des organisations autoritaires, thésaurisant un patrimoine précieux et difficile d'accès. Quoiqu'il en soit, il faut relever que la déclaration de Berlin sur l'accès ouvert au savoir scientifique de 2003 souhaitait déjà « encourager les détenteurs du patrimoine culturel à soutenir l'*open access* en mettant à disposition leurs ressources sur l'internet[6] » (« The Berlin declaration on Open Access to Scientific Knowledge » 2003). Une incitation renouvelée dans la « Déclaration de Lyon sur l'accès à l'Information et au Développement » (2014).

9. Ce sont en premier lieu les bibliothèques qui ont agi de manière volontariste dans le domaine de l'ouverture des données culturelles. Dès 2010, la Bibliothèque nationale d'Allemagne (DNB) ouvrait le banc en publiant son catalogue d'autorité sous la forme de données ouvertes et liées[7]. En juillet de l'année suivante, ce fut la bibliographie nationale britannique qui lui emboîtait le pas avec l'ouverture, sous licences CC0 (« *No Rights Reserved* »), de plus de deux millions huit cent mille notices. Puis, en septembre 2011, la Conférence des bibliothèques nationales européennes, à l'occasion de son vingt-cinquième anniversaire, adoptait à Copenhague une motion supportant l'utilisation de licences libres pour les données de catalogage.

10. Ce mouvement d'ouverture a entraîné une multiplication des interfaces pour la mise à disposition des informations sur les collections, parfois sous la forme de simples versements de jeux de données sur des plateformes de partage de code informatique, parfois par l'intermédiaire de dispositifs qui permettent à des clients informatiques de se connecter à des images techniques par la voie programmatique. Lorsque des sociétés comme Google ou Facebook donnent accès à des ressources, celles-ci les mettent généralement à disposition par l'intermédiaire d'Interfaces programmables (API). L'API devient ainsi un type spécifique d'artifice communicationnel qui définit le genre d'interactions qu'un utilisateur peut avoir avec les données, et conséquemment le type de service proposé (Goyet 2017 ; Francony 2018).

11. Suivant ce modèle, en mars 2009, le Brooklyn Museum fut sans doute l'un des premiers musées à rendre les données de sa collection lisibles par la machine en proposant une API qui fit émerger trois projets construits à partir d'elle en quelques jours. En 2011, le Rijksmuseum d'Amsterdam lançait, à son tour, une interface programmable, et la même année le British Museum inaugurait un service de données ouvertes et liées (*linked open data*[8]). À la même époque, Europeana développait un programme d'acces-

6. C'est nous qui traduisons : « *encouraging the holders of cultural heritage to support open access by providing their resources on the Internet.* »

7. « Linked Data Service ». *Deutsche National Bibliothek*. Consulté le 2 avril 2020. https://www.dnb.de/EN/Professionell/Metadatendienste/Datenbezug/LDS/lds_node.html.

8. « Our data in a nutshell ». *Rijksmuseum*. Consulté le 2 avril 2020. https://www.rijksmuseum.nl/en/data/overview.

sibilité sémantique en déployant son Europeana Data Model (EDM) pour plus de vingt millions d'objets conservés dans plus de 1500 musées, bibliothèques, archives et collections audiovisuelles à travers l'Europe[9]. Parallèlement, nombre de musées adoptaient la plateforme de partage GitHub pour diffuser les données de leur collection (le Powerhouse Museum dès 2009, en 2012 le Cooper Hewitt, en 2013 la Tate, ou encore en 2015 le MoMA).

12. En France, à partir de (2014), la « Feuille de route stratégique : Métadonnées culturelles et transition Web 3.0 » permit de soutenir une approche transversale fondée sur les technologies du Web sémantique* et du Web des données en priorisant des axes de travail stratégiques afin d'identifier les actions à initier sur le front des identifiants pérennes pour les ressources culturelles, l'identification des auteurs de ressources culturelles, l'interconnexion sémantique des grands référentiels (*graphe culture*), l'interaction avec les publics et les potentialités du Web 3.0, la traçabilité ou la sensibilisation des institutions culturelles et patrimoniales. De même, le travail engagé à partir de 2015 avec les technologies du Web sémantique allait bientôt permettre à l'American Art Collaborative de proposer un accès commun à des collections conservées dans chacune des collections des musées partenaires[10].

13. « Libre » fut incontestablement l'un des principaux mots-clés des années 2012-2013 dans le monde du patrimoine. Si dès 2008, le Powerhouse Museum de Sydney avait joint le projet frère de *Wikipedia*, The Commons, il restait relativement isolé dans cette voie. En 2012, l'accord du Walters Art Museum avec l'encyclopédie libre, puis la publication d'un premier ensemble de 125 000 œuvres librement téléchargeables sur le site du Rijksmuseum inauguraient une ère nouvelle. En juin 2013, la National Gallery de Washington lançait un nouveau site web et mettait à disposition, en accès libre, 25 000 images en haute résolution (le site en propose aujourd'hui 32 000). En août et octobre 2013, le Getty Research Institute mettait quant à lui à disposition un premier lot de 4 500 images par l'intermédiaire de son programme de contenus ouverts (sous licences CC-BY), pour une utilisation sans restriction qui s'élève aujourd'hui à plus de 100 000. En 2017, le Metropolitan Museum of Art publiait quant à lui plus de 375 000 images en CC-0 (Denoyelle *et al.* 2018 ; Hamilton et Saunderson 2017).

14. Même si parfois les licences choisies par ces institutions restent problématiques parce qu'elles relèvent encore trop souvent d'un *copyfraud* (Mazzone 2015), cette multiplication des interfaces offre de nouvelles conditions d'accès aux collections. D'autant que ces évolutions s'accompagnent de la promotion de protocoles comme ceux développés par l'International Interoperability Framework (IIIF) pour le partage d'images[11]. De tels protocoles défi-

9. « Europeana Data Model ». *Europeana Pro*. Consulté le 2 avril 2020. https://pro.europeana.eu/page/edm-documentation.

10. « Linked Open Data Initiative ». *American Art Collaborative*. Consulté le 2 avril 2020. http://americanartcollaborative.org/ et « Linked Art ». Consulté le 2 avril 2020. https://linked.art/.

11. « International Image Interoperability Framework (IIIF) ». Consulté le 12 avril 2020. https://iiif.io/.

nissent des manières de référencer une page d'un ouvrage numérisé, une séquence de pages, ou encore une zone et un cadrage d'une image à des fins de partage. Cet effort de normalisation des interfaces autorise le développement de toutes sortes de clients de visualisation par des tiers. Ils offrent de nouvelles manières de distribuer des images, y compris dans un contexte qui peut être contraint par des limitations concernant les droits d'auteur. De telle sorte que les institutions ne peuvent plus se contenter de se réfugier derrière la barrière du copyright pour ne pas diffuser leurs collections puisqu'elles peuvent aujourd'hui le faire à distance de manière contrôlée.

15. L'émergence de ce nouvel écosystème technique a permis le développement de nombreux outils pour utiliser ces ressources qui permettent dorénavant de travailler en réseau sur les métadonnées et les images d'artefacts distribuées dans plusieurs collections sur le Web. Cet environnement logiciel reposant sur ces interfaces est tout à fait prometteur pour les historiens de l'art. À titre d'exemple, on peut citer ResearchSpace développé au British Museum[12], le Getty Scholars' Workspace (Cuno 2016), la visionneuse Mirador produite par Stanford pour travailler sur les images par l'intermédiaire du protocole IIIF[13], et plus récemment l'Art Image Exploration Space (ARIES) qui a reçu le financement d'un généreux donateur par l'intermédiaire de la Fricks[14]. L'outil de publication web Wax, conçu selon les principes du *minimal computing*, démontre quant à lui que ces technologies peuvent facilement faire l'objet d'appropriations commodes[15]. Tout semble donc réuni pour l'avènement d'un environnement de travail renouvelé à même de servir directement les besoins des chercheurs.

16. À bien des égards, il devient possible d'affirmer que cette multiplication des interfaces reconfigure la signification des collections. La matérialité de ces procédures techniques et ses effets sont des réalités qui furent par exemple explorées dans une expérimentation conduite par le MetaLab en collaboration avec le musée d'Harvard (Battles et Maizels 2016). Ces interfaces autorisent aussi des exploitations inédites des collections à l'instar de la collaboration engagée par le MET avec Microsoft pour de l'étiquetage automatisé[16] ou celle du MoMA avec Google sur ses catalogues d'exposition en utilisant des technologies d'intelligence artificielle[17]. À une échelle plus large, elles font émerger un nouveau contexte distribué qui oblige les musées à affronter la question de l'éclatement de la collection.

17. Un peu partout à travers le monde, plusieurs musées de premier plan ont donc rejoint un mouvement global d'ouverture de leurs bases de données en utilisant des proto-

12. « ResearchSpace ». Consulté le 12 avril 2020. https://www.researchspace.org/.
13. « Mirador ». Consulté le 12 avril 2020. https://projectmirador.org/.
14. « ARIES ARt Image Exploration Space ». Consulté le 12 avril 2020. https://artimageexplorationspace.com/.
15. « Wax ». Consulté le 12 avril 2020. https://minicomp.github.io/wax/.
16. « The Met x Microsoft x MIT ». *The Metropolitan Museum of Art*. Consulté le 12 avril 2020. https://www.metmuseum.org/about-the-met/policies-and-documents/open-access/met-microsoft-mit.
17. « MoMA & Machine Learning ». 2018. *Experiments with Google*. Consulté le 12 avril 2020. https://experiments.withgoogle.com/moma.

coles techniques pour le partage des contenus numérisés. Même si ces services restent encore sous-utilisés, on peut dire que cette publicisation offre des perspectives particulièrement prometteuses pour les historiens de l'art. En fait, la question qu'il faut plutôt se poser aujourd'hui est celle de savoir pourquoi aussi peu de musées ont développé jusqu'à présent ce type d'accès. Ces outils sont encore peu investis par les institutions muséales françaises, par exemple. Or, répondre à cette question réclame de ne pas seulement s'intéresser aux problématiques juridiques souvent invoquées en la matière. Il faut encore envisager la question du point de vue des motifs qui ont présidé à la numérisation de ces collections et à la mise en place de ces démarches d'ouverture car celles-ci sont marquées par une certaine ambivalence.

Un détour par les politiques européennes

18. Un détour par l'histoire s'impose pour comprendre à quel point l'engagement des institutions culturelles dans des campagnes massives de numérisation découle directement de politiques menées à l'échelle européenne (Shore 2000). Marquant la création de la Communauté économique européenne, le traité de Rome de 1957 était pour l'essentiel un projet de nature économique. Mis à part deux brèves références dans son article 36 concernant les restrictions de mouvement de biens patrimoniaux lorsqu'ils sont d'intérêt national, l'accord ne déterminait rien en matière de politique culturelle. Néanmoins, on peut faire remonter le discours sur un patrimoine européen commun à 1954, année de signature de la Convention culturelle européenne. Celle-ci constitua le premier accord officiel sur les questions culturelles à l'échelle européenne et le point de départ des politiques de coopération culturelle (López 1993 ; Sassatelli 2002 ; Paganoni 2015).

19. Dès la fin des années 1970, la politique d'intégration européenne avait identifié la politique culturelle comme levier. Ce fut l'accord de Maastrich qui fit de la culture une réelle compétence légale à partir de 1992. Non sans discussions d'ailleurs, car la culture n'apparaissait pas de toute évidence comme un domaine destiné à être harmonisé, puisque pour nombre d'acteurs, il s'agit plutôt du champ de la différence et des particularismes (Bennett 2008). En 1996, le « Premier rapport sur la prise en compte des aspects culturels dans l'action de la Communauté européenne » soulignait le rôle qu'était appelée à jouer la culture dans le renforcement d'un modèle européen de société basé sur un ensemble de valeurs communes[18]. L'insistance apportée à la culture constitua un changement fondamental du discours officiel sur l'intégration européenne qui ne pouvait plus simplement résulter de l'intégration économique et d'une harmonisation réglementaire et technique (Shore 2000, 40-86). Nombre de coopérations notamment dans le domaine du numérique sur le partage de l'information muséale tels que les pro-

18. « Premier rapport sur la prise en compte des aspects culturels dans l'action de la Communauté européenne ». 1996. COM_1996_0160_FIN. Bruxelles, Belgique : Commission des communautés européennes. http://op.europa.eu/fr/publication-detail/-/publication/f4f34a44-8769-4e8f-a270-01b5e2a3d3d2.

jets Michaels, eContentPlus, etc. s'engagèrent sur cette base, au cours des années 1990[19].

20. Tout le monde a encore probablement à l'esprit le tournant marquant pris en septembre 2005. Un lobbying bien appliqué conduit notamment par l'ancien président de la Bibliothèque nationale de France, Jean-Noël Jeanneney, déboucha sur la communication de la Commission intitulée « i2010 Bibliothèques numériques » puis l'annonce, le 2 mars 2006, d'une intensification de ses efforts « pour mettre en ligne la "mémoire de l'Europe" via une bibliothèque numérique européenne ». On indiquait que cette bibliothèque reposerait sur l'infrastructure TEL[20], fixant l'objectif de deux millions de documents accessibles en ligne à l'horizon 2008 et de six millions à la fin de 2010 (Jeanneney 2006 ; Valtysson 2012).

21. C'étaient en premier lieu les bibliothèques qui étaient visées, mais les musées étaient déjà explicitement mentionnés dans le texte. Il était non seulement question de numérisation patrimoniale, mais aussi de l'accessibilité en ligne du matériel culturel et de la conservation numérique. Les mesures proposées dans cette recommandation devaient aboutir à la mise sur pied « d'une approche coordonnée parmi les états membres des questions de numérisation, d'accessibilité en ligne et de conservation numérique, et permettre de créer un point d'accès multilingue commun au patrimoine culturel diffus de l'Europe. » Dans la lignée de cette initiative, plusieurs directives et des programmes de financement spécifiques soutinrent ensuite une numérisation large des collections dans divers pays et la contribution à un projet phare à l'échelle européenne comme Europeana (Valtysson 2012).

22. Toutefois, cette politique était d'emblée marquée du sceau d'une certaine ambivalence. L'initiative sur les bibliothèques numériques visait à « permettre à tous les Européens d'accéder à la mémoire collective de l'Europe et de s'en servir à des fins éducatives, professionnelles, récréatives et créatives. » Mais il était explicitement indiqué que « les efforts déployés dans ce domaine contribueront à la compétitivité de l'Europe et étaieront l'action de l'Union européenne en matière de culture. » Les textes de la Commission étaient à cet égard sans équivoque. Si, bien sûr, les mesures recommandées étaient destinées à faire connaître la richesse et la diversité du patrimoine européen et garantir sa préservation, on indiquait qu' :

> 23. Au-delà de sa valeur culturelle intrinsèque, le matériel culturel constitue une ressource importante pour de nouveaux services à valeur ajoutée. Les mesures recommandées contribueront à favoriser la croissance dans des secteurs à haute valeur ajoutée connexes comme le tourisme, l'éducation et les médias. Le contenu numérique de haute qualité est essentiel à certaines activités industrielles à grande échelle (d'où

19. « Shaping Europe's digital future : Timeline of digitisation and online accessibility of cultural heritage ». 2014. *European Commission*. 25 août 2014. Consulté le 2 avril 2020. https://ec.europa.eu/digital-single-market/en/news/timeline-digitisation-and-online-accessibility-cultural-heritage.

20. « La Commission européenne intensifie les efforts pour mettre en ligne la "mémoire de l'Europe" via une bibliothèque numérique européenne ». 2006. Communiqué de presse. IP/06/253. Commission Européenne. https://ec.europa.eu/commission/presscorner/detail/fr/IP_06_253.

l'intérêt de la part des principaux moteurs de recherche). La numérisation et la conservation numérique sont des activités à forte intensité de connaissance qui vont sans doute se développer considérablement dans les années à venir[21].

24. Cette initiative s'inscrivait d'ailleurs dans le cadre de la stratégie de la Commission en faveur de la société de l'information (initiative i2010[22]). Il s'agit notamment « d'optimiser l'utilisation des technologies de l'information aux fins de la croissance économique, de l'emploi et de la qualité de vie. [Et] L'un des principaux objectifs politiques de l'initiative est de rendre le contenu européen plus largement accessible et plus exploitable pour de nouveaux services et produits d'information. »

25. En effet, le modèle de l'ouverture n'est pas totalement neutre politiquement. Dans une monographie récente, *Wikipedia and the Politics of Openness*, Nathanial Tkacz piste les fondements politiques de l'ouverture. Il insiste notamment sur le discours conservateur de Karl Popper dans *The Open Society and Its Enemies* de 1945 (Tkacz 2015), un ouvrage dans lequel Popper réécrivait l'histoire de la philosophie politique autour de nouvelles catégories maîtres que sont l'ouverture et la fermeture. Pour Nathaniel Tkacz, il s'agissait d'une défense du capitalisme qui prenait la forme d'une argumentation sur le savoir. Dans la foulée de Hayek, la société ouverte pouvait, selon Popper, être promue à travers la liberté offerte par la participation à un marché libre. Si cette théorie politique s'applique bien au caractère collaboratif de *Wikipédia*, il n'est pas certain qu'elle puisse servir pour analyser l'ouverture des données culturelle en Europe. Néanmoins, l'action de la Commission européenne dans le domaine numérique paraît fortement marquée par l'idéologie néo-libérale.

26. Pour mieux la comprendre, il faut essayer de restituer cette politique dans le contexte des politiques destinées à développer une économie numérique adossée à l'Internet portée par plusieurs organisations internationales (G7, OMC, Nations Unies). À partir du milieu des années 1990, celles-ci reconnurent que l'information était en train de devenir un bien économique autonome (matérialisé par la dissociation contenant/contenu), et la convergence des communications adressée et flottante, qui débouchaient sur la construction d'un patrimoine collectif (Nora et Minc 1978). En réalité, le mouvement d'ouverture des contenus culturels européens doit être plus largement réinscrit dans le cadre conceptuel de l'avènement d'une « société de la connaissance » (Musso 2002).

27. Un des concepts clefs de l'économie de la connaissance – une expression popularisée en premier lieu en 1959 par le consultant en management Peter Drucker dans son livre *The Landmarks of Tomorrow* – consiste à traiter le savoir et l'éducation, habituellement considérés comme du

21. « Recommandation de la Commission sur la numérisation et l'accessibilité en ligne du matériel culturel et la conservation numérique ». 2006. COM_2006_3808_FIN. Bruxelles, Belgique : Commission des communautés européennes. https://ec.europa.eu/transparency/regdoc/rep/3/2006/FR/3-2006-3808-FR-F-0.Pdf.

22. « Shaping Europe's digital future: European I2010 Initiative on e-Inclusion - to Be Part of the Information Society ». 2007. *European Commission*. 8 novembre 2007. Consulté le 12 avril 2020. https://ec.europa.eu/digital-single-market/en/news/european-i2010-initiative-e-inclusion-be-part-information-society.

capital humain, comme des produits commerciaux qui peuvent être exportés avec un haut retour sur investissement (Powell et Snellman 2004). Ainsi que le relèvent Gary Hall et Janneke Adema : « *Open access is currently being positioned and promoted by policy makers, funders and commercial publishers alike primarily as a means of serving the knowledge economy and helping to stimulate market competition* » (Adema et Hall 2016).

28. Quoiqu'il en soit, cette politique s'avéra assez rapidement un succès. Dès 2011, la plateforme Europeana qui avait été identifiée comme fer de lance de cette action réunit plus de quinze millions de documents. À la suite de l'initiative sur les bibliothèques numériques, le Parlement européen et le Conseil réaffirmèrent leur engagement en mai 2010. Le plan de travail 2011-2014 pour la culture soulignait la nécessité de coordonner les efforts dans le domaine de la numérisation[23]. En 2011, un Comité des Sages escompte rien de moins qu'une « nouvelle Renaissance » grâce à la mise en ligne du patrimoine culturel européen[24]. L'avis va fortement contribuer à relancer l'initiative et recommander des évolutions sur les œuvres orphelines qui s'intégreront dans le cadre plus large de la stratégie numérique pour l'Europe de la commission. Celle-ci visait à accompagner les institutions culturelles dans leur transition vers l'ère numérique et à trouver de nouveaux modèles d'activité efficaces pour accélérer la numérisation, tout en permettant une rémunération juste des ayants droit, le cas échéant.

29. Le soutien à la numérisation patrimoniale participait directement de la constitution d'un marché unique numérique. Il s'agissait de supprimer les entraves pour exploiter pleinement les possibilités offertes par Internet afin d'améliorer l'accès aux biens et services numériques, de créer un environnement propice au développement des réseaux et services numériques. Le numérique était alors considéré comme un moteur de la croissance et l'encouragement de la numérisation devait permettre à l'Europe de rester au premier plan sur la scène internationale en tirant parti de sa richesse culturelle. « Le fait de numériser les ressources culturelles et d'y donner plus largement accès offre d'énormes débouchés économiques et constitue donc une condition essentielle pour développer le potentiel culturel et créatif de l'Europe et renforcer la position de ses entreprises dans ce domaine[25] ».

30. La Commission se proposait donc de coordonner les actions pour éviter les recoupements, et de créer un environnement stable pour permettre à des entreprises d'investir. Elle reconnaissait que le coût de la numérisation de l'ensemble du patrimoine culturel européen

23. « Conclusions du Conseil et des représentants des gouvernements des États membres, réunis au sein du Conseil, sur le plan de travail 2011-2014 en faveur de la culture ». 2010. *Journal officiel de l'Union européenne*. 2 décembre 2010. Consulté le 12 avril 2020 : https://eur-lex.europa.eu/legal-content/FR/TXT/?uri=CELEX%3A42010Y1202%2801%29&qid=1613970575459.

24. « Digital Agenda : « Comité des Sages » calls for a « New Renaissance » by bringing Europe's cultural heritage online ». 2011. Communiqué de presse. IP/11/17. Commission Européenne. http://europa.eu/rapid/press-release_IP-11-17_en.htm?locale=nl.

25. « Recommandation de la Commission du 27 octobre 2011 sur la numérisation et l'accessibilité en ligne du matériel culturel et la conservation numérique ». 2011. *Journal officiel de l'Union européenne*. Consulté le 12 avril 2020. https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32011H0711&from=EN.

était élevé et ne pouvait être uniquement couvert par les moyens publics, ce qui impliquait le développement de parrainages avec le privé ou des partenariats publics-privés malgré la réaffirmation du domaine public. L'emphase fut alors mise sur les agrégateurs nationaux et la création de normes communes de numérisation pour assurer l'interopérabilité* du matériel numérisé et des métadonnées au niveau européen.

31. Europeana résidait au centre de cette politique d'harmonisation culturelle avec un objectif de trente millions d'objets numérisés en 2015. Remarquons ici que la plateforme est caractéristique de la convergence entre musées et bibliothèques rendue possible par la réduction des contenus à un même statut documentaire par la numérisation. Une convergence dont est directement issu l'acronyme GLAM et qui donna lieu à nombre de publications à partir des années 1990, à l'instar de celle de Corinne Welger-Barboza avec son *devenir médiathèque du musée* (Welger-Barboza 2001).

32. L'accès ouvert fut donc présenté et promu par les politiques institutionnelles et les bailleurs de fonds comme un moyen de servir l'économie de la connaissance et de stimuler la compétition du marché. Une vision devenue tellement prévalente, que d'aucuns, d'orientation politique diamétralement opposée, y voit un phénomène qui accompagne un processus de privatisation du savoir (cf. le communiqué récent de la CGT Culture[26]). Ce chapitre n'est pas le lieu pour faire la critique des politiques néo-libérales menées par les institutions européennes depuis les années quatre-vingt. Constatons cependant qu'en découle sans doute un certain nombre d'ambiguïtés concernant la possible commercialisation du domaine public et l'incitation à développer de nouveaux modèles de financement basés sur des partenariats public-privé. L'injonction à l'ouverture et au partage des collections intervenait donc dans une perspective de circulation des contenus relativement équivoque puisque tout en répondant à une forte demande sociale soutenue par un secteur associatif qui militait pour l'ouverture, elle était aussi destinée à alimenter une économie de la connaissance. Directement issue du monde des bibliothèques, cette politique ne reposait que très peu sur des considérations muséales. Pourtant, l'ouverture des collections présente des enjeux très directs pour les musées en ce qui concerne la valorisation de leurs collections. Alors que ces institutions se sont constituées autour de la présentation de « choses vraies » et sont souvent dépositaires d'*unica*, où peut bien être le musée s'il est partout lorsqu'il devient distribué ?

Le musée comme service d'information

33. En 1991, dans un article intitulé « The museum as information utility », George MacDonald (1938-2020) et Stefen Alsford (1952-...) expliquaient que traditionnellement les musées avaient concentré leur attention sur le passé

26. « Open data sur les images des musées et monuments : non au pillage de notre patrimoine par Google ». 2019. *La CGT Culture* (blog). 16 octobre 2019. https://www.cgt-culture.fr/open-data-sur-les-images-des-musees-et-monuments-non-au-pillage-de-notre-patrimoine-par-google-15563/.

(G. MacDonald et Alsford 1991). Cette préoccupation à l'égard des restes matériels du passé expliquait en partie leur focalisation sur les objets. Une orientation que reflète l'énumération des diverses définitions traditionnellement associées aux musées : collections, préservation, étude, exposition, interprétation. Un ensemble d'activités qui se pratiquent généralement sur site. D'après les auteurs, cette « focalisation introvertie » avait en quelque sorte convaincu les musées que leur raison d'être était leurs collections d'artefacts plutôt que de se considérer comme un outil à travers lequel nous apprenons à propos du passé[27]. Une situation qui explique aussi largement que les musées aient acquis « l'image populaire d'institutions caractérisées par les interdictions, archivant les reliques d'un passé mort, seulement accessible à une élite esthétique ou intellectuelle » (G. MacDonald et Alsford 1991).

34. Dans cet article, qui paraît après-coup particulièrement visionnaire compte tenu des développements récents de la muséologie mondiale, les auteurs proposaient de ne plus seulement servir la collection, mais la société. Selon eux, l'évolution du contexte social et politique, tout autant que celle des technologies justifiaient déjà que les institutions répondent mieux aux besoins informationnels de la société. Et de réclamer d'envisager le musée dans la perspective non pas strictement de ses fonctions, mais de ses produits qu'ils détaillent de la manière suivante :

- la génération d'informations résultant d'activités de recherche
- la perpétuation d'informations sur les collections
- l'organisation des relations entre des éléments discrets d'information, y compris la recontextualisation d'artefacts
- la dissémination

35. À partir de ce constat, George MacDonald et Stefen Alsfold proposent « un cadre d'opération global », « une théorie d'opération unifiante reflétant le cycle de vie de l'information à travers l'ensemble des départements du musée ». Si le rôle d'interface du musée entre le visiteur et l'objet est bien identifié depuis longtemps, ce qui est particulièrement intéressant ici, c'est que les auteurs, directement influencés par le théoricien des médias Marshall McLuhan (1911-1980) (G. MacDonald 1987 ; G. MacDonald et Alsford 1989), envisagent le rôle des musées dans le contexte d'une société de l'information.

36. Or, une telle conception du musée peut particulièrement être utile pour penser aujourd'hui une politique des interfaces qui répondent aux enjeux de l'institution muséale. Même si elle est fondamentalement managériale dans son énonciation, envisager le musée comme un service d'information rejoint très directement les revendications sur l'imputabilité sociale des musées qui agitent actuellement les institutions. Dans une enquête menée en 2016 auprès des directeurs de grands musées sur l'avenir des musées au Canada, tous signalent l'engagement avec le public, le rôle communautaire, et la

27. Une question qui anime toujours la communauté muséale, en témoignent les débats récents sur la définition des musées à l'ICOM.

demande d'une autorité partagée comme des questions cruciales pour leurs institutions (Maguire 2016). Le fait de rendre le patrimoine numérisé accessible à tous a souvent été associé à une démocratisation. Il est vrai que l'avènement des technologies numériques, et en particulier des réseaux sociaux, a rendu le patrimoine de plus en plus pluraliste et moins dépendant des experts (Witcomb 2006), mais sans la création d'interfaces les musées gardent le contrôle.

37. Nous avons montré combien l'ouverture numérique des collections en Europe a largement été le produit d'une politique volontariste de la Commission européenne. En ménageant des accès techniques qui permettent de manipuler ces contenus, les politiques conduites depuis la première décennie des années 2000 laissent entrevoir aujourd'hui de riches possibilités d'utilisation des collections pour les historiens de l'art. Ces accès offrent la possibilité de mobiliser les œuvres en dehors du cercle étroit de la collection et à l'appui d'une pluralité de discours. Du point de vue des usages, l'intérêt de ces approches est sans doute fortement marqué par un effet d'échelle : plus ces interfaces seront nombreuses, plus elles deviendront utiles. Ainsi, il est probablement trop tôt pour les institutions pour pouvoir réellement juger de l'efficacité de telles approches et de leur intérêt pour le grand public et les spécialistes.

38. Mais pour que les musées puissent massivement s'engager aujourd'hui dans l'ouverture et la mise à disposition de leurs collections par l'intermédiaire de dispositifs techniques, il nous semble qu'il leur faut adopter un nouveau modèle d'opérations qui leur permette de pleinement intégrer la nouvelle donne provoquée par l'éclatement qu'engendre l'ouverture de leurs collections. Si l'on souhaite réellement voir ce genre de points d'accès se multiplier, il est donc devenu impossible de seulement se focaliser sur la question des licences et des droits. Les questions juridiques et techniques ne suffisent pas à expliquer la faible proportion d'institutions qui met à disposition ses collections par le biais d'interfaces programmables. On ne peut donc plus faire l'économie d'une réflexion sur les conséquences de l'ouverture sur l'identité de leur collection.

39. Bien que les musées se soient largement saisis des opportunités de financement pour la numérisation, à l'exception de quelques-uns particulièrement avancés sur la question, ceux-ci ont rarement véritablement intégré cette ouverture dans leurs missions. De ce point de vue, l'ambivalence de la politique européenne tant sur le financement que sur la finalité des campagnes de numérisation patrimoniales n'a probablement pas facilité cette transformation. Envisager le musée comme service d'information permet toutefois d'imaginer une véritable politique des interfaces. Il s'agit donc de comprendre ce qui peut accompagner cette évolution en faisant en sorte que ce modèle puisse s'aligner avec les valeurs portées par l'institution et le projet d'établissement et prendre conscience des tensions en jeu dans les institutions pour faire avancer la belle promesse d'un accès riche et généralisé aux œuvres des collections qui est à portée de main.

Archivage du Web, un enjeu de gouvernance (d'Internet)

Francesca Musiani

Introduction

1. La plupart des institutions de collecte des archives du Web livrent en ligne un aperçu de leurs périmètres et choix de collecte, à l'instar de la BNF[1] qui distingue des collectes larges et des collectes ciblées. Par ailleurs les chercheurs ont aussi le souci d'essayer de documenter ces sélections et leurs évolutions, que ce soit en ouvrant les boîtes noires de l'archivage (Schafer, Musiani et Borelli 2016) ou en suivant les traces visibles que ces archives livrent[2].

2. En effet, non seulement les institutions, quand elles s'inscrivent dans un cadre juridique fixé, doivent faire porter leurs efforts sur un périmètre défini de sites web, mais aussi mettre en place une stratégie de collecte (en termes de récurrence, de profondeur de l'archivage des sites, de participation ou pas des internautes, etc.) qui aura un impact direct sur la représentativité de ces archives. En outre, des barrières à l'archivage peuvent apparaître, notamment pour des raisons techniques (*captcha*, mots de passe), tandis que les réseaux socio-numériques renouvellent aussi les questions de sélection et de capture. Autant d'éléments liés à la problématique de la gouvernance des archives du Web qui se pose dès lors comme un microcosme de questions plus larges de gouvernance d'Internet[3].

Des archivages en constante évolution

3. Une archive du Web est loin d'être un objet statique[4] : elle évolue sous l'effet des modalités de collecte, de la profondeur de l'exploration, ainsi que des changements au fil du temps de caractéristiques techniques – et, bien sûr, des modèles et paradigmes qui sous-tendent l'archivage.

4. Lors de l'assemblée générale de l'International Internet Preservation Consortium (IIPC) de 2014, Louise Merzeau soulignait à quel point, malgré l'histoire jusqu'ici brève de l'archivage du Web, on avait déjà pu assister à plusieurs changements aux conséquences de taille pour les

1. Voir sur le site de la BNF : https://www.bnf.fr/fr/le-depot-legal-numerique.
2. Voir (Ben-David et Amram 2018) sur le Web archivé nord-coréen.
3. Ce chapitre se fonde sur « Où commence et s'arrête l'archive ? » du livre *Qu'est-ce qu'une archive du Web ?* que l'auteur a publié début 2019 avec Valérie Schafer, Camille Paloque-Bergès et Benjamin Thierry (Marseille, OpenEdition), disponible en accès libre intégral : books.openedition.org.
4. Cette section reprend des éléments de (Schafer, Musiani et Borelli 2016).

archives. Au cours des années 1990, avec la naissance d'Internet Archive, l'archivage du Web suivait un « modèle documentaire » dont l'objectif était un archivage universel, inspiré par les modèles traditionnels et tout particulièrement celui de la bibliothèque. Ensuite, au début des années 2000, ce modèle fut brièvement remplacé par une logique plus mémorielle, accompagnée de méthodes de « bricolage » reflétant le manque d'une meilleure alternative. Une deuxième phase mit l'accent sur les aspects de préservation systématique, une sorte de « congélation » à un instant t qui consistait à sauvegarder chaque élément du corpus, pièce par pièce. Enfin, depuis la fin des années 2000, les archives du Web sont construites selon une logique d'« archive temporelle », qui cherche à capturer entièrement l'instabilité du Web – en développant des méthodes d'archivage dynamiques, tout comme le Web est dynamique. L'instabilité, qui avait été considérée comme un dysfonctionnement contingent à l'objet, est de plus en plus considérée comme une caractéristique essentielle :

5. > Paradoxalement, l'instabilité qui caractérise les flux d'information ne constitue donc pas un obstacle à leur mémorisation, mais plutôt une condition, entraînant de nouvelles procédures de sédimentation mémorielle. Parce qu'ils sont instables, les contenus doivent être dédoublés par une information sur l'information, qui anticipe, optimise et instruit leur mobilisation. Les métadonnées désormais associées à tout message ne décrivent pas seulement les énoncés : elles en permettent la segmentation, la distribution et la recomposition, chaque fragment du flux devenant une mémoire activable à volonté, pointant vers d'autres fragments. (Merzeau 2012)

6. Avec cette attention particulière prêtée aux variations du Web « vivant », le Web archivé s'éloigne progressivement de l'idée d'une restitution. Il nécessite donc une compréhension de plus en plus fine des coulisses du stockage et de la circulation des flux d'information (Merzeau 2014).

7. Le chercheur Niels Ole Finneman (2015), plaçant au cœur de ses travaux ces questions de temporalité et d'intelligibilité, remarque que tous les corpus d'archives web répondent à trois dimensions temporelles : le contenu original, son accumulation et ses transformations, et enfin l'exploration de l'archive par le spécialiste. Celui-ci devient partie intégrante de l'intelligibilité des contenus car il est inscrit dans sa propre époque et peut introduire des biais, contribuant ainsi à une lecture nostalgique ou présentiste (Schafer 2015).

8. Comme le souligne Niels Brügger (2012), un autre aspect très important réside dans le fait qu'on n'est presque jamais en train de, tout simplement, « faire une copie ». Le processus d'archivage du Web crée une série de versions uniques d'un contenu, où quelque chose peut être perdu et autre chose, qui n'était pas en ligne à cet instant t, peut être archivé avec ce contenu. Ce qui peut rendre très difficile de savoir avec certitude à quoi ressemblait effectivement une partie du Web en ligne à un moment spécifique : chaque archive web est une reconstruction (Ankerson 2015).

9. Plusieurs raisons concourent à expliquer ce phénomène. La première est la profondeur de la collecte et de la capture. Très souvent, les sites web ne sont archivés que partiellement, car le robot *crawler** est programmé pour les capturer seulement à profondeur de quelques clics – ce qui explique pourquoi les utilisateurs se trouvent régulièrement face à des pages web manquantes ou non trouvées. Cela répond à l'effort pour capturer des échantillons vastes et représentatifs du Web contemporain dans sa diversité, malgré la « superficialité » que cela entraîne. Par exemple, en France, les collectes larges de la BNF privilégient la quantité ; or, si les 4 millions et demi de sites web collectés dans une année avec ce système sont très rarement préservés dans leur intégralité, c'est aussi le cas de leurs pages web, qui sont souvent incomplètes ; des éléments tels que les publicités, les *pop-up* et les bannières sont souvent bloqués avant la collecte. Cela entraîne l'omission d'une partie intéressante et importante du patrimoine nativement numérique, avec laquelle les utilisateurs du Web ont fréquemment eu un rapport problématique, voire conflictuel, mais qui reste une illustration importante des modèles d'affaires et des stratégies de communication des firmes numériques, basés sur l'économie de l'attention (Kessous 2012).

10. Les polices et caractères peuvent aussi différer dans les archives du Web par rapport aux pages originelles ; si au moment de l'archivage une police d'une page web n'était pas inscrite explicitement dans son code source originel, mais plutôt utilisée par défaut, ce sont les paramètres établis par défaut par le navigateur dans sa version actuelle qui figurent sur la page archivée.

11. Enfin, la collecte et la sauvegarde des images peuvent poser problème dans ce paysage mouvant : plusieurs pages web des années 1990, désormais archivées, montrent des trous béants là où leurs images étaient autrefois. Probablement, la raison à la base de ce phénomène est à rechercher moins dans la difficulté technique de la capture, et plus dans « l'impatience » des robots et dans les objectifs de la collecte à l'époque : Internet Archive était liée à l'entreprise Alexa de Brewster Kahle – une firme qui avait pour objectif de classifier et d'indexer les sites web plutôt que de préserver les images. Aujourd'hui, et afin d'éviter les doublons, celles-ci ne sont pas systématiquement re-collectées ; ainsi, si leur URL n'a pas changé d'un *crawl* à l'autre, elles peuvent être récupérées du *crawl* le plus récent, au lieu d'être à nouveau capturées. Cela explique également certaines inconsistances qui peuvent surgir lorsqu'on navigue dans le Web archivé – par exemple, quand un *widget* « calendrier » montre une date différente par rapport à la date de collecte de la page web.

Le périmètre de l'archive du Web

12. C'est le regard que l'on porte sur l'archive qui, dans une certaine mesure, définit son périmètre. C'est le cas pour le regard des chercheurs, l'un des premiers publics d'usagers de l'archive du Web. L'analyse de site web a donné

lieu à des réflexions méthodologiques et épistémologiques[5], mais qui tendent à effleurer la question de l'archive du Web sans, jusqu'à récemment, la prendre en charge frontalement. Niels Brügger a lancé une nouvelle dynamique en 2009, en dessinant les contours d'un usage de l'archive web par les chercheurs (Brügger 2009, 2011) à partir d'éléments distincts : l'objet web (par exemple une image insérée dans une page web), la page web, le site web, la sphère web (un ensemble de pages web liées par une thématique), le Web dans son ensemble (ses normes, ses standards, ses institutions, ses technologies…). Ainsi, la multitude des niveaux, formats et éléments documentaires concernés par l'archivage (textes, images, sons, vidéo, graphismes, bases de données, logiciels, codes…) entre dans un périmètre plus ou moins cohérent selon la manière dont on l'analyse.

13. Toutefois, le regard du chercheur est *in fine* cadré, bien que non limité, par les dispositifs mis en place par les professionnels de l'archivage numérique en général et du Web en particulier. Jinfang Niu a proposé dès 2012 une vue d'ensemble des enjeux de l'archivage du Web, définit comme le « processus de récolte et de stockage de données enregistrées sur le World Wide Web, de leur conservation sous la forme d'une archive, et de leur mise en accessibilité pour des recherches futures » (Niu 2012).

14. Pour Niu, ce périmètre peut être décrit par les processus de travail de cet archivage, passant par :

1. L'évaluation et la sélection, qui même dans le cas de collections non discriminantes des contenus se font sur la base de postulats et de critères au moins sous-jacents. Par exemple, pour Internet Archive qui a priori ne trie pas sa récolte, c'est essentiellement le « Web de surface » (indexé par les moteurs de recherche) qui est concerné. Les collections institutionnelles sont plus sélectives, sur la base de critères géographiques, thématiques, évènementiels (comme dans le cas des périodes électorales, ou des crises terroristes), ou encore génériques (selon le type ou le format de média). Cette sélection est plus ou moins automatisée ou manuelle, plus ou moins programmée à l'avance ou ouverte à l'intervention (formulaires d'enregistrement, recommandation…). L'évaluation de la valeur peut reposer sur des méthodes très différentes : alors que la NARA (National Archives and Records Administration) américaine évalue la valeur d'un site individuel, la BNF préfère la représentativité (toutes les pages web françaises sans distinction de qualité), et le service des archives web de l'université nationale de Taiwan a recours à l'échantillonnage
2. L'acquisition : si la tradition institutionnelle de dons et dépôts est toujours d'actualité, l'archivage du Web a donné lieu à des méthodes originales, comme l'indexation de réseau (*crawling*) qui récolte les contenus par le biais du suivi d'hyperliens. La question des permissions se pose à cette étape, sauf en cas de mandat gouvernemental (en particulier le dépôt légal*, comme en France, en Nouvelle-Zélande, aux États-Unis ou encore au

5. Voir par exemple (Barats 2013).

Royaume-Uni), ou de mise en place de clauses de retrait (solutions *opt-out*, comme chez Internet Archive)

3. L'organisation et le stockage : ceux-ci doivent préserver l'intégrité du contenu, en donnant des informations sur l'origine (de la source de l'enregistrement à son adresse en tant que document vivant) et l'ordonnancement (l'agencement au sein de la structure des archives)
4. La description : les métadonnées décrivant les archives sont générées automatiquement lors de l'indexation (par exemple la signature temporelle de la récolte, la taille, le format…), ou bien induites à partir d'une extraction des métadonnées* du code des pages d'origine
5. L'accès et l'utilisation : ils sont déterminés par le contexte légal de l'archive du Web, avec une tendance à la restriction sur le modèle des « *dark archives* », qu'on ne peut consulter qu'*in situ*, « à l'ombre » des bibliothèques, par opposition aux archives ouvertes[6] (Smit, Van Der Hoeven et Giaretta 2011). Les potentialités de la recherche reposent sur la richesse des métadonnées de description, des outils d'indexation et des choix d'interface

15. Pour les professionnels, le cahier des charges d'un projet d'archivage du Web résume ces problématiques en cinq recommandations formulées par l'IIPC Preservation Working Group : la mise en place d'objectifs à buts juridiques ou scientifiques ; l'évaluation des possibilités et contraintes légales ; l'approche raisonnée de la création de collections selon des critères ; l'identification des problèmes de mise en collection (techniques et organisationnels) ; la stratégie de conservation à long terme (métadonnées, formats…).

16. La question de la création des collections révèle nombre d'autres enjeux sous-jacents. Par exemple, comment rendre la cohérence d'une collection à partir de nouveaux genres éditoriaux nativement numériques, encore mal identifiés ? Les collections de blogs ont ainsi retenu l'attention, pour les problèmes qu'ils posent en matière de droit d'auteur et de la personne, de responsabilité d'hébergement, de filtrage et d'éditorialisation des informations, de frontières floues entre production professionnelle et amateure de contenus de médias en ligne, de limites labiles entre contenu d'auteur et commentaires du public, etc. Des projets spécifiques ont été mis en place pour les prendre en charge, comme *Blog Forever*, projet collaboratif collectant, conservant, administrant et réutilisant des archives de blogs, financé par la Commission européenne[7].

17. On remarque que de nombreuses contraintes limitent le périmètre des archives du Web telles qu'elles s'offrent à l'usager. Internet Archive, qui prône une politique de numérisation massive, revendique une responsabilité civique dans l'accessibilité publique aux contenus, quitte à contourner ce que la fondation considère comme des barrières fixées par l'économie et le droit de

6. Le lecteur trouvera dans le glossaire une présentation générale de l'Open access.

7. Pour en savoir plus, consulter : https://cordis.europa.eu/project/rcn/98063_fr.html.

l'édition et des archives. Le périmètre de ses archives en est d'autant plus élargi, avec une ambition non départie d'idéaux universalistes (Paloque-Bergès 2014). C'est aussi l'approche de beaucoup d'organisations non institutionnelles, fondations privées, jeunes entreprises ou initiatives individuelles, qui étendent le périmètre de l'archive du Web aux activités culturelles sur Internet, dans une logique d'auto-archivage des productions individuelles. Par exemple, le Google Cultural Institute crée des outils accompagnant les utilisateurs dans la création de galeries de vie numérique sur leurs sites web personnels. Récusant le vocabulaire des professionnels du patrimoine, comme « commissaire d'exposition numérique », il encourage le « mariage du professionnel et de l'amateur[8] » dans le domaine de la conservation numérique. Ces approches exogènes aux institutions du patrimoine invitent à interroger la manière dont le numérique altère la perception de ce qu'est un document, une archive, ou encore une collection, au sens technique, mais aussi culturel et social. Sarah Atkinson et Sarah Whatley (2015) rappellent ainsi que les archives numériques doivent être mises en perspective avec l'espace public numérique. Ainsi, l'utilisateur et le public jouent un rôle dans la construction du périmètre de l'archive, favorisant les pratiques de l'archivage collaboratif et ouvert.

L'archivage des réseaux socionumériques, quelles spécificités ?

18. Si l'archivage du Web a bénéficié de l'initiative précoce de Brewster Kahle, le paysage numérique et ses usages ont profondément changé depuis 1996, notamment avec l'arrivée des réseaux socionumériques (RSN). Dispositifs de flux, dont Frédéric Clavert (2017) note à propos de Twitter que « collecter des tweets, notamment, via une API*, c'est transformer un flux constant en archive figée. La notion de source, flux originel intarissable, n'a jamais été une métaphore aussi actuelle », les RSN proposent par ailleurs des modalités de participation et d'accès, qui peuvent rendre l'archivage complexe : identifiants et mots de passe, statuts privés ou semi-publics des contenus, usages de protocoles spécifiques, notamment concernant les vidéos, encapsulage de liens contenant des URL parfois réduites, etc. Les contenus des RSN ne sont donc pas toujours aisément accessibles ou archivables, sans compter les changements de protocoles ou de politiques utilisateurs qu'ils introduisent fréquemment. Comme le rappelait Annick Le Follic, alors chargée de collections numériques au département de dépôt légal de la BNF, dans un entretien le 21 mars 2016 : « La limite de notre archivage des réseaux sociaux est technique : ces plateformes changent souvent de technologies et de paramètres, donc il nous faut donner à chaque fois une instruction manuelle à Heritrix[9] pour qu'il capture bien les contenus qui nous intéressent. En particulier, les

8. Kuchler, Hannah. 2014. « How to preserve the web's past for the future ». *Financial Times*, 11 avril 2014. https://www.ft.com/content/d87a33d8-c0a0-11e3-8578-00144feabdc0.

9. Robot d'indexation utilisé par la BNF mais aussi par Internet Archive. https://webarchive.jira.com/wiki/spaces/Heritrix.

protocoles HTTPS[10] nous posent parfois des problèmes, tout comme Facebook lorsqu'il utilisait des *captcha*[11] ». Les RSN n'en demeurent pas moins des témoins et supports de nos vies numériques, qui ne pouvaient rester en dehors de la réflexion sur l'archivage du Web.

19. La bibliothèque du Congrès (LOC) aux États-Unis a ainsi passé un accord en 2010 avec l'entreprise Twitter pour récupérer tous les tweets émis depuis 2006 et poursuivre ensuite cette conservation. Reste qu'à ce jour cette collection n'est pas encore accessible pour les chercheurs et soulève diverses questions, amenant même la LOC à revenir sur son projet d'exhaustivité pour se concentrer sur un périmètre plus restreint et sélectif de collecte[12]. En effet, les outils disponibles pour faire des recherches dans ces fonds gigantesques sont un enjeu majeur (le nombre de tweets journalier est passé selon la LOC de 140 millions début février 2010 à 500 millions par jour en octobre 2012). Dans un document de janvier 2013, intitulé « Update on the Twitter Archive at the Library of Congress[13] », la bibliothèque notait ainsi que réaliser une recherche sur la période 2006-2010 pouvait prendre 24 heures, et elle faisait le constat que les technologies disponibles pour accéder à ces données n'étaient pas encore aussi avancées que celles permettant de les collecter. Bien sûr l'accord entre la bibliothèque étasunienne et l'entreprise pose également la question des modalités concrètes d'accès à ces archives : leur accessibilité pour des chercheurs par exemple européens impliquera-t-elle de venir à la LOC ?

20. Des initiatives européennes ont aussi été engagées, mais avec des périmètres plus restreints, appuyés par exemple en France sur le cadre du dépôt légal du Web. La collecte de Twitter par la BNF et l'INA apporte des éléments complémentaires à une réflexion sur le patrimoine des RSN. Tout d'abord, si la BNF et l'INA archivent une partie de Twitter, elles n'ignorent pas les autres RSN, mais peuvent rencontrer plus de difficultés pour les collecter. Les deux institutions ont davantage archivé Twitter que Facebook par exemple, car les contenus de Facebook ne sont pas tous publics, outre les difficultés techniques précédemment évoquées. Et pourtant les Français sont davantage présents sur Facebook et la diversité sociologique y est mieux représentée[14]. De plus, comme pour le Web, le périmètre de collecte est aussi sélectif pour les RSN et si l'INA a pris la mesure de l'intérêt de l'archivage de Twitter et lancé des collectes dès 2014, l'équipe dédiée au DL Web (pour Dépôt légal Web) le fait dans le cadre de son périmètre lié à l'audiovisuel : elle suit ainsi les comptes d'acteurs clés du monde audiovisuel français, soit environ 13 000 utilisateurs et 400 hashtags.

10. Protocole web sécurisé.

11. Entretien mené par Marguerite Borelli et Valérie Schafer dans le cadre du projet *ASAP*, le 21 mars 2016. https://asap.hypotheses.org/168.

12. Voir l'article de Plaugic, Lizzie. 2017. « The Library of Congress will no longer archive every tweet ». *The Verge*, 26 décembre 2017. https://www.theverge.com/2017/12/26/16819748/library-of-congress-twitter-archive-project-stalled.

13. D'après « Update on the Twitter Archive at the Library of Congress ». 2017. Library of Congress. https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf.

14. Pour un aperçu des chiffres, voir Coëffé, Thomas. 2017. « Les 50 chiffres à connaître sur les médias sociaux en 2018 ». *BDM* (blog). 28 décembre 2017. https://www.blogdumoderateur.com/50-chiffres-medias-sociaux-2018/.

21. Mais son expérience s'est aussi manifestée lors des attentats de 2015, au moment où des millions de tweets ont réagi aux évènements autour de *Charlie Hebdo* puis à ceux de novembre 2015, suscitant aussi la réactivité de chercheurs qui lancent des collectes très rapidement (par exemple la collecte de Romain Badouard qui sert de base à sa réflexion sur le « Je ne suis pas Charlie » (Badouard 2016), celle du canadien Nick Ruest, dont les données sont accessibles en ligne[15], ou encore celles de Giglietto et Lee (2015). Comme le note Zeynep Pehlivan (DL Web INA) qui revient sur cet archivage réalisé en urgence :

22. > Nous avons poursuivi les collectes sur les attentats après 2015, par exemple Nice à l'été 2016. Nous avons aussi des archives relevant d'attentats qui ont eu lieu en Europe, à Bruxelles, Londres ou Manchester. En effet s'ils ne se sont pas passés en France, ils ont été profondément relayés par les médias français et sont entrés rapidement dans les *trends* [principales tendances de mots-clés] de Twitter, car les Français ont réagi. Ces tweets font partie intégrante du contexte médiatique et permettent en outre au chercheur de mettre en perspective les tweets de notre cœur de corpus du dépôt légal. Par contre on ne fait pas des collectes pour tous les attentats dans le monde, seulement pour ceux qui ont un écho fort en France, en particulier dans le monde de l'audiovisuel, qui est notre périmètre dans le cadre du dépôt légal du Web[16].

23. L'INA a pleinement conscience de l'intérêt de démarrer la collecte tôt, de ne pas rater le pic de tweets ou la montée d'un « mot-dièse » (des mots-clés précédés d'un signe « # », appelé « hashtag », permettant d'étiqueter les tweets). « Or le service est fermé la nuit ou le week-end. Aussi nous avons décidé d'archiver dorénavant automatiquement les principaux *trends* en France. Nous avons ainsi une veille automatique complémentaire, même en dehors des heures d'ouverture, sur des mots-dièses qui montent et sont en général portés ou repris dans les médias. Aujourd'hui les journalistes aussi participent et suivent en effet Twitter et ces mouvements », ajoute Zeynep Pehlivan[17].

24. Si l'aspect des archives du Web peut changer d'une institution à une autre, le cas de Twitter est particulièrement révélateur : la BNF utilise le robot de capture Heritrix développé par Internet Archive et obtient des résultats proches d'une capture d'écran, tandis que l'INA passe par l'API publique de Twitter et ne capte pas les images de fond. Les deux interfaces de programmation, API Search et Streaming, permettant pour la première à un utilisateur de remonter à un contenu particulier sur les sept derniers jours, et pour la seconde de capter un flux au fur et à mesure pour une requête précise, sont gratuites et publiques. Il est aussi possible de récupérer a posteriori les données de Twitter de façon payante. L'API publique a des limites : on ne peut collecter plus de 1 % du total des tweets émis au plan mondial à un instant t. Cette

15. Voir : https://dataverse.scholarsportal.info/dataset.xhtml?persistentId=hdl:10864/10830.

16. Entretien réalisé par Valérie Schafer fin 2017.

17. *Idem*.

limite a notamment été dépassée au moment du pic de flux lié aux attentats parisiens, et même les 20 millions de tweets conservés par l'INA sur les évènements du Bataclan ne constituent donc pas une collecte exhaustive de ce qui s'est dit sur Twitter autour du 13 novembre 2015. Ajoutons que la collecte dépend des mots-dièses sélectionnés et que certains peuvent échapper à l'archivage, qui se joue en urgence. D'autres biais ou limites ne peuvent être ignorés du chercheur : par exemple le nombre de retweets (re-publication de tweet par un autre usager) d'un message s'arrête à la date de l'archivage du tweet, impliquant donc de sérieuses précautions sur l'interprétation de cette donnée.

25. Reste qu'au-delà de ces limites, le volume archivé au moment des attentats parisiens est tel qu'il peut être considéré comme représentatif, à défaut d'être exhaustif, d'autant que l'INA s'applique à documenter sa collecte en intégrant notamment des informations sur les données manquantes, en archivant les messages signalant une restriction dans la collecte, etc. Évidemment, il faut souligner une autre limite à la représentativité : les publics de ces plateformes sont spécifiques « comme le sont les lecteurs de journaux ou les tenants de la conversation de bistrot. Mais ces traces peuvent sous certaines conditions donner accès à certains processus qu'on ne pouvait chiffrer jusqu'ici » (Boullier 2015).

Les barrières, limites, verrous à l'archivage

26. Déjà évoquées, la disparition des pages web, la volatilité des contenus et l'évolution générale des réseaux sont les limites fondamentales rencontrées par l'archivage du Web. En 2013, la durée de vie moyenne d'une URL est de 9,3 ans ; celles qui ne survivent pas entretiennent le « *link rot*[18] » (la décomposition des liens). Un « lien mort » est d'autant plus dommageable qu'il a pu servir de référence, voire de garantie institutionnelle, comme en a témoigné la même année l'affaire des articles disparus de la Cour suprême américaine révélée par *The New York Times*[19] – on parle alors de « *reference rot* ». Les liens et contenus web s'effacent au gré de la fermeture d'hébergeurs ou de plateformes, de la réorganisation de l'architecture d'un site, ou parce qu'un auteur a tout simplement choisi de supprimer un contenu, voire d'effacer complètement sa présence numérique, ce que l'on surnomme « infosuicide ».

27. Le Web peut également, tout en restant bien vivant, résister à l'archivage. Pour des raisons techniques, tout d'abord, dans la mesure où il peut être difficile pour les dispositifs d'archivage automatique de capturer contenus et objets mis en forme par des technologies non prises en charge par le dispositif ou obsolètes. Suivant

18. Summers, Ed. 2015. « The Web as a Preservation Medium ». *On Archivy* (blog). 7 mai 2015. https://medium.com/on-archivy/the-web-as-a-preservation-medium-3d697328b3b8.

19. Liptak, Adam. 2013. « In Supreme Court Opinions, Web Links to Nowhere ». *The New York Times*, 23 septembre 2013. https://www.nytimes.com/2013/09/24/us/politics/in-supreme-court-opinions-clicks-that-lead-nowhere.html.

une logique de flux, le Web dynamique tend à encapsuler des contenus hébergés ailleurs, une page n'étant que de plus en plus rarement une unité homogène. Ainsi, ces dispositifs peuvent avoir tendance à reconstituer des pages « à trous ». Par exemple, le langage JavaScript permettant une telle encapsulation de contenu a été l'un des premiers obstacles au moissonnage de données web par l'outil Heritrix, produisant des archives de pages web qui sont des coquilles vides. L'enchâssement de plusieurs types de logiciels de gestion de contenu et la superposition de plusieurs couches de code peuvent également compliquer la tâche d'une collecte numérique. C'est le cas de la re-publication ou de l'administration de forums Internet dont la vie peut être déterminée par des protocoles différents de ceux du Web : mal gérés par leurs administrateurs, difficiles à naviguer, impossibles à collecter, ils tendent à devenir des « ruines numériques » sur le Web (Paloque-Bergès 2017, 2018). Des barrières plus proactives peuvent être mises en place par les hébergeurs, les administrateurs, et les auteurs. Le problème du verrouillage par mot de passe est un classique, que l'on retrouve de manière généralisée sur les plateformes de réseaux sociaux. Le recours à un code contractuel est également une technique ancienne, comme dans le cas du protocole du *robot.txt*, une formule insérée dans le code source d'une page web par son créateur. Cette technique « a pour but principal de permettre à un éditeur d'exclure certains de ses documents du champ d'action des agents logiciels appelés *crawlers* utilisés par les moteurs de recherche pour prendre connaissance des documents » (Sire 2015). Toutefois, comme l'analyse Guillaume Sire, ce contrat de code repose sur un consensus léonin, c'est-à-dire régit par des rapports de force déséquilibrés. Google peut choisir de passer outre ce protocole tout comme certaines institutions d'archivage du Web, ces dernières en vertu des modalités du dépôt légal (Niu 2012).

28. « *Link rot* », « *reference rot* », « *infosuicide* », « *digital ruins* » : autant d'images d'un Web en décomposition, dont la logique entre pourtant dans ce que l'archéologie des médias appelle les « médias zombies », où l'information ne meurt jamais tout à fait car elle survit sous une forme ou une autre (Chun 2011). De fait, ce dépérissement stimule la résilience. Ainsi, Tim Berners-Lee lui-même a été l'un des promoteurs les plus actifs de techniques de liens pérennes au sein du monde des développeurs web, derrière le slogan « *Cool URIs* don't change* ». Des méthodes alternatives émergent pour pallier les difficultés des archivistes numériques. Les *digital forensics*, ainsi, s'intéressent à la reconstitution de documents critiques à travers les données de navigation, les courriers électroniques, l'historique des recherches, etc. (Kirschenbaum *et al.* 2010). La diplomatique numérique, elle, propose de contextualiser la valeur du document (Chabin 2012)[20]. Ces méthodes viennent tenter de répondre aux interrogations traditionnelles que les historiens renouvellent face aux archives numériques : comment dater, authentifier un document, combler les lacunes, retrouver le contexte, équilibrer les caractères externes (matériels) et

20. Voir le chapitre de Marie-Anne Chabin dans cet ouvrage : « La méthode diplomatique-face à l'information numérique ».

internes (cohérence des textes) des sources, ou encore évaluer le rapport entre échantillon et tout, singularité et représentativité.

Des enjeux de gouvernance

29. En 1980, le philosophe et sociologue Langdon Winner se demandait dans un article qui a fait école : « Est-ce que les artefacts sont politiques ? » (*Do artifacts have politic?*). Avec ce mot « politique », Winner mettait à l'épreuve la question de la neutralité technologique, pour rechercher dans les objets techniques les « arrangements de pouvoir et d'autorité dans les associations humaines, ainsi que les activités qui se passent à l'intérieur de ces arrangements » (Winner 1980). Si on cherche à appliquer cette hypothèse à l'étude des archives du Web, il s'agit de comprendre les manières dont la nature distribuée, et inscrite dans la technique, de l'archivage du Web lui permet d'incarner des formes spécifiques d'autorité et de pouvoir (DeNardis 2014), ce qui ferait de ce domaine un microcosme de la gouvernance de l'Internet au sens plus large. Cette démarche a occupé certains de nos travaux récents (Schafer, Musiani et Borelli 2016 ; Musiani et Schafer 2019)[21].

30. L'archivage du Web repose sur un modèle multi-parties prenantes. Une variété d'acteurs sont concernés par l'archivage du Web : des fondations comme Internet Archive ; des organisations transnationales, à commencer par l'IIPC ; la société civile (des militants de l'Archive Team à d'autres initiatives fondées par des communautés de chercheurs) ; et enfin, le secteur privé (par exemple, Google, qui s'est impliqué dans la conservation du patrimoine numérique natif en rendant disponible un certain nombre de groupes du forum numérique Usenet[22]). Ainsi, on retrouve dans l'archivage du Web les principales catégories d'acteurs impliqués dans la gouvernance d'Internet, ainsi que leurs tensions et leurs alliances. Des expériences de collaboration entre des institutions d'archivage et des équipes de recherche voient ainsi régulièrement le jour ; la BNF a par exemple associé notre équipe Web90[23] à une réflexion sur l'implémentation du plein texte dans les archives web des années 1990, et à un niveau plus global, le réseau RESAW[24] associe des chercheurs et des professionnels de l'archivage. L'Internet Archive va encore plus loin en promouvant explicitement des initiatives *bottom-up* destinées à revaloriser l'intervention humaine dans un monde où « les machines allaient nous sauver – parcourant le Web, numérisant les ouvrages, organisant l'information [mais ce sont] les communautés de gens qui sont au centre de l'archivage[25] » (Kahle 2014).

21. Sur lesquels cette section se base.

22. Voir notamment (Paloque-Bergès 2017).

23. Cf. https://web90.hypotheses.org/

24. Cf. http://resaw.eu/

25. La citation originale est la suivante : « *We thought the machines were going to save us – crawling the web, digitizing the books, organizing the information – but we were wrong [...] Communities of people are at the heart of curation* ».

31. L'archivage du Web n'échappe pas à des tensions ayant trait à la standardisation, un des enjeux traditionnellement le plus vif de la gouvernance d'Internet, et à des imaginaires et visions divergents, des communs aux formats propriétaires. À ce propos, il est souvent intéressant d'observer les types et les périmètres des organisations d'archivage du Web. Depuis août 2006, la mission de la BNF est de collecter et préserver une sélection de sites internet dans le cadre du dépôt légal. Cette mission doit être menée dans le respect de la propriété intellectuelle et de la protection des données personnelles, ce qui rend les collections non accessibles en ligne, et fait du Web une composante parmi d'autres d'un patrimoine éditorial français dont la mémoire doit être préservée. Ce panorama offre un contraste net avec la mission que s'est assignée l'Archive Team, précédemment évoquée, qui n'est contrainte « que » par la disponibilité des ressources informatiques et le souhait, de la part des utilisateurs, de les partager. Dans le premier cas, on voit le poids d'un héritage historique et de questions de souveraineté liées au dépôt légal et dans le second, le lien entre la capacité technique de l'individu et sa possibilité de contribuer à l'entreprise d'archivage.

32. L'archivage du Web révèle également la présence de tensions géopolitiques, illustrées de façon emblématique par les appels de Brewster Kahle lors du blocage d'Internet Archive par la Chine (Kahle 2014) ou lors des élections présidentielles américaines de novembre 2016, lorsqu'il appelle, à la suite de la victoire de Donald Trump, à un financement participatif pour créer par précaution une copie complète des collections numériques de l'Internet Archive (Kahle 2016).

33. On retrouve aussi dans l'archivage du Web certaines dynamiques qui rappellent le problème de la fracture numérique : cette communauté inclut presque exclusivement des institutions du « Nord global » (Gomes, Miranda et Costa 2011) – la présence des pays en voie de développement dans le Web archivé n'étant aucunement proportionnelle à leur présence croissante au sein du Web vivant. Un certain nombre d'associations régionales pourrait épauler l'action globale de l'IIPC et faire office de « sous-forums » pour l'échange entre acteurs autour de problèmes spécifiques à certaines régions et pour coordonner le transfert de compétences pratiques – des initiatives se développent notamment dans le Sud-Ouest de l'Asie. Cependant, il existe encore des régions du monde qui restent largement « non-archivées », en particulier en Inde, en Amérique latine et en Afrique. Comme l'expose la conférence « *The Memory of the World in the Digital Age* » (Duranti et Shaffer 2012), parmi les problèmes élémentaires de l'archivage numérique se trouvent la simple absence de ressources techniques, légale et financière, comme dans le cas de la sauvegarde des archives juridiques du Burundi. Pour pallier le risque de perdre des ressources culturelles, politiques et sociales importantes, certaines institutions « du Nord » ont entrepris d'en préserver certaines (par exemple, l'université d'Heidelberg effectue une collecte du Web socio-politique chinois) ; mais à long terme, une réponse

durable devra sans doute résider dans le développement d'initiatives locales.

34. Enfin, on retrouve dans l'archivage du Web la dialectique entre différentes pratiques et sources de normativité, de la technologie au marché, de la concertation transnationale et internationale aux standards et aux droits – une pluralité d'instruments de gouvernance qui avait déjà été identifiée pour la gouvernance d'Internet (Bygrave et Bing 2009 ; Badouard *et al.* 2013). Le « sauvetage » de Geocities opéré par l'Internet Archive suite à la fermeture de la plateforme d'hébergement de pages personnelles par Yahoo!, les collectes d'archives et de données privées par Twitter et Facebook, le dépôt légal dans plusieurs pays, la charte de l'UNESCO, l'action « standardisante » de l'IIPC : ces différents instruments de gouvernance co-existent et se superposent partiellement. L'archivage du Web réactive donc les mêmes polarisations, négociations et dynamiques qui avaient émergé lors de la naissance de la gouvernance d'Internet, notamment avec le Sommet mondial sur la société de l'information en 2003 et 2005 (« Report of the Working Group on Internet Governance » 2005).

Former « au numérique » en sciences humaines et sociales ? Propositions d'un historien

Émilien Ruiz

Introduction

1. Après une dizaine d'années de développement des « humanités numériques » en France, nos disciplines sont à la croisée des chemins. En dépit d'injonctions croissantes un double questionnement perdure : pourquoi et comment former « au numérique » des étudiants en sciences humaines et sociales[1] ?

2. Il existe, bien sûr, de plus en plus de formations professionnalisantes, en particulier les masters « Humanités numériques et computationnelles » de l'École nationale des chartes et « Métiers informatiques et maîtrise d'ouvrage » de la Sorbonne, qui s'adressent à des étudiants avec une solide formation initiale en sciences humaines. Mais qu'en est-il des formations initiales généralistes ?

3. Le monde qui nous entoure connaît, depuis plusieurs décennies, une transformation majeure. Parfois qualifiée de « grande conversion » (Doueihi 2011), elle est alors associée à l'irruption d'une nouvelle culture liée à l'apparition du Web dans nos vies quotidiennes et professionnelles. Elle peut aussi être considérée comme une nouvelle « révolution industrielle » (Caron 1997), qui s'inscrit dans une transformation de plus longue durée liée à l'informatisation. Dans tous les cas, elle affecte la société dans son ensemble. On peut la considérer, en suivant ici Dominique Cardon, comme une « rupture dans la manière dont nos sociétés produisent, partagent et utilisent les connaissances », avec des implications qui ne sont pas seulement techniques et intellectuelles, mais aussi politiques et économiques (Cardon 2019). Or, dans le même temps, dans les disciplines de sciences humaines et sociales, les réticences perdurent chez certains étudiants comme parmi leurs enseignants. Nos disciplines se perçoivent parfois elles-mêmes comme quasiment étrangères aux conséquences des mutations à l'œuvre ; mutations que l'on ne saurait résumer au développement des humanités numériques. Dès lors, la tentation d'un repli sur soi peut être grande : tant chez les partisans jusqu'auboutistes des humanités numériques que chez les tenants d'une sorte de pureté disciplinaire largement fantasmée.

4. La teneur de plusieurs échanges publics survenus au cours des derniers mois, des listes professionnelles aux

1. Si le propos qui suit n'engage que son auteur, il doit beaucoup aux conversations entretenues sur ces questions avec Paul Bertrand, Franziska Heimburger (dix ans de *Boîte à outils des historien.ne.s* !) et Claire Lemercier, que je remercie vivement pour ses commentaires sur une version préliminaire de ce chapitre.

réseaux sociaux numériques, en témoigne. Ceux qui, à partir d'une question innocente d'un étudiant en master, ont animé la liste « DH » au début de l'été 2019 ont ainsi condensé la plupart des crispations qui agitent le champ : de la question des rapports entre informaticiens et spécialistes de SHS à celle de la place des ingénieurs dans les équipes en passant par l'obstacle à la qualification CNU et au recrutement que constituerait l'adoption d'une démarche résolument interdisciplinaire[2]. Un peu plus d'un an auparavant, en mars 2018, c'est à la suite de la publication d'un billet sur la version francophone de *The Conversation* que les débats animèrent les réseaux sociaux numériques. Marcello Vitali-Rosati, faisant mine de s'interroger, « les chercheurs en SHS savent-ils écrire ? », défendait l'idée qu'écrire avec un traitement de texte plutôt qu'en XML* ou HTML relevait de l'incompétence (Vitali-Rosati 2018). Les témoignages d'une irritation légitime, notamment parmi des collègues qui ne nourrissaient pourtant pas d'hostilité particulière envers les « humanités numériques », ne se firent pas attendre : « avant on avait les gars de sciences exactes venant nous expliquer que ce qu'on fait sert à rien, maintenant EN PLUS on a les professionnels des humanités numériques qui viennent nous dire qu'on ne sait pas faire notre métier[3] ». C'est dans un tel climat que tombent les injonctions, de plus en plus nombreuses, à former nos étudiants « au numérique ».

5. Ces prescriptions sont, sans conteste, des contraintes qui viennent complexifier le cadrage des nouvelles maquettes de formation. En outre, la tâche est compliquée par une certaine confusion, parfois savamment entretenue, entre « formation aux humanités numériques » et « formation au numérique destinée aux étudiants en SHS ». Pour autant, en faire l'un des bras armés d'une offensive idéologique néo-libérale sans précédent relève d'un amalgame un peu trop rapide (Carnino et Jarrige 2019).

6. Ces injonctions peuvent, au contraire, constituer une chance pour les SHS. Si à ce stade la question des nouveaux débouchés professionnels reste ouverte, voire incertaine, intégrer pleinement « le numérique » dans nos licences et nos masters permettra de mieux former nos étudiants : qu'ils se destinent à l'enseignement, à la recherche, ou à toute autre voie, leur environnement professionnel sera, au moins en partie, numérique.

7. Cela suppose toutefois que deux conditions soient remplies. La première consiste à reconnaître le fait que nos universités sont encore aujourd'hui, en dépit des attaques réelles qu'elles subissent, l'un des rares espaces professionnels en grande partie autogéré. L'injonction à « faire du numérique » peut ainsi tout à fait être appropriée pour, justement, ne pas rester sous « hypnose numérique » ; mais pour cela il faut que nous nous impliquions collectivement dans les questions pédagogiques. La seconde

2. Les archives de la liste « DH » étant publiques, ces échanges peuvent être consultés à l'adresse suivante : https://groupes.renater.fr/sympa/arc/dh/2019-07/msg00012.html.

3. Tweet de @kinkybambou, 13 mars 2018 : https://twitter.com/kinkybambou/status/973490170740789254 ; sur les discussions qui ont suivi, je me permets de renvoyer à deux billets (Ruiz 2018, 2019) dont certains éléments sont repris dans le présent chapitre.

condition, qui me semble indispensable, consiste à placer « le numérique » au cœur de nos pratiques pédagogiques, à l'intérieur même de nos disciplines, de ne plus en faire une sorte de spécialité réservée aux promoteurs et praticiens des humanités numériques.

S'émanciper des humanités numériques

8. Tel le spectre de Marx, les « DH » hantent la formation numérique des étudiants en SHS. Si les humanités numériques m'intéressent beaucoup, que je tente d'en comprendre les enjeux, d'en enseigner certains aspects et, très modestement, de participer à leur développement, je n'ai jamais revendiqué cette appellation pour définir mes propres pratiques. Pas par rejet ; pas uniquement parce que je pense que mes compétences informatiques sont souvent en deçà de celles de certains collègues ; principalement parce que j'éprouve bien trop de difficultés à en identifier une définition à la fois simple et stabilisée, ou, à tout le moins, consensuelle.

9. Dans le contexte français, cette question sémantique a déjà fait couler beaucoup d'encre : en dépit du geste fondateur du « manifeste des *digital humanities* » (Dacos 2011) la question semble loin d'être réglée ; les réactions à la publication d'une nouvelle définition au *Journal officiel*[4] au début de l'été 2019 en ont à nouveau témoigné (Salvador 2019a, 2019b). C'est, au demeurant, une caractéristique significative de ce champ puisque, de longue date, les réflexions relatives à sa constitution et son fonctionnement le définissent comme une « auberge espagnole » où chacun trouve et projette « ses propres désirs voire ses propres fantasmes scientifiques » (Le Deuff 2012, 2018) ; un lieu « de circulation et d'échanges » sans équivalent (Mounier 2015) ou encore une « zone d'échange » un « objet- » ou un « projet frontière » (Clavert et Schafer 2019).

10. Pour tenter de saisir ce que peut être l'apport des humanités numériques pour la formation des étudiants en SHS, c'est moins sur leur définition que sur les pratiques de ceux qui s'en réclament qu'il semble nécessaire de se pencher. Aurélien Berra nous appelait déjà à une telle focalisation il y a quelques années (Berra 2012). Une façon de la mettre en œuvre pourrait consister en un examen de publications collectives récentes. En 2017, *Expérimenter les humanités numériques* offrait, par exemple, un panorama très diversifié : de l'annotation des documents vidéos à l'usage des SIG, de la pratique des réseaux sociaux à l'édition électronique, de l'usage de Zotero à la mise en place d'un plan de gestion de données (Cavalié *et al.* 2017). Une autre impression se dégage de l'examen des sommaires des deux premiers numéros de la revue *Humanités numériques*. Les 18 articles annoncés témoignent de « ce que font » ceux qui considèrent que leurs travaux s'inscrivent dans le champ des humanités numériques. Si les appels à contribution concernaient les rapports des disciplines aux HN et les questions de formation, la majorité des

4. *Journal officiel*, 9 juillet 2019 : https://www.legifrance.gouv.fr/jo_pdf.do?id=JORFTEXT000038736904.

articles portent, sous la forme de retours d'expériences pratiques ou de réflexions théoriques, sur les pratiques d'éditorialisation (Berra 2019). Ainsi, l'observation des pratiques peut, en fonction des contextes et des points d'observation, donner l'impression d'un éventail très large comme très étroit.

11. Or, tout cela ne nous dit pas grand-chose de la formation des étudiants en SHS. Pour s'en faire une idée plus précise, les masters en humanités numériques, qui se multiplient ces derniers temps, semblent constituer un bon point d'observation. Selon le portail national des masters du MESRI (ministère de l'Enseignement supérieur, de la Recherche et de l'Innovation) trouvermonmaster.gouv.fr, 26 masters relèveraient des humanités numériques à la rentrée 2019. Ils correspondent aux résultats d'une requête simple interrogeant non seulement les intitulés des formations, mais aussi les mots-clés qui leur sont associés (tout comme les noms d'établissement, les villes, etc.). Un nettoyage et un tri de ces résultats a été opérée moyennant :

- la suppression des doublons (un master co-habilité par plusieurs universités pouvant apparaître plusieurs fois)
- la suppression des formations qui n'ont pas été ouvertes en 2019 (ou pour lesquelles on ne trouve aucune fiche formation sur le site de l'établissement)
- la subdivision des divers parcours (une même mention pouvant regrouper plusieurs parcours)

12. Cela conduit à lister 39 formations de master en humanités numériques ouvertes en 2019/2020. Les informations figurant sur le portail sont le fruit d'une transmission, par les établissements, d'éléments fournis par les responsables formations. Elles permettent ainsi, non seulement de cerner les contours des formations aux humanités numériques proposées aux étudiants ; mais aussi de se faire une idée de ce qui est revendiqué comme un master en humanités numériques par leurs porteurs.

13. Telles qu'elles se présentent sur le portail, ces 39 formations peuvent relever de plusieurs domaines disciplinaires à la fois. Si la majorité s'inscrit en « Sciences humaines et sociales » (32), les domaines « Arts, lettres et langues » (15) et des « Sciences, technologie et santé » (9) sont aussi représentés. De telles catégories sont toutefois un peu trop englobantes pour nous permettre de cerner les ancrages disciplinaires de ces formations.

14. Pour tenter de les approcher de façon plus précise, la figure 1 *infra* propose une répartition par discipline d'appartenance des responsables pédagogiques des formations. Pour réaliser ce tableau, il a ainsi fallu procéder à quelques recherches complémentaires dans les descriptifs publiés sur les sites des établissements puis, une fois les responsables identifiés, sur leurs pages personnelles (figure 2). Il offre deux lectures convergentes. Si l'on concentre notre attention sur les seules sections CNU, il est très clair que les formations en humanités numériques sont d'abord portées par les « Sciences de

l'information et de la communication » (section 71), suivies de loin par l'« Informatique » (section 27). En remontant d'un cran et en observant plutôt les groupes de disciplines CNU, l'« Informatique » perd un peu de poids et ce sont toujours les SIC (associées à l'« Épistémologie et histoire des sciences » qui n'apportent qu'une formation au groupe 12) qui dominent ; suivies de près par les « Langues, littératures et sciences du langage » (groupe 03).

Groupes CNU		Sections CNU	Master "HN"
12_Pluridisciplinaire | Infocom & épistémologie			19
	dont	71_Sciences de l'information et de la communication	18
		72_Épistémologie, histoire des sciences et des techniques	1
03_SHS_Langues, littératures, sciences du langage			13
		09_Langues et littératures françaises	5
		07_Sciences du langage	4
	dont	08_Langues et littératures anciennes	1
		11_Langues et littératures anglaises	1
		12_Langues et littératures germaniques	1
		14_Langues et littératures romanes	1
04_SHS | Arts et histoire			8
		18_Arts	4
	dont	22_Histoire moderne et contemporaine	3
		21_Histoire ancienne et médiévale	1
05_Sciences		27_Informatique	6
09_Sciences		60_Génie mécanique	2
02_Droit, économie et gestion		06_Gestion	1

Lecture : en 2019/2010, 8 masters en "humanités numériques" étaient dirigés par des enseignants-chercheurs relevant du groupe 04 du CNU (SHS - Arts et histoire) dont 1 en section 21 du CNU (histoire ancienne et médiévale)

Figure 1. Groupes et sections CNU de rattachement des responsables de masters en humanités numériques ouverts à la rentrée 2019 (N = 39)

Crédit : Émilien Ruiz

15. Les responsables des formations, qui ne sont pas toujours indiqués sur le portail trouvermonmaster.gouv.fr ont été identifiés à partir des pages des formations sur les sites des établissements. Les sections CNU des responsables des 39 formations observées ici ont été identifiées par trois principaux moyens :

- le premier, le plus simple, correspond aux sections mentionnées explicitement sur les pages et CV des enseignants-chercheurs sur leurs pages professionnelles
- le deuxième, lorsque la première méthode s'avérait infructueuse, correspond au repérage du nom des responsables dans les listes électorales du CNU
- le troisième, lorsque les deux méthodes précédentes ont échoué, correspond à une extrapolation de l'auteur à partir de la discipline de la thèse des responsables, leur département d'enseignement, voire celle des thèses dirigées.

16. Une telle méthode est loin d'être infaillible. L'un des parcours les plus complexes se trouve ainsi parmi les 3 formations considérées comme dirigées par des enseignants-chercheurs relevant de la section 22. L'un des responsables, électeur et donc membre de la section 22, est docteur en physique, vraisemblablement MCF en physique, enseignant l'informatique. Chercheur associé dans un laboratoire d'histoire, sa production dans cette discipline est principalement liée à l'engagement dans des recherches collectives en histoire des sciences. Le choix de partir des sections CNU d'appartenance et non des disciplines effectives d'enseignement m'a ainsi conduit à le placer en section 22. D'autres méthodes auraient pu être adoptées : partir des disciplines principales enseignées ou de celles des publications. Elles m'ont semblé plus aléatoires et lacunaires.

17. Il faudrait bien entendu aller plus loin que cette première exploration – en examinant, notamment, les programmes de l'ensemble des formations pour affiner leur ancrage disciplinaire[5]. Néanmoins, à ce stade, il est clair que les masters en humanités numériques sont avant tout des formations portées par les SIC et les lettres.

18. En outre, ces premiers résultats sont aussi intéressants par les absences qu'ils révèlent : aucune trace de la sociologie, l'économie ou la science politique. Ces disciplines furent pourtant pionnières et restent à la pointe de l'utilisation de méthodes et l'exploration d'objets qui sont souvent considérées comme relevant des humanités numériques. Je pense, par exemple, à l'analyse et la visualisation de réseaux (Briatte 2016) ou aux questions soulevées par le *Big data*, les réseaux sociaux numériques, etc. (Bastin et Tubaro 2018).

19. Enfin, l'examen des débouchés annoncés par ces formations est, lui aussi, très instructif : l'essentiel des masters en humanités numériques ne se destinent pas à des métiers relevant de la recherche ou de l'appui à la recherche. Parmi les 39 formations identifiées : 25 n'annoncent que des débouchés professionnels hors de la recherche ; 13 mentionnent des débouchés principalement hors de l'enseignement supérieur et de la recherche, tout en laissant cette possibilité ouverte (à travers, notamment, une possible poursuite d'études en doctorat). Parmi toutes les formations identifiées, le master « Humanités Numériques » de l'université Paris Sciences & Lettres (principalement porté par l'École des chartes) fait figure d'exception. Il s'agit en effet de la seule formation qui affiche explicitement un positionnement « résolument tourné vers la recherche », qui vise « avant tout à former des étudiants souhaitant placer leur parcours dans le contexte d'une poursuite en doctorat[6] ».

20. La très grande majorité des formations observées ici n'entend pas répondre aux défis que doivent relever nos disciplines face aux transformations numériques qui les affectent. Leur objectif principal est de pourvoir à des métiers perçus comme « nouveaux » (mais que l'on peine à identifier) ou à des métiers particulièrement atteints par leur « numérisation ». Dans les listes de métiers citées dans les fiches formations, les intitulés qui reviennent le plus souvent dans les descriptifs sont ainsi « chef de projet » ou « consultant », les métiers de l'édition et de la culture sont ainsi très représentés, autour des fonctions de « médiation ».

21. Il ne s'agit pas de contester l'intérêt et même la nécessité de telles formations : c'est une façon, pour des masters professionnels, de mieux préparer les étudiants aux transformations numériques qui affectent les métiers de

5. Des présentations de programmes de masters en humanités numériques portés en « Lettres et en Sciences du langage » ont été faites à l'occasion de l'édition 2019 de DHNord. Voir l'enregistrement vidéo : https://publi.meshs.fr/ressources/former_aux_humanites_numeriques/.

6. Présentation sur le site de l'École nationale des chartes : http://www.chartes.psl.eu/fr/cursus/master-humanitesnumeriques ; notons que les masters de Lyon et Rennes s'inscrivent dans une logique de double-diplôme (les étudiants sont à la fois inscrits en HN et dans un master disciplinaire) dont les ambitions sont assez proches.

l'art, de la culture, de l'édition, de la communication, etc. Mais ces formations laissent une question fondamentale en suspens : que proposons-nous à l'immense majorité des étudiants en SHS qui ne poursuivront pas leurs études dans ces filières ?

Pour un recentrage disciplinaire de la formation numérique en SHS

22. En partant d'une réflexion sur le cas de la formation en histoire, Paul Bertrand a récemment répondu à cette question de la façon suivante : « s'agissant de l'enseignement, les humanités numériques restent un mirage. » Développant son analyse, il insistait sur la nécessité pour nous, enseignant l'histoire dans le supérieur, « de nous mettre à l'ouvrage ». L'historien plaidait alors « pour un abandon sur le terrain de l'enseignement du concept d'humanités numériques » (Bertrand 2019).

23. Je souscris totalement à cette approche, qui dépasse assez largement la seule question du « numérique ». De ce point de vue, on trouvera difficilement meilleur guide que Bernard Lepetit. Directeur d'études à l'EHESS, il fut co-directeur du Centre de recherches historiques (CRH) et de la belle collection « L'évolution de l'humanité » de 1992 à son décès prématuré en 1996. Moderniste, spécialiste de l'histoire des villes, élève de Jean-Claude Perrot, il fut membre du comité de direction de la revue des *Annales* et, à ce titre, l'un des principaux artisans du « tournant critique » qui donna son titre à la dernière livraison de l'année 1989 (*Annales. Économies, Sociétés, Civilisations : Histoire et sciences sociales. Un tournant critique* 1989). Au cours de cette période, il joua un « rôle fondamental » (« Bernard Lepetit » 1996) dans la création et l'animation de la revue *Histoire & mesure*, dans une période où, comme il l'écrivait lui-même, « l'histoire quantitative [n'était] plus à la mode » (Lepetit 1989) ; et contribua de façon décisive aux renouvellements de l'histoire sociale (Lepetit 1993, 2013).

24. Pourquoi convoquer la figure de Bernard Lepetit ? D'abord parce que ses réflexions sur la connaissance historique et le métier d'historien, ses propositions relatives à l'interdisciplinarité, aux usages de la quantification sont d'une actualité saisissante et, pour moi, une inspiration constante. Convoquer les réflexions de cet historien me permet, ensuite, de clarifier certaines de mes prises de positions concernant les humanités numériques et, en particulier, d'exposer ici les raisons pour lesquelles un recentrage disciplinaire me semble nécessaire.

25. Bernard Lepetit plaidait pour une « pratique restreinte de l'interdisciplinarité » qui revenait, finalement, à rappeler l'importance de la diversité des approches des phénomènes sociaux pour leur compréhension. Non pour créer des silos mais pour confronter les regards et les méthodes de travail. Il prenait l'exemple de « l'économie qui étudie le mouvement des prix au XVIII^e siècle » ou de la philosophie s'intéressant à « la naissance des structures d'enfermement ». L'historien insistait alors : « s'ils se font historiens et adoptent les habitudes des historiens, la nouveauté radicale de leur regard s'éteint, leur capacité

de provocation s'émousse ». C'est que, pour Bernard Lepetit, envisager la disparition des disciplines par « annulation des différences, c'est croire que la compréhension des sociétés progresse par la réduction du nombre et de la complexité des commentaires explicatifs tenus sur elles ». Plaidant « pour l'attitude inverse », il rappelait le fait qu'une discipline relevait non seulement d'un « mode de structuration de la réalité décrite », mais aussi d'un « métier, c'est-à-dire un ensemble de procédures éprouvées qui constituent une première garantie d'un discours cohérent. » (Lepetit 1990, 1999, 303-313).

26. Or, en externalisant, vers les masters précités par exemple, la formation « au numérique » d'un nombre forcément restreint de leurs étudiants, certaines disciplines de sciences humaines et sociales autolimitent leurs champs d'investigations futures. Plus grave, elles abandonnent à d'autres les réflexions critiques qui s'imposent concernant les transformations, non seulement des disciplines mais de la société dans son ensemble. Dit autrement, s'en remettre aux seules humanités numériques et à leurs spécialistes, comme s'il s'agissait d'une autre discipline, c'est témoigner d'un incroyable manque de confiance dans les capacités de nos disciplines : à s'approprier des outils informatiques qui lui sont devenus indispensables ; mais aussi à faire du « numérique » un objet légitime de ses investigations. C'est là que se situent, à mon sens, le *repli* (et non le recentrage) disciplinaire et la position de rejet. Pour former « au numérique » leurs étudiants, les disciplines doivent s'approprier les outils informatiques et investir les objets de recherche qui en relèvent. Comme le notait récemment sur Twitter l'historien médiéviste Nicolas Perreaux : « Il est temps enfin d'aller au-delà des étiquettes. Je ne suis pas un "humaniste numérique", mais un historien qui utilise des machines pour explorer les documents du passé. "Juste" un historien, qui a pris acte des évolutions technologiques qui bouleversent nos sociétés[7]. »

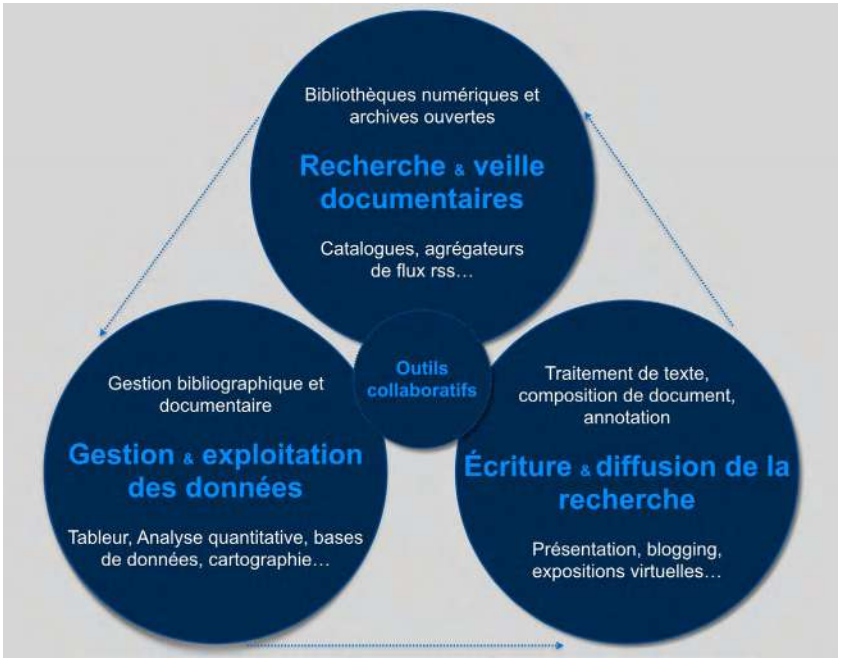


Figure 3. Un socle commun de formation numérique des historien.ne.s ?
Mise en schéma des propositions publiées dans (Heimburger et Ruiz 2011)

27. Sans présumer des intentions de Frédéric Clavert et Caroline Muller, c'est aussi à un tel recentrage heuristique que j'aurais tendance à associer leur *Goût de l'archive à l'ère*

7. Tweet de @N_Perreaux, 11 juin 2019 https://twitter.com/N_Perreaux/status/1138342158535135232.

numérique (Clavert et Muller 2017). Or, c'est à transposer ce type de propositions dans les formations disciplinaires qu'il me semble nécessaire de travailler. C'est ce que nous proposions il y a maintenant 8 ans, avec Franziska Heimburger, à travers l'idée d'un « socle commun de formation aux outils numériques » (Heimburger et Ruiz 2011) représenté sous forme de schéma dans la figure 3 *supra.*

28. Nous avions conçu cette proposition à partir de notre expérience de formation d'étudiants en master et en thèse à l'EHESS. L'idée était la suivante : partant de toutes les questions qui nous étaient posées entre deux portes dans un couloir ou dans nos séminaires de méthodologie et d'historiographie, quelles étaient les compétences minimales qu'il serait idéal de fournir à toute apprentie historienne et tout apprenti historien pour qu'ils puissent mener leurs recherches en toute autonomie, quitte à devoir suivre des formations avancées ensuite.

29. Une telle approche reposait sur la conviction selon laquelle l'usage du numérique en histoire ne relevait pas d'une spécialisation et, en tous cas, ne devait pas relever d'une formation non-historienne. Il s'agissait de partir du cœur de nos pratiques professionnelles (recherche documentaire ; exploitation des sources ; restitution et diffusion) et de les confronter au champ des possibles offert par le numérique, tout en adoptant une démarche critique et réflexive.

30. Au-delà de la métaphore de la « boîte à outils » que nous affectionnons particulièrement, l'accent devait ainsi être mis sur les démarches historiographiques et les questionnements scientifiques et non sur les logiciels. Dans un texte de 2012, Olivier Ertzscheid illustrait parfaitement ce parti pris :

31. > Former à Facebook, à Google, à Twitter est certes encore utile mais lorsque ces outils disparaîtront ou seront remplacés par d'autres, à quoi cela aura-t-il servi ? Il existe pourtant une solution simple : il faut enseigner la publication. De sa naissance jusqu'à sa mort, le Web fut et demeurera un média de la publication. Enseigner l'activité de publication et en faire le pivot de l'apprentissage de l'ensemble des savoirs et des connaissances. Avec la même importance et le même soin que l'on prend, dès le cours préparatoire, à enseigner la lecture et l'écriture. Apprendre à renseigner et à documenter l'activité de publication dans son contexte, dans différents environnements. Comprendre enfin que l'impossibilité de maîtriser un « savoir publier », sera demain un obstacle et une inégalité aussi clivante que l'est aujourd'hui celle de la non-maîtrise de la lecture et de l'écriture, un nouvel analphabétisme numérique hélas déjà observable. (Ertzscheid 2012)

32. Dans une perspective similaire, nous affirmions que ce n'est pas le fait d'enseigner l'usage d'un tableur ou d'un logiciel de base de données qui devait primer, par exemple, mais bien celui de former à la mise en lignes et en colonnes. Autrement dit, d'enseigner la mise en données de documents, voire de corpus documentaires, qu'ils soient textuels, statistiques, visuels ou autre, en vue de leur traitement et de leur analyse historique.

33. Mais à travers les formations que nous avons conçues et dispensées depuis, nous avons finalement, peut-être, contribué à renforcer l'idée selon laquelle les « enseignements au numérique » devaient être l'affaire, sinon des seuls informaticiens ou professionnels des sciences documentaires, à tout le moins de celles et ceux qui, parmi les historiennes et historiens s'y intéressaient le plus. Dit autrement et de façon familière, même envisagée au sein des disciplines, la formation au numérique reste souvent l'apanage des *geeks* de service.

Décloisonner le numérique et généraliser la formation par la recherche

34. Pourtant, en faisant des enseignements théoriques et pratiques du numérique une affaire de spécialistes, on place ces derniers en périphérie de la discipline et l'on met en danger la pérennité de ces formations. Cette situation devient en effet délétère lorsque les enseignements concernés doivent faire face aux restrictions budgétaires et à la diminution des heures d'enseignement, au départ des « M. ou M^{me} numérique » locaux ou, cela peut arriver, à l'hostilité de principe des responsables de formation. On imagine mal, quand elles existent, les formations à la lecture critique de sources, à la rédaction de commentaires de documents ou de dissertations disparaître d'une licence en histoire. Un enseignement de paléographie médiévale trouvera toujours un titulaire, en particulier en L3 ou en master. En revanche, on aura beaucoup moins de difficulté à voir disparaître une initiation au tableur en histoire économique faute de « spécialiste » dans l'équipe, et l'on considérera souvent qu'instaurer une formation à l'écriture d'un mémoire (que ce soit avec un papier et un crayon ou avec LaTeX) relève de la perte de temps.

35. Le recentrage disciplinaire ne suffit donc pas. En plus de faire en sorte que les formations « au numérique » en histoire soient dispensées par des historiens, il faut, de surcroît, qu'elles soient pleinement intégrées aux enseignements « traditionnels », qu'ils soient thématiques ou méthodologiques. Sans renoncer à tout atelier ou enseignement méthodologique spécialisé, il faut « décloisonner » les enseignements numériques.

36. Interrogée à ce propos en 2012, Claire Lemercier répondait que le véritable défi était celui de la « normalisation : permettre l'appropriation par tous des outils de la "*digital history*", pour que ce terme, à la limite, devienne inutile, parce qu'elle se confondrait avec l'histoire tout court » (Grandi et Ruiz 2012). Ce défi n'a jamais été autant d'actualité car, en dépit d'efforts réels dans de nombreuses universités, il ne va toujours pas de soi qu'une formation solide au métier d'historien suppose la maîtrise d'un environnement technique numérique et l'acquisition de réflexes critiques qui permettent de questionner ce que change, par exemple, le fait de travailler sur un corpus numérisé, du point de vue de la critique de la source elle-même.

37. Une illustration, aussi anecdotique qu'évocatrice, permet de prendre la mesure de la distance qui nous sépare encore de cette normalisation. En janvier 2019, Sébastien Poublanc publiait un retour d'expérience pédagogique où il présentait la façon dont il avait tenté d'intégrer les questions « numériques » dans les travaux dirigés d'histoire de l'Europe moderne dont il était chargé. Il avait adopté deux partis pris. Le premier consistait à demander aux étudiants des travaux qui supposaient l'application de formations reçues au semestre précédent (traitement de texte et logiciel de bibliographie). Le second consistait à demander aux étudiants de réaliser leurs travaux d'histoire moderne en exploitant de la documentation numérique ; ce qui impliquait de les accompagner dans l'apprentissage de la lecture critique et historienne de ce type de ressources (Poublanc 2019). Partagé sur les réseaux sociaux, ce billet y avait été commenté par une personne ayant statut de professeur des universités dans un établissement où l'offre de formation au « numérique pour historiens » se limitait à un total de 24 h obligatoires pour l'ensemble des trois années de licence (soit 1,6 % du total en équivalent TD). Au billet tentant de réfléchir aux façons de « faire du "numérique" dans un TD d'histoire » moderne, il fut répondu, « et faire de l'histoire dans un TD d'histoire, c'est possible ? »

38. S'il est difficile d'objectiver le caractère plus ou moins diffus de ces réticences, il est essentiel de les dépasser. Car il est indispensable de procéder, au moins partiellement, à une dé-spécialisation des enseignements dits « numériques » en histoire. Cela revient donc à affirmer, non seulement la nécessité de faire dispenser ces formations par des historiens, mais aussi de les intégrer pleinement aux enseignements « traditionnels ». En effet, il n'y aura pas de « conversion numérique » de nos disciplines et, surtout, de leurs étudiants, si l'ensemble des équipes enseignantes ne se saisissent pas de ces questions. C'est d'autant plus envisageable que la généralisation de l'usage de « l'ordinateur-personnel-relié-à-Internet » (Delalande et Vincent 2011) a conduit à une « numérisation du métier » (Muller 2018), des pratiques de recherche, qu'il faudrait davantage mobiliser dans nos enseignements, dès le premier cycle.

39. Seul un décloisonnement de la formation « au numérique » permettra de normaliser ces pratiques. Cela passe d'ailleurs aussi par une réaffirmation de l'importance du collectif dans la fabrique de nos programmes et nos méthodes d'enseignement. Non seulement pour construire ensemble des programmes cohérents, mais aussi pour mutualiser nos compétences et, surtout, pour faire en sorte que les étudiants saisissent d'emblée l'intérêt des formations d'apparence « technique » qu'ils reçoivent.

40. Le corollaire d'un tel positionnement, c'est de plaider pour un renforcement et une généralisation de la formation par la recherche dès le premier cycle. Si les historiens ne sont pas les seuls à travailler avec des corpus d'archives et autres types de documents, on peut sans trop de risque avancer qu'ils en ont une pratique relativement ancienne et réflexive. Or, cela ne se traduit pas toujours explicite-

ment dans nos pratiques pédagogiques. Je me souviens ainsi de l'un de mes enseignants en DEUG qui, il y a près de 20 ans déjà, pestait contre le sacro-saint commentaire de document en TD. Non pour affirmer qu'il était inutile, mais pour souligner qu'il ne correspondait en rien aux pratiques réelles : la réunion, le croisement, la confrontation de *plusieurs* documents. Cette expertise de la critique des sources, à l'identification, au traitement et à l'exploitation de documents, devrait être beaucoup plus mise en avant dans les formations « au numérique ».

41. Quelques exemples permettent d'illustrer le potentiel d'une telle approche, qu'il s'agisse d'intégrer « le numérique » aux formations classiques ou de faire appliquer des techniques apprises dans des enseignements dédiés. Lorsque j'étais chargé des enseignements « numériques » pour les étudiants en deuxième année de licence en histoire à l'université de Lille, l'une des demi-UE concernées permettait d'initier les étudiants aux rudiments de l'analyse quantitative avec un tableur. Je dispense ce type d'enseignements en premier cycle depuis 2006. Or, bien que j'aie toujours tenté de familiariser les étudiants à ces techniques en lien direct avec des problématiques historiques, une partie non négligeable validait ce module en travaillant correctement pendant le semestre concerné, puis oubliait tout. Il est ainsi arrivé que des étudiants en master qui avaient suivi ces enseignements me redemandent une formation car, cette fois, ils en avaient « vraiment » besoin. On peut, c'est un sport très pratiqué, blâmer les étudiants pour leur manque de mémoire et d'imagination. On peut aussi s'interroger sur nos propres pratiques pédagogiques et nous dire qu'il y manque peut-être quelque chose. Ce quelque chose, c'est la coordination avec les autres membres de l'équipe enseignante pour la construction de convergence avec d'autres cours.

42. J'en ai particulièrement pris conscience lorsqu'une collègue médiéviste m'a raconté la réaction de ses étudiants lorsqu'elle leur a demandé d'ouvrir un tableur dans le cadre d'un enseignement de L3 sur les sources du Moyen Âge. La quasi-totalité des étudiants avait suivi mon enseignement de deuxième année mais, lorsque Esther Dehoux les confrontait aux rouleaux des morts et les enseignes de pèlerinage, alors que la question de leur mise en données se posait, ils ne pensaient pas d'emblée à mobiliser un tableur (ni même simplement leur ordinateur), qu'ils associaient peut-être trop étroitement à l'histoire économique et sociale contemporaine. En outre, ils se trouvaient par ailleurs bien démunis face à la question de leur cartographie, aucun enseignement de ce type n'étant proposé en licence d'histoire.

43. Autre exemple lillois, Matthieu De Oliveira dispense depuis plusieurs années un enseignement de méthodologie en L3 visant à sensibiliser les étudiants à la création, la collecte et la conservation des archives. Dans ce cadre, il développe chaque année un projet collectif en lien avec les AD59 (archives départementales du Nord). Au second semestre 2019, il a donné un petit aperçu, sur Twitter, du travail de ses étudiants sur les notices biographiques de militants anarchistes du Nord pour la rédaction d'entrées

proposées au *Maitron*, le dictionnaire biographique du mouvement ouvrier[8]. J'étais alors déjà en détachement à Sciences Po, mais j'ai immédiatement imaginé le potentiel d'une collaboration avec la mise en place d'un atelier méthodologique sur la mise en données d'informations biographiques et sur les bonnes pratiques et les outils mobilisables pour une approche prosopographique. Mais, même si j'avais été sur place, j'aurais dû faire face à une difficulté : il n'existait pas d'enseignements numériques en L3 histoire.

44. On peut aussi imaginer ce type de convergences au niveau master. Avec Franziska Heimburger par exemple, pour un atelier obligatoire d'histoire quantitative et histoire numérique en master histoire à Sciences Po, nous avons décidé de mettre en application cette approche. Conjointement avec Odile Gaultier-Voituriez, historienne à Sciences Po et responsable de la documentation du CHSP et du CEVIPOF, et Émeline Grolleau, archiviste du CHSP, nous avons identifié cinq fonds, riches et facilement accessibles pour les étudiants[9]. L'atelier permet d'aborder les différentes méthodes quantitatives, notamment à travers la discussion d'articles. Au terme du semestre, les étudiants présenteront une esquisse d'exploitation quantitative à partir du fond de leur choix, avec l'une des méthodes évoquées. En outre, ils bénéficient, en parallèle, d'un autre atelier obligatoire sur les archives animé par Odile Gaultier-Voituriez.

45. Ces trois exemples, limités et imparfaits, donnent une idée du potentiel de l'approche consistant à placer la formation « au numérique » au sein de dispositifs pédagogiques centrés sur la formation par la recherche. C'est, en outre, une piste qui me semble à même de réaffirmer la spécificité, la richesse et l'importance des parcours de formation en sciences humaines et sociales.

Conclusion

46. Les propositions qui viennent d'être énoncées ne peuvent prendre de sens que si elles sont le fruit d'une réelle collaboration des équipes enseignantes et d'un véritable souci des échanges pédagogiques. Cet enjeu de la concertation est, en réalité, celui qui me semble devoir primer sur tous les autres. Si l'importance du collectif est de plus en plus affirmée concernant la recherche (comme en témoignent nombre de pages de « remerciements » des thèses par exemple ; mais aussi des projets relevant des humanités numériques) il n'en va pas de même concernant l'enseignement.

47. Il faut dire que les historiennes et historiens du supérieur éprouvent parfois quelques difficultés à reconnaître l'importance des réflexions collectives en matière pédagogique. Étienne Anheim le soulignait récemment en ouverture du savoureux chapitre « Enseigner » du livre tiré de

8. Tweet de @MatthieudeO, 11 février 2019 : https://twitter.com/MatthieudeO/status/1095056130257158146.

9. Il s'agit des Fonds Hubert Beuve-Mery ; Robert Blum ; Couve de Murville ; Émile Mayer ; et André Siegfried. Pour plus d'informations voir : http://chsp.sciences-po.fr/fonds-archives/les-fonds.

son mémoire de synthèse pour l'habilitation à diriger des recherches : nous faisons rarement de nos pratiques d'enseignement un « objet intellectuel à part entière en discutant publiquement ses fondements méthodologiques et ses liens intellectuels avec nos activités de recherche » (Anheim 2018). À ce titre, il faut saluer les trop rares initiatives qui tentent d'amorcer des réflexions et discussions collectives (notamment Anheim et Girault 2015 ; Mazeau 2016 ; Galvez-Behar 2017). Informelles ou institutionnalisées, elles sont pourtant indispensables à la progression de la qualité de nos enseignements et à une meilleure adaptation aux besoins des étudiants.

48. Deux ans, jour pour jour, avant la journée d'ouverture de l'édition 2019 de DHNord, avec Paul Bertrand, historien médiéviste à l'université de Louvain, nous concluions, dans les mêmes locaux, une journée d'études qui avait consisté à faire échanger des collègues lillois et louvanistes à propos de l'enseignement de la critique des sources[10]. Cette journée avait été une sorte de « révélation » : partis pour échanger principalement autour du numérique, ce sont les discussions pédagogiques qui dominèrent les débats. Tous les participants s'accordèrent sur un constat : nous n'avions jamais échangé de la sorte sur nos pratiques d'enseignement. Les aléas des mobilités professionnelles nous ont, pour le moment, retardés dans l'organisation d'autres initiatives de ce type, mais nous y reviendrons certainement à l'avenir et je ne peux que vous encourager à mener des expériences similaires, quitte à le faire à l'échelle des facultés, voire des départements. Parlons pédagogie et mutualisons nos compétences !

49. Salutaire d'un point de vue général, une telle démarche permettra aussi, j'en suis persuadé, d'identifier tous les « M. et Mme Jourdain » sans l'aide de qui nous n'arriverons jamais à faire en sorte que l'histoire « numérique » se confonde avec l'histoire « tout court » ; que faire des « humanités numériques » ne soit pas autre chose que faire des humanités, à l'ère numérique.

Pour consulter les données mobilisées dans le chapitre, voir **https://hnso-corpus.nakala.fr/**

10. Enseigner la critique historique à l'ère numérique, annonce sur le site de la MESHS Lille Nord de France : https://www.meshs.fr/page/enseigner_la_critique_historique_a_ere_numerique.

Données

Les données mobilisées dans certains chapitres sont librement accessibles en ligne dans un entrepôt Nakala à l'adresse suivante : **https://hnso-corpus.nakala.fr/**

Pour cet ouvrage, les chapitres concernés sont :

- Frédéric Glorieux : « Le corpus de tous les livres depuis les débuts de l'imprimerie, tous comptes faits… ».
- Alix Chagué, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie : « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non uniforme ? ».
- Andrea Del Lungo et Karolina Suchecka : « Le projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles ».
- Émilien Ruiz : « Former "au numérique" en sciences humaines et sociales ? Propositions d'un historien ».

Structurer automatiquement un corpus homogène issu de la reconnaissance d'écriture manuscrite : les jugements du Conseil des prud'hommes des tissus parisiens

Victoria Le Fourner, Alix Chagué, Manuela Martini et Anaïs Albert

Le contexte de recherche

1. Le projet *TIME-US*[1] s'inscrit dans la lignée d'une historiographie en plein essor sur l'industrialisation en France et en Europe attentive à la place accordée au travail des femmes et des enfants (Martini et Albert 2021)[2]. Dans la mouvance d'un intérêt renouvelé pour l'étude des salaires et de l'économie domestique des ménages ouvriers, le projet « Rémunérations et usages du temps des femmes et des hommes dans l'industrie textile en France de la fin XVII^e siècle au début du XX^e siècle », aussi nommé *Time-Usage* ou *TIME-US*, est un programme de recherche débuté en 2017, soutenu par l'Agence nationale de la recherche (ANR), et coordonné par Manuela Martini, professeure d'histoire contemporaine à l'Université Lumière Lyon 2 et membre du Laboratoire de recherche historique Rhône-Alpes (LARHRA). Associant des spécialistes en histoire économique, en ethnométhodologie, en humanités numériques et en informatique, le projet s'est intéressé aux enjeux socio-économiques liés à la définition des rémunérations dans l'industrie textile en France pendant la première et la deuxième industrialisation. Dans le but d'étudier les liens entre rémunération, genre et statut des travailleurs, les quatre équipes réunies autour de *TIME-US* se sont interrogées sur la nature du travail réalisé par les différentes personnes impliquées dans le processus productif, sur la place du salaire dans les revenus des ménages, sur l'évolution des tâches et des modes de paiement et, plus largement, sur l'évaluation du travail effectué et les conflits que cela pouvait engendrer.

2. Pour ce faire, la recherche s'est orientée d'emblée non seulement vers la collecte de séries quantitatives mais également vers le repérage de différents corpus de sources qualitatives imprimées et manuscrites : manuels et enquêtes publiés, mémoires individuels et presse ouvrière, pétitions et rapports de police, sources du

1. *TIME-US*, « Rémunérations et usages du temps des hommes et des femmes dans le textile en France de la fin du XVII^e au début du XX^e siècle » (2017-2021), ANR appel à projets générique 16-CE26-0018-01 2016.
2. Voir le chapitre d'Alix Chagué, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie dans cet ouvrage : « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non uniforme ? ». Pour consulter les données mobilisées dans le chapitre, voir https://hnso-corpus.nakala.fr/.

contentieux au travail, faillites et registres des contraventions des corporations pour l'époque moderne. L'ampleur des corpus a rendu nécessaire l'usage de technologies numériques afin de traiter la masse de documents rassemblés et ainsi en extraire des informations de même nature pour les comparer dans les régions industrielles étudiées (Lyon, Paris, Lille-Tourcoing principalement). L'objectif était ainsi de rassembler un ensemble de sources pour une période longue sur les salaires et les revenus en nature des travailleurs et travailleuses, selon le temps d'exécution, les tâches accomplies, le type de rémunération, les périodes d'activité, le statut, le sexe, l'âge (dans la mesure du possible) en les situant dans leurs lieux de production (atelier, usine, domicile).

3. C'est dans ce contexte que les jugements du Conseil des prud'hommes des tissus de Paris ont été repérés et numérisés puisqu'il est possible d'en tirer des informations sur les modalités de rémunération et de travail des ouvriers et ouvrières. Les affaires présentant le plus grand intérêt pour le projet *TIME-US* sont celles qui offrent une confrontation entre les parties. Les donneurs d'ordre et les ouvriers et ouvrières doivent alors détailler l'objet du litige et incluent des précisions sur les salaires et les conditions de travail.

4. Institution fondée en 1806 à Lyon, et présente à partir des années 1840 dans plusieurs départements, le Conseil des prud'hommes est un organe clé pour étudier les enjeux liés à la rémunération, au temps d'exécution des produits et plus largement aux conflits autour des usages propres à chaque métier ou à la fabrication de produits. À Paris, le Conseil des prud'hommes du textile voit le jour en 1847 (Cottereau 1987). Destiné à concilier les différends entre fabricants et ouvriers, ou entre les chefs d'atelier et leurs travailleurs, il exerce un contrôle administratif sur le nombre des métiers existants et détient un pouvoir de police en cas de désordre grave dans l'atelier. L'industrie textile parisienne est alors très importante et Paris est l'un des principaux centres français de la confection (Albert 2021). Les prud'hommes sont une justice de conciliation où il est prévu plusieurs étapes de règlement des conflits : les plaignants passent d'abord devant le bureau particulier, qui est destiné à trouver une solution à l'amiable entre les deux parties. En cas d'échec, l'affaire passe ensuite devant le bureau général qui rend un jugement. Les conseillers sont habilités à entendre des témoins et peuvent également se rendre dans les ateliers en cas de besoin (Cottereau 2006).

5. Entre mars et juillet 2019 le corpus des jugements des prud'hommes de Paris a été photographié, transcrit et structuré[3]. Le corpus sur lequel porte ce *data paper* est issu d'archives publiques, conservées aux Archives de Paris. Ce sont des registres, rédigés par le greffier du tribunal – appelé aux prud'hommes « secrétaire » – à partir des notes prises lors d'une audience, qui mentionnent parfois de précédentes tentatives de conciliation auprès du Bureau Particulier[4]. Le registre suit le déroulement

3. Les années 1847-49, 1858, 1868 et 1878 ont été intégralement photographiées.
4. Pour plus de détails, l'article de blog rédigé à ce sujet écrit par Kevin Champougny en 2019 peut être consulté : https://timeus.hypotheses.org/567.

chronologique des audiences, jugement par jugement. Ces minutes sont des actes authentiques, conservés par l'autorité qui les a produits et faisant mention de la décision de justice, jusqu'à leur versement aux archives départementales[5].

Le protocole de production des données

Collecte des numérisations

6. La constitution et collecte des données numériques pour l'ensemble du projet *TIME-US* ont déjà fait l'objet d'une description dans un autre chapitre (Chagué *et al.* 2022). Il faut toutefois rappeler ici la chaîne de traitement suivie pour les registres du Conseil des tissus. Un dépouillement exhaustif de l'année 1858 a permis aux historiens de travailler sur la conflictualité des travailleurs[6]. Cette année a retenu l'attention des chercheurs puisqu'il s'agit d'un moment clé dans l'institutionnalisation du Conseil et dans l'essor de la confection parisienne. Afin de pouvoir établir des comparaisons diachroniques, les registres des années 1847-1849[7] ont également été collectés, ainsi que ceux de 1868 et 1878. Pour faciliter la transcription et la comparaison, un échantillonnage a été réalisé sur les mois de janvier et juin de chacune des années retenues[8].

7. Après avoir repéré les sources, la collecte – phase cruciale du projet puisque l'ensemble des traitements dépendait de sa bonne réalisation – a été réalisée en photographiant les pages des documents à l'aide d'une ScanTent[9], à partir de 2019. Une uniformisation des fichiers numériques était nécessaire. Avec la ScanTent et le logiciel de prise de vue associé, ScanTailor[10], il était possible de mieux cadrer les images pour limiter le bruit. Un dossier partagé hébergé par le service de stockage ShareDocs, obtenu par l'intermédiaire de l'Infrastructure de recherche Huma-Num, a permis de centraliser les numérisations afin que tous les chercheurs y aient accès pendant toute la durée du projet. Les images étaient rassemblées par lots en fonction de leur côte et de leur établissement d'origine (Chagué 2018)[11].

5. Ce sont dans les années 1960 que sont intervenus les premiers versements d'archives du conseil de prud'hommes aux Archives de la Seine. Ils concernent les minutes de jugements et les dossiers pour les années 1908 à 1954 (Lainé 2006).
6. Archives départementales de Paris, D1U10 386.
7. Le nombre d'audiences pour le début de l'institution était moins important que pour les années 1858 et 1878, c'est donc l'ensemble qui a été sélectionné afin d'avoir des sets de données équivalents pour les trois périodes.
8. Il s'agissait ici d'éviter la morte-saison du textile.
9. Cf. https://readcoop.eu/scantent/
10. Cet outil permet notamment de couper ou de recadrer des pages, de compenser l'angle d'inclinaison ou d'ajouter/supprimer des champs de contenu et des marges. Une fois le traitement effectué sur toutes les prises de vues d'un même ensemble documentaire, il est possible de récupérer le scénario des traitements effectués afin de l'appliquer sur un autre ensemble documentaire ou encore de connaître le détail des modifications subies par les fichiers.
11. Ce service destiné aux chercheurs permet de stocker, mettre à jour les fichiers et les échanger au sein du programme de recherche de façon sécurisée. Il ne permet pas d'exposer les données mais uniquement de les stocker. Cet espace de stockage commun est essentiel afin de donner un aperçu global de toutes les ressources disponibles mais aussi d'éviter de perdre des données lors de l'échange entre chercheurs.

8. La deuxième phase consistait à acquérir la transcription des comptes rendus, à les structurer, puis à leur appliquer des méthodes de TAL en vue de leur analyse pour la troisième année du projet. Cette phase du projet mobilisait essentiellement les ingénieurs et chercheurs en informatique, mais également les historiens : il s'agissait d'identifier leurs besoins au fur et à mesure que les données étaient traitées. Par ailleurs, l'expertise des historiens pour la validation des transcriptions et l'élaboration de la modélisation de la structure était irremplaçable.

Transcription et transformation

9. La première chaîne de traitements (*workflow*) a été élaborée à partir de 2018. Un système d'extraction du texte s'appuyant sur le logiciel Transkribus[12], alors gratuit, permettait de faciliter la transcription des documents[13]. L'utilisation d'un modèle de transcription affiné petit à petit permettait d'obtenir une première version du texte qu'il s'agissait ensuite de corriger manuellement. Une fois cette correction effectuée directement sur l'interface du logiciel Transkribus, le programme ExportFromTranskribus[14], développé par Alix Chagué pour le projet, permettait de récupérer un fichier XML TEI unique étant donné un ensemble de pages (1 page = 1 image). Le programme s'appuie sur la fonctionnalité d'export de Transkribus vers XML TEI. Un second programme, StructurationMinute développé par Victoria Le Fourner[15], permettait de modifier cet arbre XML de manière à recomposer la structure logique, interne à chaque registre du Conseil des tissus parisiens (Le Fourner 2019). Les deux programmes, associés à un scénario de transformation XSLT, ont ainsi permis la création d'un prototype de visualisation d'une partie du jeu de données déployé sur les serveurs hébergés à l'Inria[16].

10. Si le choix de Transkribus en tant que logiciel de reconnaissance de caractères manuscrits s'est avéré pertinent au début du projet, celui-ci a fait l'objet de nombreuses discussions tout au long du projet, notamment liées à l'horizon d'un changement de modèle économique pour le logiciel, devenu payant à l'automne 2020. En outre, les taux d'erreur des modèles de transcription n'étaient alors pas pleinement satisfaisants. Pour effectuer des traitements comme la détection automatique d'expressions précises dans le texte, celui-ci doit être parfaitement lisible. Un texte avec un taux d'erreur par caractère passant en dessous du seuil des 20 % commence à deve-

TIME-US avait demandé un espace de 80 gigaoctets (Go) à l'ouverture de l'espace de stockage et a ajouté en juin 2019 110 Go afin d'accueillir des données supplémentaires. Cf. https://documentation.huma-num.fr/sharedocs-stockage/.

12. Développé par le projet READ (Recognition and Enrichment of Archival Documents) à l'université d'Innsbruck, Transkribus est une plateforme de reconnaissance automatique de la structure et du texte de documents. Cf. https://readcoop.eu/transkribus/.

13. L'entraînement du modèle a été réalisé grâce à des transcriptions manuelles réalisées par des étudiants stagiaires des masters d'histoire des universités de Lyon 2 et Paris Cité (Laurie Vanneau, Laura Bey et Kevin Champougny).

14. Cf. https://gitlab.inria.fr/almanach/time-us/ExportFromTranskribus

15. Cf. https://gitlab.inria.fr/almanach/time-us/schema-tei/-/tree/master/E%20-%20Structuration%20automatique/E.1%20-%20StructurationMinute

16. Cf. http://timeusage.paris.inria.fr/prudhommes-paris-19e/home.html

nir lisible par un humain, mais au-dessus de 10 %, les erreurs étaient encore trop nombreuses pour nos traitements informatiques[17].

11. Des corrections manuelles post-traitements ont donc été nécessaires sur certaines informations cruciales pour le projet. Travailler avec un logiciel en partie fermé posait des difficultés pour comprendre comment améliorer nos données afin d'obtenir de meilleurs modèles de transcription. Avec l'apparition des premières versions fonctionnelles de l'application eScriptorium[18] début 2020, le choix est devenu évident de basculer vers une solution libre et gratuite, pour laquelle il était possible de développer des fonctionnalités répondant aux besoins du projet *TIME-US*. Grâce à la collaboration d'une partie des membres de l'équipe ALMANACH (dont Alix Chagué et Lionel Tadjou) avec l'équipe SCRIPTA en charge du développement d'eScriptorium, les données *TIME-US* ont ainsi permis de réaliser des tests et d'élaborer des propositions de fonctionnalités, notamment visant à la correction et à la structuration des textes et ensuite à la mise en place d'une interface de consultation des données. L'idéal était de se tourner vers une solution libre et gratuite afin de pouvoir développer des fonctionnalités propres à *TIME-US*[19]. Les reprises manuelles post-traitement sur les données ont donc été très importantes à ce moment du projet afin de pouvoir avancer vers l'analyse des sources par les historiens.

12. Cette première chaîne de traitements avait conduit à un prototype de visualisation des résultats destiné à aider les historiens dans la navigation au sein des données numériques mais elle n'était pas optimale et nécessitait d'être retravaillée. Elle a été révisée en 2021 d'une part en raison des modifications non rétroactives apportées par Transkribus à ses fonctionnalités d'export (avec par exemple le passage d'ALTO 2 à ALTO 4) mais surtout en raison du transfert des données vers eScriptorium qui ne proposait pas encore, en février 2022, de fonctionnalité d'export vers de la TEI, malgré plusieurs travaux en cours. Le programme de structuration automatique, révisé par Alix Chagué à partir du programme initial, apportait plusieurs modifications. Les deux plus importantes étaient d'une part le point de départ (l'application eScriptorium déployée par l'Inria) et d'autre part le format de départ (du texte brut au lieu d'un fichier XML structuré). Ce nouveau format de départ correspond à l'état du document au milieu de l'exécution du programme StructurationMinute[20]. Reprendre la chaîne de traitements a ainsi été l'occasion d'optimiser la formulation des expressions régulières (*regex*)[21] et la construction de l'arborescence XML. La structura-

17. Les taux d'erreurs par caractères relevés avec Transkribus étaient alors d'environ 15 % pour les jugements du Conseil des tissus de Paris.

18. E-Scriptorium est un environnement virtuel de transcription, servant d'interface graphique à Kraken, un moteur de transcription automatique. Cf. https://escriptorium.fr/.

19. La compatibilité des exports Transkribus avec les imports dans eScriptorium s'est améliorée mais elle n'est pas encore complète et basculer en cours de route de Transkribus vers eScriptorium a représenté un effort de conversion et de contrôle non négligeable qui a ralenti la généralisation de la transcription au reste du corpus.

20. Ligne 373 du code.

21. Chaîne de caractères suivant une syntaxe précise qui permet de repérer un ensemble de chaînes de caractères précis au sein d'un texte.

tion de l'arborescence a été établie en reprenant celle du premier *workflow*. Un travail similaire a également été entrepris sur le corpus lyonnais des comptes rendus des audiences prud'homales présents dans la presse ouvrière entre 1831 et 1850 pour pouvoir ensuite comparer les éléments sur des années ou périodes similaires[22].

Établissement de la structure et des annotations sémantiques

13. Les jugements du Conseil des tissus possèdent une structure logique interne qu'il fallait pouvoir identifier afin de structurer les transcriptions et pour pouvoir cibler les sections intéressant le projet *TIME-US* (des expressions ponctuelles ou des passages longs). La grille de structuration, précise et formalisée par un schéma XML TEI, a été élaborée à partir d'échanges entre les historiens et les spécialistes des traitements informatiques ainsi que grâce à l'analyse d'un échantillon d'exemples issus du corpus. Le but était d'obtenir la structure la plus généralisable possible. Les premiers exemples ont été annotés à la main[23], puis le programme de structuration automatique a été mis en place pour reproduire ces annotations. En élargissant progressivement les exemples annotés par le programme, on en testait les limites afin de renforcer la robustesse des motifs de recherche. En effet, annoter la structure logique d'un ensemble de documents nécessite d'en avoir étudié plusieurs et de pouvoir assurer la cohérence entre tous (Fort, Ehrmann et Nazarenko 2009).

14. Parmi les niveaux de structuration essentiels aux historiens, on trouve les audiences et les jugements (ou affaires) : le changement de date signale le début d'une nouvelle audience, tandis que les signatures signalent la fin d'une affaire. Le passage de l'image au texte brut ne nous permettait pas de nous appuyer sur des indices visuels, l'ajout de balises TEI rendant compte de ces transitions devait pallier ce manque de repère afin de retrouver la structure. L'ajout de ce balisage permet aussi de rendre mieux visibles des transitions pour lesquelles les indices visuels étaient faibles.

15. Afin de conserver les données annotées de manière pérenne et de les rendre interopérables[24], l'usage d'un cadre comme celui proposé par la Text Encoding Initiative (TEI) semblait le plus judicieux. Toutefois, il est important de souligner que, même si tous les textes peuvent être structurés à l'aide de ce standard, il est davantage préconisé pour le traitement de textes littéraires plutôt que de textes juridiques. Nous avons pu

22. Les revues suivantes ont été entièrement dépouillées afin d'extraire les comptes rendus des audiences du Conseil des prud'hommes des tissus de Lyon : *L'Écho de la Fabrique* (1831-1834) ; *L'Écho des travailleurs* (1833-1834) ; *L'Indicateur* (1834-1835) ; *La Tribune lyonnaise* (1834-1835) ; *L'Écho des ouvriers* (1840-1841) ; *L'Écho de la Fabrique de 1841* (1841-1845) ; *L'Écho de l'industrie* (1845-1846) ; *L'Avenir* (1846-1847) ; *Tribune prolétaire* (1845-1850). Pour un exemple d'étude qualitative sur ces comptes rendus voir (Martini 2021).

23. L'annotation correspond à l'ajout de métadonnées de différentes natures sur un document. Pensée pour être lue et interprétée par une machine, elle permet de retrouver une information précise.

24. L'interopérabilité désigne le fait que des systèmes informatiques ou téléphoniques puissent s'adapter afin de collaborer avec d'autres systèmes indépendants, afin de créer un réseau et de faciliter le transfert de données.

nous appuyer sur la régularité des minutes parisiennes qui sont toujours structurées de la même manière peu importe l'année[25].

16. Cette structure reprend celle d'un procès : le dispositif[26], l'appel du rôle, les plaidoiries puis le jugement. Ainsi, pour chaque minute de jugement d'affaires on trouve le déroulé suivant : tout d'abord, la présentation par le secrétaire des parties du procès (demandeurs d'un côté et défendeurs de l'autre) en précisant leurs noms, qualités et adresses ; ensuite, un rappel du déroulé des audiences précédentes, appelé « point de fait », devant les Bureaux Particuliers. C'est dans cette partie qu'il est possible d'obtenir des informations sur l'objet des litiges et les argumentaires de chaque partie. Après le rappel des faits et des argumentaires, intervient le « point de droit »[27]. Cette partie est composée d'une ou deux phrases interrogatives. Enfin viennent les arguments des juges, puis la décision du conseil sur l'affaire. C'est l'utilisation de formules récurrentes par le secrétaire du conseil dans chaque affaire qui a permis d'automatiser le repérage de la structure logique à un niveau très fin. Une audience est identifiée comme le texte, c'est-à-dire la chaîne de caractères, débutant et se concluant par l'expression « Audience Du [jour de l'audience] ». Une présentation de la date de l'audience et des juges suit et au cours de chacune d'entre elles, plusieurs affaires se succèdent. Elles démarrent toujours par un rappel préalable « dudit jour », suivi de la date de l'audience, et ensuite « entre… » pour donner l'identification des parties ; excepté la première affaire de l'audience qui débute directement par « entre ». Suivant le souhait des historiens du projet, le prototype de visualisation rend immédiatement visible la distinction des différentes parties avant de présenter l'affaire[28]. La structure pour chaque audience est alors la suivante :

```
<div type="courtHearing">
  <div type="case">
        <opener/>
        <div type="identificationParties">
            <span type="part"/>
            <span type="part2"/>
        </div>
        <div type="pointDeFait"/>
        <div type="pointDeDroit"/>
        <div type="judgement"/>
  </div>
</div>
```

17. Outre la structuration générale des données permettant d'établir des analyses quantitatives, le projet *TIME-US* avait également pour objectif l'annotation des données à l'échelle des mots et des phrases. Différents types d'élé-

25. Contrairement aux comptes rendus de presse lyonnais dont la structuration faisait l'objet, en parallèle, d'un travail spécifique.

26. Décision des jugements portant sur les affaires plaidées antérieurement en présence des justiciables.

27. Les questions juridiques auxquelles les conseillers prud'homaux doivent apporter une réponse.

28. C'est sur cet élément que le prototype de visualisation s'est appuyé afin de présenter clairement au-dessus du texte la détection automatique des parties en présence.

ments porteurs de sens, définis comme des entités nommées, peuvent être annotés (Poibeau 2005). Il est possible de les grouper en catégories :

- Les expressions liées à l'identité des parties : statut dans le processus de production (apprenti ou apprentie, maître, ouvrier ou ouvrière), nom du métier, adresse, statut juridique (mineur ou majeure), statut matrimonial (célibataire, séparé de corps et de biens, etc.), place dans le procès (demandeur, défendeur), présence au procès (comparant, défaillant)
- Les expressions liées au procès : les débats à la suite des accusations de vol ou de perte de marchandise, les étapes du procès (bureau particulier, en dernier ressort)
- Les expressions de ce qui peut être mesuré : temps, rémunérations et coûts, marchandises

18. À partir de ces multiples entités nommées et les moments possibles de leurs apparitions dans le texte, il a été possible d'établir un schéma avec une granularité fine autorisant un certain nombre de possibilités contraintes. La documentation de ce schéma est incluse dans le protocole de contrôle des données grâce à un ODD (*One Document Does it all*). Il faut distinguer la structuration automatique réalisée grâce aux programmes automatiques mentionnés plus haut et l'annotation qualitative qui peut être réalisée à la main ou bien automatiquement à l'aide d'outils de TAL plus complexes. En effet, la structuration automatique a été réalisée mais notons que ce niveau d'annotation suppose un regard de spécialiste et ne peut être réalisé qu'en collaboration étroite avec des historiens initiés aux questions similaires à celles traitées par le projet *TIME-US*.

Protocole de contrôle qualité sur les données

19. Une fois l'ensemble des règles de structuration générale et d'annotation sémantique élaborées et testées sur des fichiers au cas par cas, la conformité des fichiers obtenus en bout de chaîne avec les règles définies grâce à une grammaire devait être systématiquement vérifiée. Afin de réaliser cette grammaire (Rahtz et Burnard 2013), l'équipe du projet a choisi de suivre la recommandation du TEI Council pour l'utilisation de plusieurs schémas de validation, ici la structure générale et l'annotation sémantique, tout en ayant une documentation dans un seul document : ODD[29]. Bien qu'un ODD puisse contenir un très grand nombre de personnalisations de la TEI, l'hétérogénéité même du corpus a poussé à écrire deux ODD distincts en fonction de la nature de la source traitée. En effet, un travail similaire a été mené sur le corpus des comptes rendus des séances des prud'hommes publiés dans la presse ouvrière lyonnaise, sans parvenir à une granularité aussi fine étant donné la très grande variété des structures adoptées par les différents journaux. Pour les comptes rendus lyonnais comme pour les

29. Grammaire spécialisée utilisée pour définir des schémas de validation, des ensembles d'éléments TEI, d'autres éléments XML non compris dans la TEI et une documentation des usages spécifiques à un projet. Le fichier ODD contient ainsi des listes d'éléments autorisés, leur contexte d'utilisation, des attributs autorisés et leurs valeurs ainsi que la définition des éléments avec des exemples.

minutes des prud'hommes de Paris toutefois, la structure logique a été formalisée par un ODD en essayant chaque fois de trouver des dénominateurs communs. Si deux ODD distincts sont nécessaires pour la structuration, un schéma commun a été établi pour les annotations fines propres au projet. Chaque historien peut ainsi trouver les informations qui l'intéressent, liées aux objectifs de *TIME-US*, dans un cadre XML TEI aussi régulier que possible. Une fois ces différents choix établis, un ODD *chaining* a été mis en place afin de lier les différents ODD[30].

Description du jeu de données

Choisir un entrepôt et ordonner les données

20. Comme le rappellent Michaël E. Sinatra et Marcello Vitali-Rosati (2014) dans *Pratiques de l'édition numérique*, la place d'Internet dans le développement des humanités numériques a entraîné la modification de l'ensemble des pratiques de la communauté savante. En effet, il ne s'agit plus seulement d'une question technique, mais d'une question de structuration et d'organisation du savoir en général. Afin de contribuer à celles-ci, le dépôt des données dans un entrepôt dédié est fortement recommandé. Il est demandé aux projets financés par les infrastructures de recherche comme l'ANR de déposer les données produites dans le cadre du projet dans un entrepôt ouvert afin de faciliter leur réutilisation ultérieure. *TIME-US* a choisi de déposer l'ensemble des données du projet dans Zenodo.

21. Service du CERN (Organisation européenne pour la recherche nucléaire) financé par la Commission européenne, l'entrepôt Zenodo est destiné au partage des données issues de toutes les communautés scientifiques. L'entrepôt est ouvert à tous, quelle que soit l'institution ou la source de financement. Selon les préconisations de la science ouverte, tout document produit dans le cadre de la recherche, même s'il n'a pas servi à une publication, est nécessaire à la compréhension du processus de travail scientifique. L'intérêt du recours à cet entrepôt réside également dans la facilité d'intégration du dépôt dans les rapports de recherche. Contrairement à d'autres entrepôts de données, l'interface web de Zenodo permet de télécharger facilement le contenu sans passer par une API. Choisir Zenodo pour publier les données d'un projet de recherche permet de s'assurer de les publier en accord avec les principes du *FAIR data*[31] et des valeurs de la science ouverte. Un ensemble de métadonnées obligatoires est associé à chaque dépôt ainsi qu'un identifiant pérenne permettant de citer la ressource facilement et de manière unique[32]. Zenodo accepte tous les types de formats de fichiers pour le

30. Cf. Burnard, Lou. 2017. « ODD Chaining for Beginners ». *TEIC GitHub IO Repository*. 2017. http://teic.github.io/TCW/howtoChain.html.

31. Il s'agit des données qui respectent les principes de l'ouverture des données appelés FAIR (Facile à trouver, Accessible, Interopérable et Réutilisable).

32. Par l'attribution d'un DOI, pour *Digital object identifier*.

dépôt (texte, code, audio, vidéo, images…) et permet de visualiser certains formats. Il faut également mentionner que Zenodo incite également au dépôt de résultats positifs et négatifs afin de rendre la science la plus ouverte possible dans tous les domaines de recherche. Le choix des licences est flexible et des fonctions d'embargo sont aussi prévues. Les métadonnées associées aux jeux de données sont en consultation libre. Ces fichiers ainsi que leurs métadonnées sont copiés par le CERN afin d'assurer une sauvegarde de sécurité.

22. Les données liées à ce *data paper* ont été classées dans un dossier dédié aux prud'hommes de Paris au sein du dossier principal « Codes et Applications ». L'ensemble de la structuration du dépôt est détaillé dans un fichier à la racine du dépôt. Ainsi « Textes_extraction_escriptorium » contient les textes (au format TXT) obtenus immédiatement après avoir été transcrits automatiquement dans eScriptorium, « Schema_validation_ODD » contient les schémas de l'ODD *chaining*[33] et « Structuration_automatisee » contient le programme (dans un fichier python) ainsi que le Jupyter Notebook associé (au format IPYBN) pour automatiser la structuration et l'annotation des sources.

Rendre reproductible le code

23. Les principes de la science ouverte invitent à rendre les données de la recherche et le code disponibles, ce qui était déjà le cas du projet grâce à la plateforme Gitlab de l'Inria, mais nous souhaitions également rejoindre le mouvement de reproductibilité de la science. En effet, dans la lignée du colloque #dhnord2021 sur les nouvelles formes d'écriture de la recherche, les mérites de l'*executable paper* ont été vantés notamment pour des questions pédagogiques. C'est pourquoi le même programme est disponible sous la forme d'un fichier python exécutable mais aussi sous celle d'un *notebook* Jupyter. La lecture de celui-ci peut se faire dans MyBinder[34]. Les nombreux commentaires permettent de documenter le code pour des utilisateurs non familiers de la syntaxe python et désireux de reproduire l'expérience sur leur propre corpus.

Réutiliser les données de *TIME-US* sur le conseil des prud'hommes de Paris

24. Le projet *TIME-US* et les données collectées sur le Conseil des tissus parisien ne sont pas concernés par le Règlement général sur la protection des données (RGPD) et les questions éthiques qui peuvent lui être liées.

33. Avec un fichier au format RNG et un fichier au format ODD pour l'ODD. Il est accompagné d'un fichier au format HTML pour la documentation et d'un fichier XML pour les besoins de la compilation de l'ODD *chaining*.

34. Cf. https://mybinder.readthedocs.io/en/latest/introduction.html

25. Il ne semble *a priori* pas y avoir de difficulté à la réutilisation de ces données notamment grâce à leur format de publication. Les fichiers en texte sont faciles à transformer et peuvent être réutilisés facilement car leur lecture ne dépend d'aucun éditeur de texte spécifique. Il en va de même pour les fichiers en python et Jupyter Notebook. Les fichiers de contrôle du jeu de données, *ODD chaining*, sont rédigés dans une syntaxe XML, interopérable et dans le respect des standards du TEI Council. Une documentation complète en HTML du schéma est attachée et le rend donc facilement compréhensible et réutilisable. Ainsi, l'ensemble des données publiées peut être lu ou réutilisé pour être modifié. L'édition numérique de ces textes peut alimenter des études sur les relations sociales dans le monde du travail, le langage des conflits sociaux ou de la réglementation du travail, mais aussi d'autres projets de recherche portant sur l'histoire urbaine ou l'histoire des techniques au XIX^e^ siècle. Plus largement, la publication de ces données permet de rendre accessible une partie des archives d'une institution faisant régulièrement la une de la presse et des médias mais dont le passé est méconnu par le grand public.

Phœbus e-Balzac : édition numérique exhaustive d'un monument littéraire

Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

Le contexte de recherche

1. « Avant de concevoir *La Comédie humaine* et de se battre avec l'état civil, Balzac s'est battu avec le roman de son temps. [...] La création n'est pas le prix d'une victoire du romancier sur la vie, mais sur le monde de l'écrit dont il est habité », écrivait André Malraux (1977, 155). Cette affirmation trouve sa réalisation quelques années plus tard grâce à l'avènement du numérique et du projet e-Balzac coordonné par Andrea Del Lungo[1].

2. Le site ebalzac.com, ouvert en avril 2017, propose une édition électronique de *La Comédie humaine* d'Honoré de Balzac en libre accès et dans une version inédite en ligne, ainsi que des outils d'interrogation textuelle, comme un moteur de recherche lexicale ou un comparateur de versions. L'objectif est de valoriser le patrimoine écrit français en employant des méthodes parmi les plus innovantes dans le domaine des humanités numériques. Ce site constitue une première réalisation du projet *Phœbus* (Projet d'hypertexte de l'œuvre de Balzac reposant sur l'utilisation de similarités), financé par l'ANR (Agence nationale de la recherche) pour la période 2015-2019[2].

3. La partie principale de ce projet éditorial consiste en la numérisation d'une quantité importante d'œuvres. *La Comédie humaine* est un monument littéraire composé de 95 textes, dont le projet e-Balzac ambitionne la mise en ligne, dans une version philologiquement exacte, des différents états imprimés publiés du vivant de l'auteur. Cependant, la spécificité du corpus balzacien repose sur la multiplication par l'auteur des supports de publication (livres, volumes collectifs, feuilletons) et la réutilisation de ses textes antérieurs. La numérisation de différentes versions est donc un défi de taille, tant pour l'établissement d'une chaîne de traitement efficace que pour la mise en œuvre du contrôle de la qualité des données. La formation numérique des acteurs engagés au sein du projet au fil des années est variable et nécessite souvent le recours à des outils intuitifs et faciles à prendre en

1. Voir le chapitre d'Andrea Del Lungo et Karolina Suchecka dans cet ouvrage : « Le projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles ». Pour consulter les données mobilisées dans le chapitre, voir https://hnso-corpus.nakala.fr/.

2. Ce projet est porté par les équipes CELLF (Centre d'études de la langue et de la littérature françaises) et LIP6 (Laboratoire d'informatique de Paris 6) de Sorbonne Université et par l'équipe ALITHILA (Analyses littéraires et histoire de la langue) de l'université de Lille.

main, que ce soit pour l'établissement du texte ou pour son exportation dans les formats HTML ou EPUB.

Le protocole de production des données

4. Les données conçues au sein du projet peuvent être divisées en trois sous-parties :

1. Le corpus principal constitué des différents états imprimés des œuvres de Balzac publiés du vivant de l'auteur (la dernière édition retenue étant celle dite « Furne corrigé »)
2. Le corpus comparatif qui regroupe les fichiers combinant deux versions d'un texte
3. Le corpus secondaire des auteurs contemporains ou antérieurs à Balzac

5. Pour chaque sous-ensemble, ainsi qu'au sein de ces derniers, la méthode d'acquisition des données varie en fonction de l'accessibilité des œuvres. Tandis que, pour le corpus principal, la totalité de la chaîne du traitement, à commencer par la numérisation du fac-similé, a été prise en charge au sein du projet, une partie des œuvres composant le corpus secondaire a pu être numérisée semi-automatiquement à partir des formats déjà structurés, comme le format adaptable Daisy DTBook, disponible sur Gallica pour certaines œuvres les plus connues, et le format EPUB mis à disposition par des bibliothèques numériques ou d'autres projets de recherche, comme le projet ANR *Chapitres*[3]. Quant au corpus comparatif, il a été établi automatiquement à partir des documents composant le corpus principal.

6. Les attentes concernant la qualité des données étaient aussi variables. Afin d'atteindre l'exactitude philologique de chaque texte pour le corpus principal, leurs établissements ont été très rigoureusement suivis, tandis que quelques erreurs d'océrisation du corpus secondaire ont été acceptées, dans la mesure où ce dernier est destiné à l'exploitation avec le logiciel de détection automatique des réutilisations textuelles TextPAIR (Del Lungo et Suchecka 2022) et non pas à la mise en ligne au sein de l'édition numérique. Nous décrivons ci-dessous la chaîne du traitement mise en place pour le corpus principal, dont l'édition des sources est désormais disponible sur l'entrepôt Nakala[4].

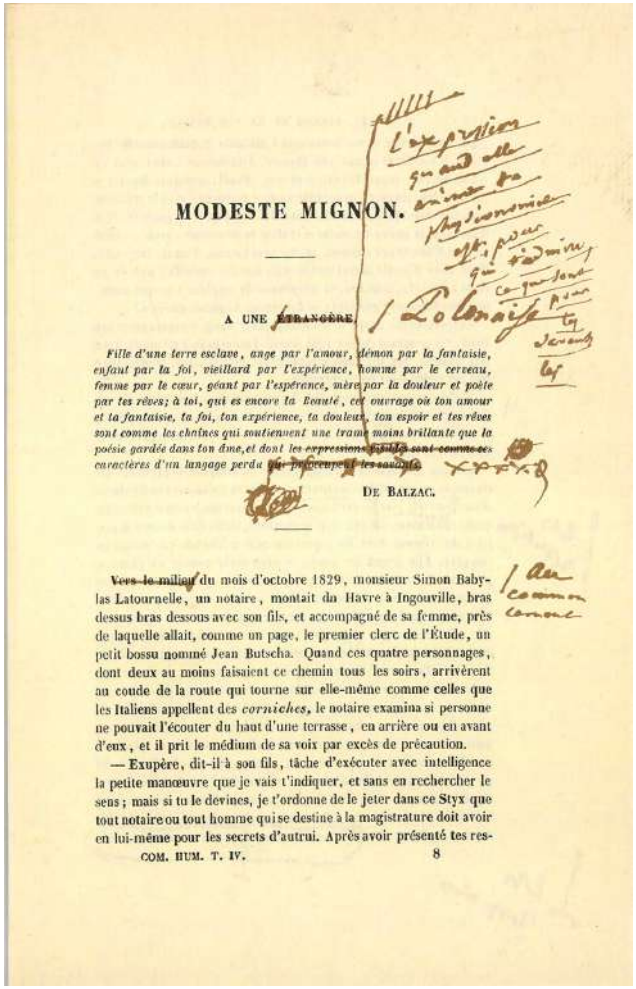
L'établissement du texte

7. La version de référence de *La Comédie humaine*, considérée comme le dernier état de l'œuvre conforme à la volonté de l'auteur, est celle appelée « Furne corrigé » (ci-après : FC). Son statut est particulier puisqu'elle intègre les corrections apportées par Balzac sur son

3. Pratiques et poétiques du chapitre du 19e au 21e siècle : génétique, rhétorique de la lecture et transmédialité – CHAPITRES, Aude Leblond (dir.), THALIM (Théorie et histoire des arts et des littératures de la modernité), Sorbonne Université, 2015-2018.

4. Cf. https://nakala.fr/ et « § La description du jeu de données ».

exemplaire personnel de la dernière édition imprimée de son vivant (figure 1), celle du Furne (ci-après : F).



MODESTE MIGNON.

A UNE ÉTRANGÈRE

Fille d'une terre esclave, ange par l'amour, démon par la fantaisie, enfant par la foi, vieillard par l'expérience, homme par le cerveau, femme par le cœur, géant par l'espérance, mère par la douleur et poète par tes rêves; à toi, qui es encore la Beauté, cet ouvrage où ton amour et ta fantaisie, ta foi, ton expérience, ta douleur, ton espoir et tes rêves sont comme les chaînes qui soutiennent une trame moins brillante que la poésie gardée dans ton âme, et dont les expressions fugitives sont comme ces caractères d'un langage perdu que préoccupent les savants.

DE BALZAC.

Vers le milieu du mois d'octobre 1829, monsieur Simon Babylas Latournelle, un notaire, montait du Havre à Ingouville, bras dessus bras dessous avec son fils, et accompagné de sa femme, près de laquelle allait, comme un page, le premier clerc de l'Étude, un petit bossu nommé Jean Butscha. Quand ces quatre personnages, dont deux au moins faisaient ce chemin tous les soirs, arrivèrent au coude de la route qui tourne sur elle-même comme celles que les Italiens appellent des *corniches*, le notaire examina si personne ne pouvait l'écouter du haut d'une terrasse, en arrière ou en avant d'eux, et il prit le médium de sa voix par excès de précaution.

— Exupère, dit-il à son fils, tâche d'exécuter avec intelligence la petite manœuvre que je vais t'indiquer, et sans en rechercher le sens; mais si tu le devines, je t'ordonne de le jeter dans ce Styx que tout notaire ou tout homme qui se destine à la magistrature doit avoir en lui-même pour les secrets d'autrui. Après avoir présenté tes res-

COM. HUM. T. IV. 8

Figure 1. Exemple d'une page de l'édition Furne

Annotation manuelle de Balzac

Source : https://www.ebalzac.com/romans/05-modeste-mignon/furne-corrige/scans/p113.jpg Licence CC-BY-NC-ND-4.0

8. Cette spécificité pose des problèmes quant à l'emploi des logiciels d'océrisation. Les annotations manuelles, situées généralement en marge et éloignées des passages qu'elles modifient, sont difficilement lisibles, même pour un œil humain. Les ratures et les signes de correction orthotypographique, quant à elles, empêchent la bonne reconnaissance de caractères imprimés. Le travail d'océrisation a donc débuté par la version non annotée de l'édition F, disponible sur Gallica tant en format image qu'en format texte[5]. Alors que l'édition numérique de la version FC reste inédite avant la création du projet e-Balzac, la F est très facilement accessible en ligne, que ce soit au sein des éditions numériques dédiées à Balzac (notamment celle réalisée par le Groupe international de recherches balzaciennes, la mairie de Paris et l'université de Chicago en 2004[6]), sur Wikisource[7], voire au sein des répertoires mettant à disposition des œuvres du domaine libre en format EPUB[8].

9. La disponibilité des œuvres en format texte permet d'accélérer la chaîne du traitement de manière importante. Cependant, elle présente un risque lié à la qualité des données mises à disposition. Pour le texte disponible sur Gallica, par exemple, le taux d'erreur d'océrisation

5. Cf. catalogue.bnf.fr/ark:/12148/cb30051006q
6. Cf. https://www.maisondebalzac.paris.fr/vocabulaire/furne/protocole.htm
7. Cf. https://fr.wikisource.org/wiki/La_Com%C3%A9die_humaine. L'édition citée est celle d'Alexandre Houssiaux, publiée de manière posthume, mais sans reprendre les corrections manuscrites de Balzac. Le texte est donc identique à celui de la version Furne (sauf erreurs d'édition).
8. Cf. par exemple https://www.ebooksgratuits.com/ebooks.php.

demeure important, alors que celui de Wikisource risque de contenir de nombreuses erreurs d'édition[9]. La comparaison de deux versions du texte disponibles en ligne (figure 2) avec le logiciel MEDITE[10] permet l'établissement d'un texte de qualité beaucoup plus satisfaisante.

Figure 2. Interface de comparaison avec le logiciel MEDITE
Version texte de Gallica à gauche, version texte de Wikisource à droite
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

10. MEDITE est un outil de comparaison des versions d'une œuvre qui s'appuie, entre autres, sur l'algorithme de l'alignement par fragments grâce à la détection des homologies (une méthode utilisée initialement pour l'alignement des macromolécules). Les blocs communs sont analysés et les différentes variantes sont signalées grâce à des codes de couleur. Les remplacements sont marqués en bleu, les suppressions en rouge, les insertions en vert et les déplacements en gris. Quelques modifications des paramètres initiaux sont possibles. Par défaut, le traitement est sensible à la casse, aux séparateurs ainsi qu'aux signes diacritiques. La longueur initiale des blocs communs est de cinq caractères. Pour que deux variantes soient considérées comme un remplacement, le ratio de la longueur des deux chaînes repérées doit être supérieur ou égal à 50 % (« abcd » est remplacé par « efgh », mais « a » est supprimé et « efgh » est inséré). Enfin, dans le cas de fortes densités des blocs communs et des variantes, les premiers sont insérés dans la variante si la différence de leur longueur par rapport à la longueur des variantes est supérieure ou égale à 50 %. Pour nos traitements, nous modifions uniquement le ratio des remplacements à 1 %, en considérant que les suppressions et les insertions sont des blocs qui n'ont pas leurs homologues dans l'autre texte. Les autres paramètres gardent les valeurs par défaut.

11. Dans l'exemple présenté en figure 3, nous observons une série de remplacements dus à la mauvaise reconnaissance du texte de la version Gallica (*armis/armés*, *coisi/choisit*, *invetions/inventions*, etc.). Le logiciel permet également de constater plusieurs erreurs typographiques de cette version, comme l'omission des espaces avant

9. Celles-ci sont accumulées par les reprises successives de la version Furne, à commencer par l'édition Houssiaux, jusqu'à celle des utilisateurs engagés dans la mise en ligne du texte.

10. Cf. http://obvil.lip6.fr/medite/ et (Ganascia et Bourdaillet 2006; Ganascia 2011; Ganascia, Glaudes, et Lungo 2014).

les ponctuations doubles (*aigre:*) ou l'ajout de celles-ci avant les ponctuations simples (*pieu ,*). Simultanément, nous comprenons que la version Wikisource contient quelques modernisations des graphies, comme c'est le cas des *falottes* des versions F et FC remplacées par *falotes* chez Houssieux. Ainsi, l'emploi du logiciel rend le travail du contrôle moins laborieux, notamment grâce aux codes couleur qui soutiennent la vigilance du correcteur, ce qui est particulièrement utile par exemple pour la correction des mots outils à graphies proches fréquemment confondus par les logiciels d'océrisation, comme *on* et *ou*.

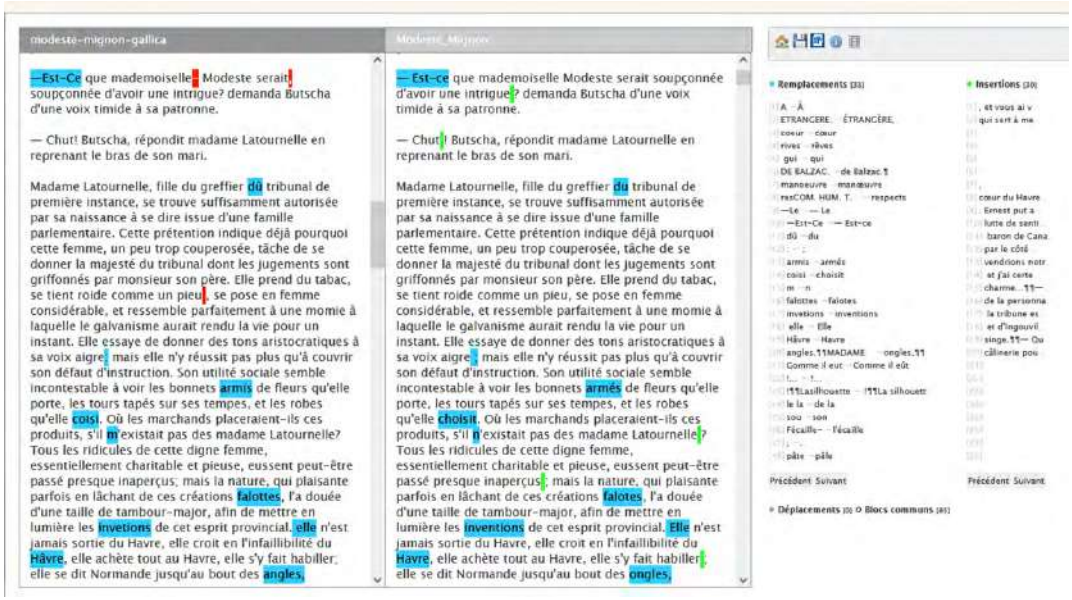


Figure 3. Extrait des résultats de MEDITE

Comparaison entre la version Gallica et celle de Wikisource de *Modeste Mignon*

Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

Le stylage dans un logiciel de traitement textuel

12. Après l'établissement d'une version satisfaisante du texte de l'édition F, nous suivons la chaîne de traitement Odette/Teinte (Glorieux 2015) au sein du laboratoire OBVIL (Observatoire de la vie littéraire[11]). Appuyée sur une structure restreinte issue du standard XML-TEI, Teinte, une feuille de style dédiée[12] a été proposée pour permettre de travailler avec les logiciels de traitement de texte. Même si les outils WYSIWYG[13] sont peu performants pour les structures complexes, ils ont l'avantage de faciliter l'intégration dans la chaîne éditoriale des acteurs externes – stagiaires, doctorants, vacataires ou chercheurs expérimentés –, sans qu'ils soient initiés au langage XML.

13. Optimisée pour les textes romanesques, cette méthode compose avec les fonctionnalités initiales des outils (mise en italique, style du paragraphe par défaut) et les noms de styles particuliers (mis entre chevrons) permettant l'annotation sémantique d'un bon nombre d'éléments textuels (citations, illustrations, correspondance, poèmes, etc.). Le développement est accompagné d'un guide

11. Actuellement ObTIC (Observatoire des textes, des idées et des corpus).
12. Pour celle utilisée au sein du projet e-Balzac, cf. https://sharedocs.huma-num.fr/wl/?id=RuGwKHl5KJPgDWSjXg8tJNBCOCqvIsFc.
13. *What You See Is What You Get*, acronyme qui renvoie aux éditeurs visuels de texte.

d'emploi détaillé[14]. Le nombre restreint des éléments limite l'hétérogénéité des annotations choisies par les éditeurs impliqués dans le processus, dont l'interprétation de la structure textuelle peut varier.

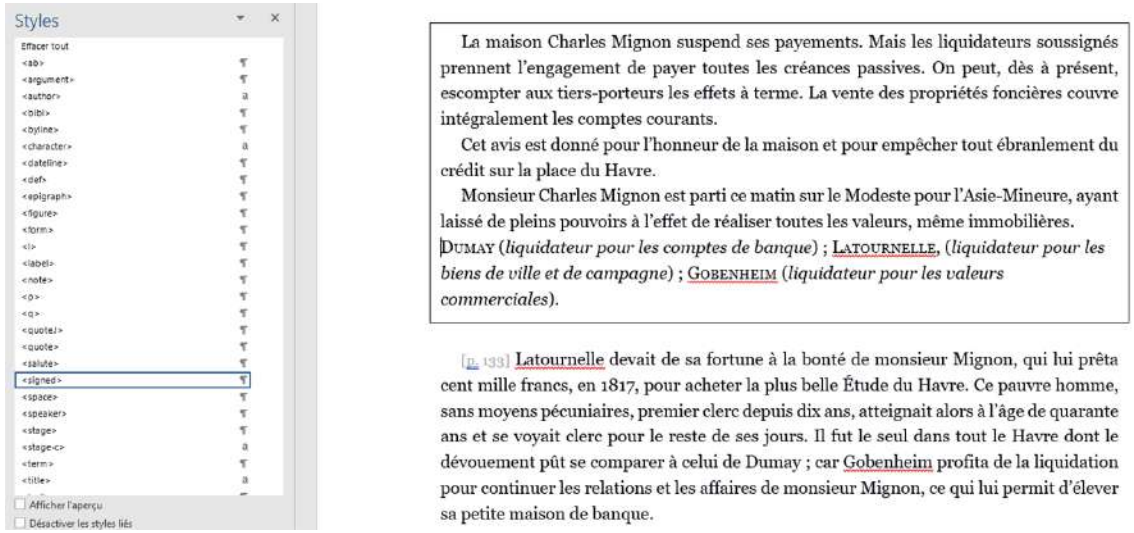


Figure 4. Application des styles. Exemple des éléments <q> et <signed> dans *Modeste Mignon*
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

14. En figure 4[15], les bordures entourent le texte d'un article du *Courrier du Havre*. L'enchâssement d'autres textes dans le récit, notamment d'articles et de lettres, est très fréquent dans *La Comédie humaine*. Différenciés du corps du texte par une mise en page spécifique[16], ces éléments pourraient être considérés comme des citations, et donc à signaler, en suivant le standard XML-TEI Teinte, par <quote>. Or, du point de vue philologique, une citation est définie comme une partie de texte externe à l'œuvre, alors que, chez Balzac, il s'agit souvent d'extraits fictionnels, comme des lettres écrites par des personnages ou encore des articles fictifs tirés de journaux bien réels[17]. Nous choisissons donc de les signaler avec l'élément <q>, qui, dans le standard TEI : « contient un fragment qui est marqué (visiblement) comme étant d'une manière ou d'une autre différent du texte environnant, pour diverses raisons telles que, par exemple, un discours direct ou une pensée, des termes techniques ou du jargon, une mise à distance par rapport à l'auteur, des citations empruntées et des passages qui sont mentionnés, mais non employés[18]. »

14. Cf. https://obvil.github.io/Teinte/teinte.html

15. Pour consulter le document complet en format DOCX, cf. https://sharedocs.huma-num.fr/wl/?id=BwvWJnbdnROWwND1xGBP42YCzoCJnHxK.

16. Dans l'exemple à la figure 4, le texte de l'article est entouré des guillemets et séparé du corps du texte par deux lignes horizontales, cf. https://gallica.bnf.fr/ark:/12148/bpt6k6116517k/f145.highres.

17. *Le Courrier du Havre* a été publié quotidiennement de 1839 à 1906, cf. https://data.bnf.fr/fr/32751243/courrier_du_havre/.

15. Pour répondre aux besoins spécifiques du projet e-Balzac, le schéma Teinte a modifié cette définition en précisant que « le contenu de <q> ne peut pas être attribué à une origine extérieure au texte (origine fictionnelle ou non

18. Cf. https://tei-c.org/release/doc/tei-p5-doc/fr/html/ref-q.html

identifiable). Exemples : récit enchâssé, lettre insérée, article de presse, acte notarié, poème, traité, axiomes[19]... »
La structure interne de l'élément textuel codé par <q> est également balisée à l'aide de styles dédiés. En figure 4, par exemple, les signatures des trois liquidateurs sont marquées avec le style du paragraphe <signed> et leurs noms sont mis en petites capitales.

16. Le texte corrigé et stylé de l'édition F est ensuite repris pour l'établissement de la version FC en intégrant manuellement les annotations de Balzac. Cette méthode, certes chronophage et minutieuse, garantit l'exactitude philologique de la transcription et permet de corriger d'éventuelles erreurs des autres éditions scientifiques appuyées sur ce texte de référence et disponibles notamment sur le marché du livre imprimé. L'éditeur travaille sur un texte auquel les styles ont déjà été appliqués, ce qui lui permet de focaliser toute son attention sur les annotations de Balzac. Cela implique également que l'expert de la main de l'auteur ne doit pas simultanément maîtriser les techniques de l'édition numérique. Enfin, une relecture finale, appuyée par la comparaison avec MEDITE est effectuée avant la transformation vers le format XML[20].

La transformation automatique en format XML-TEI Teinte avec Odette

17. Le texte stylé dans un traitement de texte et enregistré en format ODT est ensuite transformé en format XML-TEI grâce au logiciel Odette[21] (figure 5). Une fois encore, la mise en place d'un outil intuitif et disponible en ligne augmente l'autonomie des éditeurs sans demander des compétences informatiques particulières.

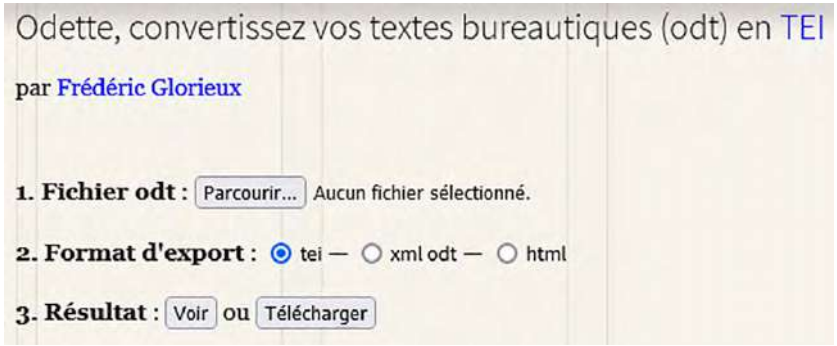


Figure 5. Interface Web de l'outil Odette
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

18. Grâce à Odette, trois étapes simples (chargement du fichier stylé, choix du format d'export, puis téléchargement du résultat en format choisi) suffisent pour transformer rapidement un document ODT vers le format XML-TEI Teinte. Pour l'extrait textuel présenté en

19. Cf. https://obvil.github.io/Teinte/teinte.html#el_q

20. Pour l'établissement des versions antérieures, l'éditeur peut s'appuyer sur le texte d'une édition déjà établie et, par exemple, la comparer avec l'océrisation de la version en cours de préparation.

21. Cf. http://obvil.lip6.fr/Odette/

figure 4, par exemple, nous obtenons la structuration XML suivante :

```
<q type="press">
  <p>La maison Charles Mignon suspend ses payements. [...]</p>
  <p>Cet avis est donné pour l'honneur de la maison [...].
  </p>
  <p>Monsieur Charles Mignon est parti ce matin [...].
  </p>
  <signed>
    <hi rend="sc">Dumay</hi> (<hi rend="i">liquidateur pour les
    comptes de banque</hi>) ;
    <hi rend="sc">Latournelle</hi>, (<hi rend="i">liquidateur pour
    les biens de ville et de campagne</hi>) ;
     <hi rend="sc">Gobenheim</hi>, (<hi rend="i">liquidateur
     pour les valeurs commerciales</hi>).
  </signed>
</q>
<p>
  <pb n="133" xml:id="p133"/>
 Latournelle devait de sa fortune à la bonté de monsieur Mignon,
 [...].
</p>
```

19. Pour les utilisateurs plus expérimentés, le code source du logiciel est disponible en ligne[22]. Le logiciel est conçu principalement à l'aide des feuilles de style XSL et scripts PHP. Il est possible, en l'utilisant localement, d'appliquer la transformation massivement à l'aide d'une ligne de commande exécutée à partir du dossier avec les ressources Odette :

20. user@user:~/odette$ php Odt2tei.php "*.odt" XML/?

21. La variable "*.odt" indique que tous les fichiers en format ODT présents dans le dossier doivent être traités. En revanche, il n'est pas permis de les regrouper dans un sous-dossier : la variable "sous-dossier/*.odt" ne sera pas reconnue. Les résultats peuvent être sauvegardés dans un sous-dossier spécifique (XML/ dans notre exemple).

22. La qualité de la transformation est très satisfaisante (à condition, bien sûr, que le stylage soit fait correctement et en respect du schéma Teinte). Sporadiquement, des balises auto-fermantes <anchor/> peuvent apparaître dans le XML. Il est également conseillé de revoir les éléments de l'emphase avec la requête xPath //hi[@rend='i'] et //emph, car la mise en italiques des longs fragments textuels produit parfois l'accumulation inutile des balises. Le <teiHeader> doit être modifié et complété en fonction des usages de chaque projet.

La transformation vers des formats éditoriaux (HTML et EPUB)

23. Il est également possible, dans la continuité de la chaîne du traitement Odette/Teinte, de générer des fichiers HTML et EPUB à partir du document XML ainsi obtenu. La procédure devient, à ce point, plus complexe, notamment si l'on souhaite utiliser le développement sur un

22. Cf. https://github.com/oeuvres/odette

serveur Web[23]. Toutefois, pour une utilisation basique et locale, les compétences de base en XSLT sont suffisantes.

24. À partir du dossier des sources, le document XML traité doit être transformé à l'aide de la feuille de style « tei2html.xsl ». Pour ce faire, il est possible d'utiliser une ligne de commande, en exécutant, à partir du dossier Teinte :

25.
```
user@user:~/teinte$ xsltproc tei2html.xsl texte.xml > ../HTML/texte.html
```

26. Pour que le fichier HTML ainsi généré puisse être mis en forme (à l'aide d'une feuille de style CSS « tei2html.css »), il doit être enregistré dans un dossier frère de celui contenant les ressources Teinte. La visualisation proposée est basique, mais offre tout de même quelques fonctionnalités facilitant la lecture dans l'environnement numérique (figure 6). Par exemple, le clic sur le numéro de page affiché sur la marge de gauche permet d'afficher le fac-similé correspondant[24], la génération de la table de matières accélère la navigation dans le texte, et la mise en forme des éléments balisés facilite la révision et le contrôle qualité du document.

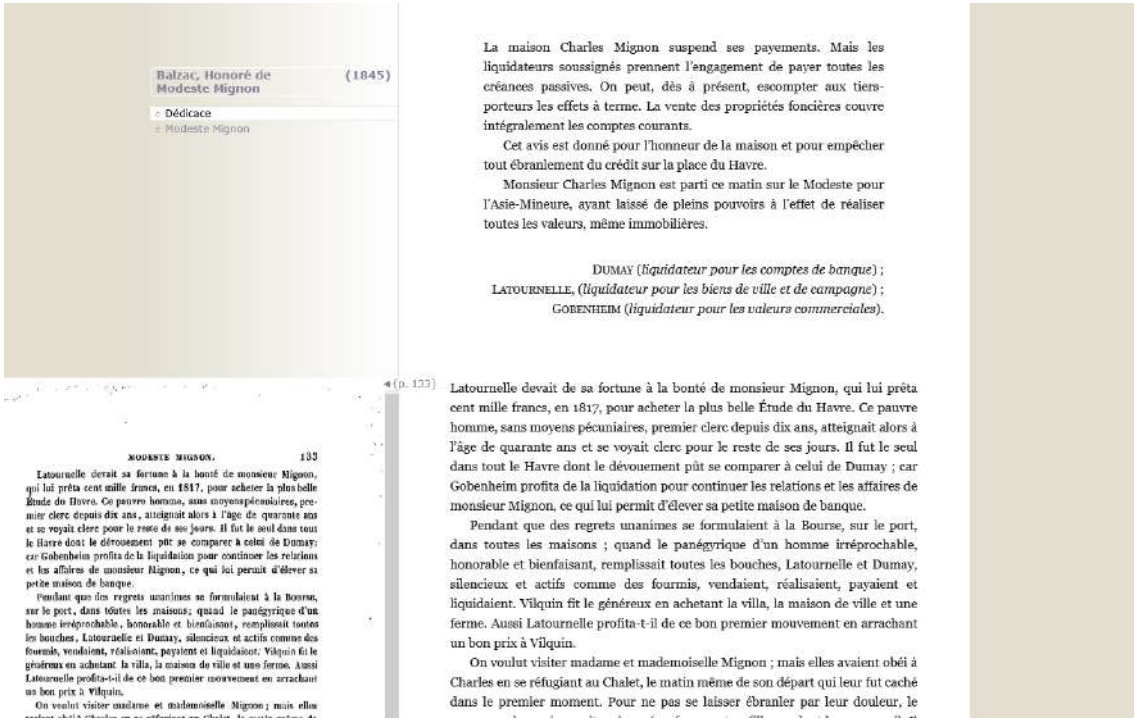


Figure 6. Visualisation HTML avec Teinte. Exemple d'affichage pour *Modeste Mignon*
Crédit : Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

23. Dans ce cas, nous renvoyons les lecteurs vers la documentation détaillée disponible sur https://github.com/oeuvres/teinte. Pour l'utilisation avec Omeka, cf. https://github.com/oeuvres/Bookmeka.

24. À condition que le lien vers l'image soit renseigné en tant que valeur de l'attribut @facs de l'élément <pb>. Dans l'exemple présenté à la figure 6, la page 133 est déclarée ainsi dans le XML final : <pb n="133" xml:id="p133" facs="https://gallica.bnf.fr/ark:/12148/bpt6k6116517k/f146.highres"/>. Nous ajoutons les attributs @facs automatiquement après la génération du XML, cf. https://nakala.fr/10.34847/nkl.d1066aod.

La description du jeu de données

27. Le jeu de données décrit dans cet article a été déposé dans l'entrepôt de données Nakala suivant une modélisation des métadonnées préparée dans le cadre du dépôt. Avant d'aller plus loin dans la description du jeu de données, précisons qu'il s'agit ici d'informations sous format numérique, sorties de leur contexte de recherche, qu'il est nécessaire de traiter, structurer et présenter d'une certaine manière.

Le choix de l'entrepôt

28. Nakala est un service de l'IR* Huma-Num[25] permettant aux chercheurs, enseignants-chercheurs ou équipes de recherche de partager, publier et valoriser tous types de données numériques documentées (fichiers texte, sons, images, vidéos, objets 3D, etc.) dans un entrepôt sécurisé afin de les publier en accord avec les principes du *FAIR data*[26] et des valeurs de la science ouverte. Choisir Nakala comme entrepôt de données permet de s'inscrire dans la logique du Web de données ouvertes (*Linked Open Data*), qui rend possible la connexion à d'autres entrepôts existants, et prolonge le travail sur l'intertextualité autour de l'œuvre de Balzac.

29. Une fois les données publiées sur Nakala, il est possible de les récupérer, moissonner[27], et exposer ailleurs grâce au protocole documentaire OAI-PMH[28], au modèle RDF[29] ou à une API REST[30]. Par ailleurs, Nakala fait partie d'un dispositif cohérent de services mis en place par Huma-Num suivant la chaîne de traitement des données. L'accès, le signalement, la conservation et l'archivage à long terme des données de la recherche en SHS sont facilités par les outils développés par Huma-Num.

30. Nakala est destiné au stockage de données stabilisées (décrites et complètes) et est utilisé par les porteurs de projets une fois le processus de collecte des données terminé. Afin de réaliser une collecte optimale, les projets sont encouragés à :

- mettre en sécurité les données sur un outil de stockage externe (disques durs...)
- ordonner et organiser les données

25. L'Infrastructure de Recherche (IR*) des humanités numériques (anciennement TGIR, Très Grande Infrastructure de Recherche), cf. https://www.huma-num.fr/.

26. Il s'agit des données qui respectent les principes de l'ouverture des données appelés *FAIR* (*Findability*, *Accessibility*, *Interoperability* et *Reusability*).

27. Cela veut dire que les données peuvent être récoltées et incluses dans une base de données regroupant les documents aux références bibliographiques similaires. Seulement les données encodées avec les mêmes procédés techniques peuvent être moissonnées.

28. *Open Archives Initiative – Protocol for Metadata Harvesting*, protocole pour la collecte des métadonnées de l'initiative pour les archives ouvertes, est un dispositif permettant l'échange des métadonnées entre plusieurs institutions qui donnent accès aux documents numériques, cf. http://www.openarchives.org/OAI/openarchives-protocol.html. Il est utilisé, par exemple, par la Bibliothèque nationale de France, cf. https://www.bnf.fr/fr/les-entrepots-oai-de-la-bnf.

29. *Resource Description Framework* est un modèle descriptif des données Web et des métadonnées attachées.

30. Interface de Programmation d'Application (*Representational State Transfer*) est un style d'architecture logicielle.

- consigner toutes les informations disponibles sur les données
- mettre en place un plan de nommage harmonisé[31] des fichiers

31. Nakala accepte tous les types de formats de fichiers pour le dépôt et permet d'en visualiser certains[32]. Les données sont respectées dans leur intégrité et ne subissent pas de modification une fois déposées. De plus, Huma-Num effectue une copie sécurisée des données et des métadonnées au sein de son infrastructure. Celles-ci sont identifiées grâce à l'attribution d'identifiants pérennes[33]. Il est ainsi possible de citer chaque donnée de manière précise et exacte grâce au DOI[34]. Par exemple, l'édition électronique du texte de *Modeste Mignon* parue chez Furne (1845) est connue sous l'identifiant 10.34847/nkl.defelc0z. Il est également possible de citer la référence complète de la donnée (Balzac 2021). Même si cette manière de référencer est encore à retravailler, notamment en ce qui concerne la reprographie des textes imprimés, l'identifiant alphanumérique assure que la citabilité de chaque donnée et chaque collection soit pérenne et aussi exacte que possible. Afin de suivre toutes les recommandations *FAIR*, le projet e-Balzac s'est également assuré que les ressources déposées ne contiennent pas des informations personnelles, dont la diffusion est encadrée par le Règlement général sur la protection des données (RGPD). Par conséquent, il n'est pas nécessaire de mettre en place des procédures de pseudonymisation[35] ou de *mash-up* (application composite) des données. Elles doivent tout de même être mises sous une licence. Celle reprise pour la publication des données se conforme au choix initial effectué lors de la création du site ebalzac.com : il s'agit de la licence Creative Commons Attribution Non Commercial No Derivatives 4.0 International[36] (CC-BY-NC-ND-4.0).

32. Outre la facilitation du respect des principes *FAIR*, d'autres fonctionnalités de Nakala visent à simplifier la gestion collective du dépôt. Tous les utilisateurs pourvus des droits d'édition peuvent déposer dans les collections et modifier les métadonnées avant la publication. Cette gestion fine des droits d'accès aux dépôts, collections et fichiers est un atout de taille pour répartir le travail entre deux personnes n'appartenant pas à la même institution et ne pouvant se voir régulièrement. Les droits sur les données sont partagés entre le déposant[37] (ROLE_OWNER) et les administrateurs ajoutés[38] (ROLE-ADMIN). La vue et l'édition de la donnée dépendent ainsi des rôles attribués par la suite, qui peuvent être modifiés par le déposant. Il est aussi possible d'ajouter des listes d'utilisateurs autorisés à avoir certains rôles.

31. Cela implique que les fichiers sont toujours nommés de la même manière, par exemple : Type_date_numéro.
32. Par exemple l'utilisation de la visionneuse d'images OpenSeadragon (cf. https://openseadragon.github.io/).
33. Depuis 2020, il s'agit du DOI. Auparavant, c'est Handle qui était plus largement utilisé.
34. Le *Digital Object Identifier* permet d'identifier les ressources de manière unique.
35. Moins forte que l'anonymisation, qui empêche totalement et de manière irréversible la possibilité d'identifier une personne, la pseudonymisation permet d'empêcher l'identification d'une personne en fonction des données présentes.
36. Cf. https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode
37. Ici : Victoria Le Fourner.
38. Ici : Karolina Suchecka.

Les métadonnées

33. La modélisation des métadonnées a été effectuée à partir des informations présentes sur chaque page d'une œuvre sur le site ebalzac.com, les informations présentes dans le <teiHeader> des fichiers XML ainsi qu'à partir des métadonnées obligatoires pour chaque dépôt dans Nakala. Chaque donnée déposée peut contenir plusieurs fichiers et est décrite par un certain nombre de métadonnées. La clarté dans l'exposition des métadonnées est fondamentale pour une réutilisation et une compréhension par autrui. Aussi, notre travail insiste sur la précision de la description de chaque donnée. Cinq métadonnées sont rendues obligatoires pour le dépôt : le type de la donnée, le titre, l'auteur, la date de création et la licence (tableau 1). Il est admis qu'un auteur soit anonyme ou qu'une date soit inconnue. Le choix du type de dépôt est restreint par les valeurs présentes au sein d'une liste déroulante. Le type « Édition de sources » convient à la majorité des données du projet. Les choix de mettre Balzac en créateur principal et de renseigner la date de l'édition originale de la source ont été basés sur celles du projet *Nénufar*, sur lequel nous nous sommes appuyés[39]. Nous avons également suivi les recommandations du consortium CAHIER pour les éditions numériques (Galleron et al. 2018) afin de savoir s'il est permis de garder la structuration des informations contenues dans le <teiHeader> ou s'il est préconisé de respecter les usages du standard Dublin Core. La multiplication des mots-clés nous a semblé importante pour un meilleur référencement dans le moteur de recherche de Nakala et pour un moissonnage extérieur.

34. Le choix de mettre dans une donnée tous les types de fichiers générés à partir du XML initial (plutôt que de créer une collection par type de fichier) est motivé par la volonté de faciliter la recherche et l'extraction ciblée des informations par un utilisateur externe. Par ailleurs, le type de fichiers présents dans les données varie en fonction de l'édition de chaque source. Un fichier image (en format PNG) est attaché uniquement aux documents de l'édition FC. Autrement, 171 fichiers HTML sont publiés, 92 EPUB, 179 XML et 93 PNG. Le fichier attendu est le fichier XML, les autres n'étant pas produits pour toutes les éditions. Le volume total des données publiées est de 193,6 Mo. Elles ont été déposées en août 2021 et publiées en septembre 2021. Toutes les données sont disponibles dans l'entrepôt Nakala du projet sous l'identifiant 10.34847/nkl.89fa9i03[40]. Enfin, un tableau qui spécifie, pour chacune des 95 œuvres, son appartenance aux collections et les identifiants DOI des différentes versions éditoriales est disponible dans la collection « e-Balzac »[41].

39. Cf. http://nenufar.huma-num.fr/presentation/

40. Cf. https://nakala.fr/u/collections/10.34847/nkl.89fa9i03

41. Cf. https://nakala.fr/10.34847/nkl.b2afnpoa

42. Dans ce cas spécifique, nous ne renseignons pas de date pour l'édition FC, qui n'a pas été publiée du vivant de Balzac, pour la différencier de celle du F (cf. « § L'établissement du texte »).

Tableau 1. Liste des métadonnées remplies pour chaque dépôt

L'astérisque (*) indique le caractère obligatoire de la métadonnée lors du dépôt.

Nom de la métadonnée	Type	Explication	Exemple (*Modeste Mignon*)
Type de dépôt* nakala:title	Liste	Type de la donnée déposée	Édition de sources
Titre* nakala:type	dcterms:box	Titre de la version déposée	*Modeste Mignon*, dans *La Comédie humaine*, *Études de mœurs*, *Scènes de la vie privée*. Furne corrigé
Auteurs* nakala:creator		Auteur de la source	Honoré de Balzac
Date de création* nakala:created		Date de création de la source. Celle-ci peut-être inconnue.	Inconnue[41]
Licence* nakala:licence	Liste	Licence attribuée sur le dépôt, en lien avec le type de donnée	Creative Commons Attribution Non Commercial No Derivatives 4.0 International
Description dcterms:description	dcterms:Box	Description de l'édition électronique reprise des indications du <teiHeader> du fichier	Cette édition électronique de *La Comédie humaine* constitue la première édition en ligne de l'œuvre de Balzac dans la version dite du « Furne corrigé », qui intègre les corrections manuscrites apportées par l'auteur sur son exemplaire personnel de la première édition de *La Comédie humaine* parue chez Furne de 1842 à 1847.
Mots-clés dcterms:subject	dcterms:Box	Liste de mots-clés destinés au référencement sur Nakala	Édition en ligne ; Édition électronique ; Balzac ; Modeste Mignon ; Comédie humaine ; e-Balzac ; …
Langues	dcterms:language	Langue pour les fichiers de données	Français
dcterms:publisher	dcterms:Box	Éditeur	e-Balzac

Nom de la métadonnée	Type	Explication	Exemple (*Modeste Mignon*)
dcterms:contributor	dcterms:Box	Éditeur(s) scientifique(s)	Directeurs du projet : Andrea Del Lungo, Jean-Gabriel Ganascia et Pierre Glaudes Éditeur : Maxime Perret Correction OCR : Dimitri Julien Établissement du texte et stylage TEI : Maxime Perret Édition XML-TEI : Amélie Canu Informatique éditoriale : Frédéric Glorieux Traitement des images : Claire Carpentier Dépôt Nakala : Victoria Le Fourner et Karolina Suchecka
dcterms:relation	dcterms:URI	Lien vers un document en relation	https://nakala.fr/10.34847/nkl.defelc0z
dcterms:abstract	dcterms:Box	Résumé	Au Havre, Modeste Mignon et sa mère attendent patiemment le retour de Charles Mignon de La Bastie, parti tenter de rétablir sa fortune après une faillite fracassante. En l'absence de son père, Modeste tombe sous le charme de la poésie de Melchior de Canalis et trouve le moyen d'initier une correspondance avec son grand homme. Elle ne se doute pas que c'est le secrétaire du poète, Ernest de La Brière, qui a emprunté le nom de Canalis et qui est le véritable interlocuteur de cet échange épistolaire.
dcterms:available	dcterms:URI	Lien vers la ressource disponible sur le site e-Balzac	https://www.ebalzac.com/edition/05-modeste-mignon/furne-corrige
dcterms:issued	dcterms:Box	Date de publication du fichier électronique	2017

L'organisation en collections

35. En choisissant Nakala comme entrepôt de données, il faut composer avec la spécificité de celui-ci : les collections n'ont pas de niveaux hiérarchiques. En revanche, il est possible de mettre les données dans autant de collections que nécessaire (une collection publique n'acceptant que des données publiées et non déposées[43]). Sans hiérarchisation, il est a priori difficile de rendre compte de la composition conceptuelle de *La Comédie humaine*. En effet, l'œuvre se subdivise en *Études* puis en *Scènes* (et parfois même en diptyques). La subtilité du projet réside également en la présence de plusieurs éditions d'un même texte, qui doivent donc être reliées de manière compréhensible. Cette hiérarchie, bien visible sur le site du projet, est pensée comme une arborescence, alors que Nakala demande de la représenter avec un graphe. Pour ce faire, nous choisissons d'explorer les champs Dublin Core (préfixés par dcterms), qui permettent de reconstituer le lien entre les collections, notamment grâce à dcterms:hasPart, dcterms:isPartOf ou encore dcterms:relation. L'utilisateur peut alors retrouver l'appartenance d'une collection à une autre ou voir le lien entre les textes[44].

36. Ainsi, il est nécessaire de mettre chaque donnée dans différentes collections afin de la restituer clairement dans *La Comédie humaine*. Par exemple, la donnée *Modeste Mignon* (10.34847/nkl.1cffmy50) appartient aux collections « e-Balzac » (10.34847/nkl.89fa9103), « Édition Furne Corrigé » (10.34847/nkl.19e543l9), « Comédie humaine » (10.34847/nkl.bb45t7np), « Études de mœurs » (10.34847/nkl.7d1b0m32) et « Scènes de la vie privée » (10.34847/nkl.da7c094z). La donnée est mise en relation avec une donnée similaire, *Modeste Mignon* de l'édition Furne (10.34847/nkl.defelc0z). Les métadonnées entre elles sont sensiblement les mêmes. Toutefois, ces données ne recouvrent pas la même réalité matérielle. Par conséquent, une description philologique et littéraire est ajoutée à chaque collection afin d'accompagner l'utilisateur au mieux dans l'exploration de la structure complexe de *La Comédie humaine* et dans les différentes éditions des œuvres. Les données sont ainsi disséminées dans 28 collections (tableau 2). La collection « e-Balzac » regroupe directement toutes les ressources produites dans le cadre du projet. Elle est destinée à être reliée avec la collection *Humanités numériques et science ouverte*.

La navigation dans l'entrepôt avec Nakala_Press

37. Une fois l'ensemble des collections publiées, il est possible de mettre en place un site Nakala_Press[45]. Ce site

43. Une donnée déposée est une donnée accessible uniquement au déposant et aux personnes ayant un accès à cette donnée. Une donnée publiée est une donnée visible par n'importe quel utilisateur de Nakala. Si la donnée est visible, cela signifie que ses métadonnées sont visibles, mais le contenu peut être masqué et mis en embargo. Les utilisateurs sont toutefois au courant de son existence.

44. Ces liens sont désormais visibles directement grâce à l'interface Nakala. Ils n'étaient pas directement apparents au moment du dépôt en août 2021.

45. Si le remplaçant de Nakalona n'a pour le moment que peu d'options de stylage et de valorisation des données, il n'est plus lié à Omeka, ce qui permet d'éviter le problème de pérennité lié à la mise à jour et à l'obsolescence des technologies.

Tableau 2. Liste des collections, leurs identifiants DOI et leurs liens pérennes

Nom de la collection	Id Nakala	Lien pérenne vers la collection
e-Balzac	10.34847/nkl.89fa9i03	https://nakala.fr/u/collections/10.34847/nkl.89fa9i03
Comédie humaine	10.34847/nkl.bb45t7np	https://nakala.fr/u/collections/10.34847/nkl.bb45t7np
[Édition] Furne	10.34847/nkl.d3dazxns	https://nakala.fr/u/collections/10.34847/nkl.d3dazxns
[Édition] Furne Corrigé	10.34847/nkl.19e543l9	https://nakala.fr/u/collections/10.34847/nkl.19e543l9
[Édition] Mame	10.34847/nkl.ff47434q	https://nakala.fr/u/collections/10.34847/nkl.ff47434q
[Édition] Béchet	10.34847/nkl.814e3gt6	https://nakala.fr/u/collections/10.34847/nkl.814e3gt6
[Édition de] La Presse	10.34847/nkl.ea325v1w	https://nakala.fr/u/collections/10.34847/nkl.ea325v1w
[Édition du] Constitutionnel	10.34847/nkl.f9182h84	https://nakala.fr/u/collections/10.34847/nkl.f9182h84
[Édition du] Pétion	10.34847/nkl.06f00638	https://nakala.fr/u/collections/10.34847/nkl.06f00638
[Édition du] Siècle	10.34847/nkl.f4f5e445	https://nakala.fr/u/collections/10.34847/nkl.f4f5e445
[Édition de l'] Union	10.34847/nkl.b5d6k760	https://nakala.fr/u/collections/10.34847/nkl.b5d6k760
[Édition du] Canel	10.34847/nkl.7dcc76uu	https://nakala.fr/u/collections/10.34847/nkl.7dcc76uu
[Édition] Chlendowski	10.34847/nkl.8babp549	https://nakala.fr/u/collections/10.34847/nkl.8babp549
[Édition de l'] Europe Littéraire	10.34847/nkl.e5579r98	https://nakala.fr/u/collections/10.34847/nkl.e5579r98
[Édition] Charpentier	10.34847/nkl.4bbczmc6	https://nakala.fr/u/collections/10.34847/nkl.4bbczmc6
Études de mœurs	10.34847/nkl.7d1b0m32	https://nakala.fr/u/collections/10.34847/nkl.7d1b0m32
Études philosophiques	10.34847/nkl.4b5e8zw2	https://nakala.fr/u/collections/10.34847/nkl.4b5e8zw2
Études analytiques	10.34847/nkl.3ff5mulk	https://nakala.fr/u/collections/10.34847/nkl.3ff5mulk
Scènes de la vie privée	10.34847/nkl.da7c094z	https://nakala.fr/u/collections/10.34847/nkl.da7c094z
Scènes de la vie de province	10.34847/nkl.2fadvhq8	https://nakala.fr/u/collections/10.34847/nkl.2fadvhq8
Scènes de la vie parisienne	10.34847/nkl.6c2ef1n9	https://nakala.fr/u/collections/10.34847/nkl.6c2ef1n9
Scènes de la vie politique	10.34847/nkl.fc77m86e	https://nakala.fr/u/collections/10.34847/nkl.fc77m86e
Scènes de la vie militaire	10.34847/nkl.cca809ji	https://nakala.fr/u/collections/10.34847/nkl.cca809ji
Scènes de la vie de campagne	10.34847/nkl.fcf53ybd	https://nakala.fr/u/collections/10.34847/nkl.fcf53ybd
Les Célibataires	10.34847/nkl.adfbt5u7	https://nakala.fr/u/collections/10.34847/nkl.adfbt5u7
Les Parisiens en province	10.34847/nkl.357cr836	https://nakala.fr/u/collections/10.34847/nkl.357cr836
Les Rivalités	10.34847/nkl.76a6ll98	https://nakala.fr/u/collections/10.34847/nkl.76a6ll98
Les Parents pauvres	10.34847/nkl.f31403ds	https://nakala.fr/u/collections/10.34847/nkl.f31403ds

n'a pas pour vocation de remplacer le site original du projet, mais de faciliter la lecture et la navigation au sein des données partagées. Dans la même veine que les expositions du *Research Data Journal for the Humanities and Social Sciences* (Brill), le but de cette présentation des données est de rendre la recherche des ressources dans l'entrepôt moins fastidieuse pour les personnes non initiées aux services et outils d'Huma-Num. Le site où les données peuvent être consultées, e-balzac.nakala.fr, a été construit en relation avec le projet *Humanités numériques et science ouverte* et a été connecté au site principal de celui-ci[46], permettant ainsi de restituer le chapitre du présent ouvrage dans son contexte. Le site principal oriente l'utilisateur vers les données des différents projets. Elles sont liées, mais ne se mélangent pas. La modélisation des données reste également propre à chaque collection et adaptée à chaque projet. Elle est renseignée à l'aide du standard Dublin Core.

38. Différents types de contenu peuvent être créés à partir de la page d'accueil (lien vers une ressource externe, liste de données, métadonnées, visualisation du contenu…). Les données contenues dans la collection et les métadonnées sont mises en forme, ce qui facilite leur lecture. En cliquant sur « Chronologie des publications », il est également possible de filtrer les données par année de création. Ainsi, un utilisateur extérieur peut facilement récupérer uniquement un échantillon chronologique du corpus qui l'intéresse. De même, il est possible de voir les différentes collections auxquelles sont liées les données. Une donnée pouvant être contenue dans plusieurs collections, il est plus aisé pour un utilisateur extérieur, par exemple de naviguer directement dans la collection « Études de mœurs » à partir de la collection « e-Balzac » depuis laquelle le site Nakala_Press est généré. Enfin, une recherche par type de licence est également possible.

39. Des pages annexes ont été ajoutées pour structurer davantage le site. Il s'agit notamment des liens vers le site du projet et vers les outils exploités dans le cadre du projet. Un moteur de recherche interne à cette collection est également disponible. Ainsi, il s'agit non seulement de rendre les données disponibles, mais aussi visibles, et surtout lisibles par des personnes extérieures au projet, voire peu familières des outils et méthodes des humanités numériques. La consultation de l'entrepôt n'est peut-être pas encore un réflexe, mais la navigation sur un site Web est une pratique généralement démocratisée.

La réutilisation des données

40. Le projet e-Balzac s'appuyant sur la production littéraire du XIX^e^ siècle, toutes les œuvres mises à disposition dans le cadre de l'édition numérique sont dans le domaine public et ne sont pas restreintes par les droits d'auteur. Cela limite de manière importante d'éventuelles questions éthiques soulevées par la collecte des données. Plus encore, alors que des éditions numériques des sources ont déjà suscité des controverses liées au travail édito-

46. Cf. https://hnso.nakala.fr/

rial, critique ou scientifique[47], le FC a été établi pour la première fois à la fin du XIXe siècle[48]. Par ailleurs, le projet e-Balzac a pris en charge l'établissement du texte à partir de l'exemplaire personnel de Balzac. La version de référence de la Pléiade a été consultée dans certains cas épineux (mais l'interprétation des corrections manuscrites a parfois été divergente).

41. Dans le cadre du projet, les données sont employées de nombreuses manières. Grâce au partenariat avec le projet ARTFL de l'Université de Chicago, il est possible d'interroger le corpus à l'aide d'un moteur de recherche lexicale. L'adaptation du logiciel MEDITE et la création des fichiers XML regroupant les variantes des deux versions différentes d'un texte ont permis de proposer une comparaison informatique de différents états du texte qui facilite une étude génétique de l'œuvre de Balzac. Enfin, grâce à la réutilisation des données mises à disposition par d'autres projets ou des archives numériques, il a été possible de débuter le dernier axe du projet, orienté vers une édition hypertextuelle et visant à recomposer une bibliothèque virtuelle qui comprend l'ensemble des textes littéraires et non littéraires dont Balzac a pu s'inspirer dans la création de *La Comédie humaine*. Mais les possibilités sont loin d'être épuisées. Alors que nous nous intéressons aux inspirations balzaciennes, il paraît stimulant d'investiguer l'influence de l'auteur sur ses successeurs, et comparer *La Comédie humaine* à un corpus postérieur. Le corpus comparatif, composé des fichiers XML qui regroupent deux versions d'un texte et précisent le type des variantes[49], pourrait être exploré massivement, par exemple en extrayant les variantes seules pour déterminer ce qui caractérise les modifications les plus

47. Cf. par exemple (Rageot 2014).

48. Il s'agit de l'édition Michel Lévy (puis Calmann-Lévy) parue de 1869 à 1876. Elle a été ensuite complétée par Conar en 1912-1940, puis perfectionnée enfin par la deuxième édition de la Pléiade (1976-1981), devenue depuis la version de référence. Cf. (Del Lungo 2017).

49. Pour ce faire, nous employons l'élément <choice> dont @ana précise le type de modification (remplacement, suppression, insertion, déplacement). <choice> peut avoir deux fils, <orig>, qui code le contenu du texte publiée antérieurement, et <reg>, qui contient la variante du texte postérieur. Voici, par exemple, la manière dont sont codées les modifications du dédicace de *Modeste Mignon* entre l'édition F et celle du FC (figure 1) :

```
<div type="dedication">
    <pb n="113" xml:id="p113"/>
    <salute>À UNE
            <choice ana="remplacement">
                <orig>ÉTRANGÈRE</orig>
                <reg>POLONAISE</reg>
            </choice>
    </salute>
    <p rend="noindent">Fille d'une terre esclave, ange par l'amour, [...]
            <choice ana="remplacement">
                <orig>les expressions visibles</orig>
                <reg>l'expression quand elle anime ta [...]</reg>
            </choice>
    sont
            <choice ana="remplacement">
                <orig>comme ces</orig>
                <reg>pour les savants les</reg>
            </choice>
    caractères d'un langage perdu
            <choice ana="remplacement">
                <orig> qui préoccupent les savants.</orig>
                <reg>. </reg>
            </choice>
    </p>
            <signed><hi rend="sc">De Balzac.</hi></signed>
</div>
```

fréquentes. Le cas de Balzac est réplicable pour d'autres auteurs, aussi bien Zola que Hugo par exemple.

42. Par ailleurs, différents niveaux de réplicabilité peuvent être mis en avant à travers ce *data paper*, que ce soit au niveau du jeu de données, que des pratiques et méthodes décrites. À la multiplication des utilisations, nous ajoutons la multiplication des formats : la transformation du XML-TEI dans les formats HTML et EPUB permet la diffusion des textes au public plus large et l'utilisation de l'édition au-delà du milieu universitaire (dans le cadre scolaire, voire plus une lecture de plaisir). L'emploi du standard TEI garantit l'interchangeabilité et l'ouverture des données produites dans le cadre du projet et l'emploi de la syntaxe Teinte, plus restreinte, mais adapté au projet et développé avec lui, permet de proposer des textes dotés d'une structuration sémantique. L'établissement rigoureux du format XML a été priorisé, puisque celui-ci est considéré comme un format pivot, à partir duquel nous procédons aux transformations automatiques vers d'autres formats éditoriaux. L'automatisation de la chaîne du traitement garantit par ailleurs la conformité des versions des fichiers et facilite la correction plus rigoureuse d'éventuelles coquilles signalées par les utilisateurs.

43. Comme le constatent Marcello Vitali-Rosati et Michael Sinatra (2014, 60) : « [I]l ne s'agit pas seulement de choisir, de légitimer, de mettre en forme et de diffuser un contenu, mais il s'agit aussi de réfléchir à l'ensemble des techniques que l'on va utiliser ou créer pour le faire, ainsi qu'aux contextes de circulation produits par l'espace numérique. Si les humanités numériques s'occupent de produire des outils et de réfléchir à leur impact sur la production et la circulation du savoir, alors l'éditorialisation devient l'objet central de leur travail. »

44. Le concept de la science ouverte invite à donner accès non seulement aux résultats, mais aussi aux chaînes de traitement qui ont permis leur production. Nous sommes persuadés que le dépôt des données collectées et produites au sein de chaque projet, accompagné d'une documentation facilitant leur réutilisation, permet de remettre l'éditorialisation au centre de la réflexion sur les objets et outils numériques.

Bibliographie

La bibliographie est accessible au format Zotero par ce lien.

Abdul-Rahman, Alfie, Glenn Roe, Mark Olsen, Clovis Gladstone, Richard Whaling, Nicholas Cronk, Robert Morrissey et Min Chen. 2017. « Constructive Visual Analytics for Text Similarity Detection: Constructive Visual Analytics for Text Similarity Detection ». *Computer Graphics Forum* 36 (1) : 237-248. https://doi.org/10.1111/cgf.12798.

Adam, Jean-Michel. 2006. « Autour du concept de texte. Pour un dialogue des disciplines de l'analyse des données textuelles ». Dans *JADT'06 : Actes en ligne*, édité par Jean-Marie Viprey. Paris, France : Lexicometrica. http://lexicometrica.univ-paris3.fr/jadt/JADT2006-PLENIERE/JADT2006_JMA.pdf.

—. 2008. *La Linguistique textuelle : introduction à l'analyse textuelle des discours*. 2e éd. Paris, France : Armand Colin.

— et Ute Heidmann (éd.). 2005. *Sciences du texte et analyse de discours : enjeux d'une interdisciplinarité*. Genève, Suisse : Slatkine.

Adema, Janneke et Gary Hall. 2016. « Posthumanities: The Dark Side of "The Dark Side of the Digital" ». *Journal of Electronic Publishing* 19 (2). https://doi.org/10.3998/3336451.0019.201.

Agamben, Giorgio. 2007. *Qu'est-ce qu'un dispositif ?* Traduit par Martin Rueff. Paris, France : Payot et Rivages.

Ågren, Maria (éd.). 2017. *Making a Living, Making a Difference: Gender and Work in Early Modern European Society*. Oxford, Royaume-Uni : Oxford University Press.

Aijmer, Karin et Bengt Altenberg (éd.). 2004. *Advances in Corpus Linguistics*. Amsterdam, Pays-Bas : Rodopi.

Albert, Anaïs. 2021. « Les conflits du travail dans l'industrie textile à Paris sous le Second Empire ». *Le Mouvement Social* 276 (3) : 107-128. https://doi.org/10.3917/lms1.276.0107.

Allen, Timothy, Clovis Gladstone et Richard Whaling. 2013. « PhiloLogic4: An Abstract TEI Query System ». *Journal of the Text Encoding Initiative* (5 [avril]). https://doi.org/10.4000/jtei.817.

Andro, Mathieu, Marion Chaigne et Franck Smith. 2012. « Valoriser une bibliothèque numérique par des choix de référencement et de diffusion ». *Les Cahiers du numérique* 8 (3) : 75-90. https://doi.org/10.3166/lcn.8.3.75-90.

Anheim, Étienne. 2018. *Le Travail de l'histoire*. Paris, France : Éditions de la Sorbonne.

— et Bénédicte Girault (éd.). 2015. « Recherche historique et enseignement secondaire ». *Annales. Histoire, Sciences Sociales* 70 (1) : 141-214.

Ankerson, Megan Sapnar. 2015. « Take Me Back! Web History as Chronotourism of the Digital Archive ». Communication présentée à *Times and Temporalities of the Web*. Paris, France.

Annales. Économies, Sociétés, Civilisations : Histoire et sciences sociales. Un tournant critique. 1989. Vol. 44. Paris, France : Éditions de l'EHESS. https://www.persee.fr/issue/ahess_0395-2649_1989_num_44_6.

Atkinson, Sarah et Sarah Whatley. 2015. « Digital Archives and Open Archival Practices ». *Convergence* 21 (1) : 3-7. https://doi.org/10.1177/1354856514560292.

Azémard, Ghislaine. 2013. *100 notions pour le crossmédia et le transmédia*. Paris, France : Les Éditions de l'Immatériel.

Bachelard, Gaston. 1993. *La Formation de l'esprit scientifique : contribution à une psychanalyse de la connaissance*. Réédition. Paris, France : Vrin.

Bachimont, Bruno, Fabien Gandon, Gautier Poupeau, Bernard Vatant, Raphaël Troncy, Stéphane Pouyllau, Ruth Martinez, Michèle Battisti et Manuel Zacklad. 2011. « Enjeux et technologies : des données au sens ». *Documentaliste-Sciences de l'Information* 48 (4) : 24-41. https://doi.org/10.3917/docsi.484.0024.

Badouard, Romain. 2016. « "Je ne suis pas Charlie". Pluralité des prises de parole sur le web et les réseaux sociaux ». Dans *Le Défi Charlie. Les médias à l'épreuve des attentats*, par Pierre Lefébure et Claire Sécail. Paris, France : Lemieux Éditeur. https://hal.archives-ouvertes.fr/hal-01251253.

—, Francesca Musiani, Cécile Méadel et Laurence Monnoyer-Smith. 2013. « Towards a Typology of Internet Governance Socio-technical Arrangements ». Dans *Normative Experience in Internet Politics*, par Françoise Massit-Folléa, Cécile Méadel et Laurence Monnoyer-Smith, 99-124. Paris, France : Presses des Mines ; OpenEdition.

Balzac, Honoré de. 2021. « *Modeste* Mignon, dans *La* Comédie humaine, Paris, Furne, 1845, t. IV : *Études de mœurs*, *Scènes* de la vie privée, p. 113-345 ». Édition des sources. Nakala. août 2021. https://nakala.fr/10.34847/nkl.defelcoz.

Barats, Christine (éd.). 2013. *Manuel d'analyse du web en Sciences Humaines et Sociales*. Paris, France : Armand Colin. https://doi.org/10.3917/arco.barat.2013.01.

Bardiot, Clarisse. 2018. « Happy APIs : Débridons les APIS pour développer les humanités numériques ». *DORRA-DH* (blog). 7 septembre 2018. https://dorradh.hypotheses.org/66.

Barrellon, Vincent. 2017. « A Generic Approach Towards the Collaborative Construction of Digital Scholarly Editions ». Thèse de doctorat, Lyon, France : INSA, Université de Lyon. https://tel.archives-ouvertes.fr/tel-02090792.

Bastin, Gilles et Paola Tubaro (éd.) 2018. « Big data, sociétés et sciences sociales ». *Revue française de sociologie* 59 (3) : 375-557.

Battles, Matthew et Michael Maizels. 2016. « Collections and/of Data: Art History and the Art Museum in the DH Mode ». Dans *Debates in the Digital Humanities 2016*, édité par Matthew Gold et Lauren Klein. *Debates in the Digital Humanities* 2. Minneapolis, États-Unis : University of Minnesota Press. https://dhdebates.gc.cuny.edu/read/untitled/section/7cdd40a6-9ef4-4aca-8f53-bfee4cd9ed0e.

Baudens, Stéphane et Alexandre Jeannin. 2009. « Les pièces de procédure des archives du parlement de Flandre : rapport d'activité ». *Revue du Nord* 382 (4) : 739-744.

Beaulande, Véronique. 2019. « Goût de l'archive, goût des lieux, goût des gens ». *Le Goût de l'archive à l'ère numérique* (blog). 5 février 2019. http://www.gout-numerique.net/table-of-contents/gout-de-larchive-gout-des-lieux-gout-des-gens.

Ben Smida, Kaouther. 2016. « Production of First Domain Ontology in CIDOC CRM Format from Heterogeneous Metadata ». Mémoire de master *Data, Knowledge and Dis-*

tributed Systems, Jendouba, Tunisie : Université de Jendouba.

Ben-David, Anat et Adam Amram. 2018. « The Internet Archive and the Socio-technical Construction of Historical Facts ». *Internet Histories* 2 (1-2) : 179-201. https://doi.org/10.1080/24701475.2018.1455412.

Benedetti, Julien. 2018. « La salle de lecture, hors du temps et de l'espace ? » *Le Goût de l'archive à l'ère numérique* (blog). 5 septembre 2018. http://www.gout-numerique.net/table-of-contents/la-salle-de-lecture-hors-du-temps-et-de-lespace.

Bennett, James. 2008. « Interfacing the Nation: Remediating Public Service Broadcasting in the Digital Television Age ». *Convergence* 14 (3) : 277-294. https://doi.org/10.1177/1354856508091081.

Bentz, Bruno et Benjamin Ringot. 2009. « Jacques Rigaud et les recueils des Maisons royales de France ». *Nouvelles de l'estampe* (224) : 23-34.

« Bernard Lepetit ». 1996. *Histoire & Mesure* 11 (1-2). https://www.persee.fr/doc/hism_0982-1783_1996_num_11_1_1771.

Berra, Aurélien. 2012. « Faire des humanités numériques ». Dans *Read/Write Book 2 : Une introduction aux humanités numériques*, édité par Pierre Mounier : 25-43. Marseille, France : OpenEdition Press. http://books.openedition.org/oep/238.

—. 2019. « Revue Humanités numériques : sommaires des numéros 1 et 2 ». *Humanistica* (blog). 6 juillet 2019. http://www.humanisti.ca/revue/revue-hn-sommaires-des-numeros-1-et-2/.

Bertrand, Paul. 2019. « La fin nécessaire et heureuse des Humanités numériques #DHIHA8 ». *MDVZ 3* (blog). 11 juin 2019. https://medievizmesblog.wordpress.com/2019/06/11/la-fin-necessaire-et-heureuse-des-humanites-numeriques/.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge, Royaume-Uni : Cambridge University Press.

—. 1995. *Dimensions of Register Variation: A Cross-linguistic Comparison*. Cambridge, Royaume-Uni : Cambridge University Press.

—. 2004. « Conversation Text Types: A Multi-dimensional Analysis ». Dans *JADT04 : Le poids des mots*, édité par Gérald Purnelle, Cédrick Fairon et Anne Dister, 1 : 15-31. Louvain-la-Neuve, Belgique : Presses universitaires de Louvain.

—, Susan Conrad et Randi Reppen. 1998. *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge, Royaume-Uni : Cambridge University Press.

Billen, Roland, Benoît Jonlet, Andrea Luczfalvy Jancsó, Romain Neuville, Gilles-Antoine Nys, Florent Poux, Muriel Van Ruymbeke, Mathieu Piavaux et Pierre Hallot. 2018. « La transition numérique dans le domaine du patrimoine bâti : un retour d'expérience ». Dans *Bulletin de la Commission royale des Monuments, Sites et Fouilles*, 30 : 119-148. Liège, Belgique : Commission royale des Monuments, Sites et Fouilles de la Région wallonne.

Blandin, Claire et Isabelle Garcin-Marrou. 2018. « Le temps long des archives de presse ». Dans *En quête d'archives. Bricolages méthodologiques en terrains médiatiques.*, édité par Sarah Lécossais et Nelly Quemener : 43-50. Bry-sur-Marne, France : INA Publications.

Bloch, Marc. 1959. *Apologie pour l'histoire ou métier d'historien*. 2e édition. Cahiers des Annales 3. Paris, France : Armand Colin. http://classiques.uqac.ca/classiques/bloch_marc/apologie_histoire/apologie_histoire.html.

Blouin, Francis. 1996. « Cadre de réflexion pour la prise en compte de la diplomatique dans l'environnement électronique ». *La Gazette des archives* 172 (1) : 71-87. https://doi.org/10.3406/gazar.1996.3390.

Bolter, Jay David et Richard Grusin. 1999. *Remediation: Understanding New Media*. Cambridge, États-Unis : MIT Press.

Boukhaled, Mohamed-Amine, Zied Sellami et Jean-Gabriel Ganascia. 2015. « Phoebus : un Logiciel d'Extraction de Réutilisations dans des Textes Littéraires ». Dans *TALN 2015 : 22e Conférence sur le Traitement Automatique des Langues Naturelles*. Atala. https://hal.sorbonne-universite.fr/hal-01198411.

Boullier, Dominique. 2015. « Charlie est un phénomène de 3e génération (aussi) par D. Boullier ». Billet. *SHS 3G* (blog). 1 juin 2015. https://shs3g.hypotheses.org/114.

Bourlet, Caroline, Lucie Fossier, Jean-Philippe Genet, Christiane Klapisch-Zuber, Jacques Lefort, Josette Metman et Gian Piero Zarri. 1979. « Éditorial ». *Le médiéviste et l'ordinateur* 1 (1) : 1-2.

Briatte, François. 2016. « Les réseaux de cosignatures législatives de quinze parlements européens ». Dans *Le réseau. Usages d'une notion polysémique en sciences humaines et sociales*, par Rosemonde Letricot, Mario Cuxac, Maria Uzcategui et Andrea Cavaletto : 215-228. Louvain-la-Neuve, Belgique : Presses universitaires de Louvain. https://halshs.archives-ouvertes.fr/halshs-01435495.

Broughton, Vanda. 2010. « Emergent Vocabulary Control in Web 2.0 ». *Les Cahiers du numérique* 6 (3) : 49-75.

Brügger, Niels. 2009. « Website History and the Website as an Object of Study ». *New Media & Society* 11 (1-2) : 115-132. https://doi.org/10.1177/1461444808099574.

—. 2011. « Web Archiving: between Past, Present, and Future ». Dans *The Handbook of Internet Studies*, édité par Mia Consalvo et Charles Ess : 24-42. Oxford, Royaume-Uni : Wiley-Blackwell. https://doi.org/10.1002/9781444314861.ch2.

—. 2012. « Web History and the Web as a Historical Source ». *Zeithistorische Forschungen. Studies in Contemporary History* 9 (2) : 316-325. https://doi.org/10.14765/zzf.dok-1588.

Brunet, Étienne. 1981. « Le vocabulaire français de 1789 à nos jours d'après les données du Trésor de la Langue Française ». Dans *Actes du Congrès international informatique et sciences humaines 1981*, édité par Louis Delatte : 111-119. Liège, Belgique : Laboratoire d'analyse statistique des langues anciennes, Université de Liège. https://hal.archives-ouvertes.fr/hal-01432697.

—. 2011. *Ce qui compte : méthodes statistiques*. Édité par Céline Poudat. Paris, France : Honoré Champion.

—. 2012. « Au fond du GOOFRE, un gisement de 44 milliards de mots ». Dans *JADT 2012 : 11e Journées internationales d'Analyse statistique des Données Textuelles*, édité par Anne Dister, Dominique Longrée et Gérald Purnelle : 7-21. Liège, Belgique : Université de Liège. https://hal.archives-ouvertes.fr/hal-01790505.

— et Laurent Vanni. 2014. « GOOFRE version 2 ». Dans *JADT 2014 : 12e Journées internationales d'Analyse statistique des Données Textuelles*, édité par Jean-Michel Daube, Mathieu Valette et Serge Fleury : 106-119. Vincennes, France. https://hal.archives-ouvertes.fr/hal-01196595.

« Budapest Open Access Initiative ». 2002. https://www.budapestopenaccessinitiative.org/boai15-1.

Burnard, Lou. 2016. « ODD Chaining for Beginners ». Guide pratique. http://teic.github.io/PDF/howtoChain.pdf.

Bygrave, Lee et Jon Bing. 2009. *Internet Governance: Infrastructure and Institutions.* Oxford, Royaume-Uni : Oxford University Press. http://www.SLQ.eblib.com.au/patron/FullRecord.aspx?p=430398.

Calas, Frédéric (éd.). 2006. *Cohérence et discours*. Paris, France : Presses de l'université Paris-Sorbonne.

Cameron, Fiona et Sarah Kenderdine (éd.). 2010. *Theorizing digital cultural heritage: a critical discourse*. Media in transition. Cambridge, États-Unis : MIT Press.

Cardon, Dominique. 2013. « Dans l'esprit du PageRank. Une enquête sur l'algorithme de Google. » *Réseaux* 177 (1) : 63-95. https://doi.org/10.3917/res.177.0063.

—. 2019. *Culture numérique*. Presses de Sciences Po. https://www.cairn.info/culture-numerique--9782724623659.htm.

Cardon, Rémi. 2017. « Extraction d'informations pour construire une base de connaissances sur le patrimoine industriel textile à partir de sources de données hétérogènes ». Mémoire de master Sciences du langage, Lille, France : Université Lille 3.

Cardoni, Fabien. 2012. « Aux sources du budget domestique selon Le Play ». *Les Études sociales* 155 (1) : 11-46.

Carnino, Guillaume et François Jarrige. 2019. « L'université sous hypnose numérique ». Dans *Critiques de l'école numérique*, édité par Cédric Biagini : 293-316. Montreuil, France : L'Échappée.

Caron, François. 1997. *Les Deux révolutions industrielles du XX^e siècle*. Paris, France : Albin Michel.

Carusi, Annamaria et Torsten Reimer. 2010. « Virtual Research Environment Collaborative Landscape Study: A JISC Funded Project ». Rapport de recherche. King's College London ; University of Oxford. http://www.jisc.ac.uk/media/documents/publications/vrelandscapereport.pdf.

Cavalié, Étienne, Frédéric Clavert, Olivier Legendre et Dana Martin (éd.). 2017. *Expérimenter les humanités numériques : Des outils individuels aux projets collectifs*. Montréal, Canada : Presses de l'université de Montréal. http://books.openedition.org/pum/11091.

Cazals, Géraldine. 2018. *L'Arrestographie flamande : jurisprudence et littérature juridique à la fin de l'Ancien Régime (1668-1789)*. Genève, Suisse : Droz.

Chabin, Marie-Anne. 2000. *Le Management de l'archive*. Paris, France : Hermès science.

—. 2007. *Archiver, et après ?* Paris, France : Djakarta.

—. 2011. « Peut-on parler de diplomatique numérique ? » *Le Blog de Marie-Anne Chabin* (blog). 2011. http://www.marieannechabin.fr/diplomatique-numerique/.

—. 2012. « L'ère numérique du faux ». *Médium* 31 (2). https://doi.org/10.3917/mediu.031.0046.

—. 2013. « Peut-on parler de diplomatique numérique ? » Dans *Vers un nouvel archiviste numérique*, par Valentine Frey. Paris, France : L'Harmattan.

—.2018a.« L'image à la une et la désinformation subliminale ». *Le Blog de Marie-Anne Chabin* (blog). 2018. http://www.marieannechabin.fr/2018/12/limage-a-la-une-et-la-desinformation-subliminale/.

—. 2018b. « Analyse d'un faux (arnaque au RGPD par fax) ». *Arcateg, méthode d'archivage par catégorie* (blog). 27 juin 2018. https://www.arcateg.fr/2018/06/27/analyse-dun-faux-arnaque-au-rgpd-par-fax/.

—. 2020. « The Potential for Collaboration between AI and Archival Science in Processing Data from the French Great National Debate ». *Records Management Journal*, février. https://doi.org/10.1108/RMJ-08-2019-0042.

Chagué, Alix. 2018. « Constituer un corpus pour la fouille de texte – de la transcription des documents d'archives à l'annotation. Exploration d'une méthodologie par l'ANR Time Us ». Thèse de master, École nationale des chartes.

—, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie. 2022. « Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non uniforme ? » Dans *La Fabrique numérique des corpus en sciences humaines et sociales*, édité par Clarisse Bardiot, Émilien Ruiz et Esther Dehoux. Humanités numériques et science ouverte 1. Presses universitaires du Septentrion.

—. 1983. « Coherence as a Principle in the Interpretation of Discourse ». *Text* 3 (1). https://doi.org/10.1515/text.1.1983.3.1.71.

Charolles, Michel. 1995. « Cohésion, cohérence et pertinence du discours ». *Travaux de Linguistique* 29 : 125-151.

Charron, Jean et Jean De Bonville. 2004. « Les mutations du journalisme : modèle explicatif et orientations méthodologiques ». Dans *Nature et transformation du journalisme : théorie et recherches empiriques*, édité par Colette Brin, Jean Charron et Jean De Bonville : 87-120. Québec, Canada : Presses de l'université Laval.

« Charte Nizhny Tagil pour le patrimoine industriel ». 2003. Nizhny Tagil, Russie : The International Committee for the Conservation of the Industrial Heritage (TICCIH). https://www.icomos.org/18thapril/2006/nizhny-tagil-charter-f.pdf.

Chaudiron, Stéphane, Bernard Jacquemin et Éric Kergosien. 2019. « L'apport du Web sémantique à la sauvegarde du patrimoine immatériel : le cas du textile, de la musique et de la mine ». Dans *Information, communication et humanités numériques : Enjeux et défis pour un enrichissement épistémologique : Actes du 23e colloque bilatéral franco-roumain en sciences de l'information et de la communication*, édité par Ioan Roxin, Federico Tajariol, Ioan Hosu et Nicolas Pélissier : 311-328. Cluj-Napoca, Roumanie : Accent Publisher.

Cheniti, Tarek. 2009. « Global Internet Governance in Practice. Mundane Encounters and Multiple Enactments ». Thèse de doctorat, Oxford, Royaume-Uni : University of Oxford.

Chun, Wendy Hui Kyong. 2011. *Programmed visions: Software and memory*. Software studies. Cambridge, États-Unis : MIT Press.

Clavert, Frédéric. 2017. « Le goût de l'API ». *Le Goût de l'archive à l'ère numérique* (blog). 20 octobre 2017. http://www.gout-numerique.net/table-of-contents/gout-api.

— et Caroline Muller. 2017. « Introduction : Le goût de l'archive à l'ère numérique ». *Le Goût de l'archive à l'ère numérique* (blog). 23 octobre 2017. http://www.gout-numerique.net/.

— et Caroline Muller (éd.). 2019. *La Gazette des archives : Le Goût de l'archive à l'ère numérique*. Vol. 253. Paris, France : Association des archivistes français.

— et Valérie Schafer. 2019. « Les humanités numériques, un enjeu historique ». *Quaderni* (98 [février]) : 33-49. https://doi.org/10.4000/quaderni.1417.

Cleyet-Michaud, Rosine. 2009. « Le fonds du parlement de Flandre : historique de la conservation ». *Revue du Nord* 382 (4) : 683-685. https://doi.org/10.3917/rdn.382.0683.

Cohen, Deborah. 2015. « Silent Changes to the History Manifesto ». *Deborah Cohen* (blog). 23 mars 2015. http://www.deborahacohen.com/profile/?q=content/silent-changes-history-manifesto.

« Convention concernant la protection du patrimoine mondial culturel et naturel ». 1973. *Museum International* 25 (1-2). https://doi.org/10.1111/j.1755-5825.1973.tb02107.x.

« Convention pour la sauvegarde du patrimoine culturel immatériel ». 2003. Rapport de conférence. Paris, France : UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000132540_fre.

Cordell, Ryan. 2015. « Reprinting, Circulation, and the Network Author in Antebellum Newspapers ». *American Literary History* 27 (3) : 417-445. https://doi.org/10.1093/alh/ajv028.

—. 2017. « "Q i-jtb the Raven": Taking Dirty OCR Seriously ». *Book History* 20 (1) : 188-225. https://doi.org/10.1353/bh.2017.0006.

— et David Smith. 2017. « The Viral Texts Project: Mapping Networks of Reprinting in 19th-century Newspapers and Magazines ». *Viral Texts*. http://viraltexts.org.

Corpora. 2006. Revue en ligne. 15 vol. Edinburgh University Press. https://www.euppublishing.com/loi/cor.

Corpus. 2002. Revue en ligne. 21 vol. Bases, corpus, langage (UMR 7320). https://journals.openedition.org/corpus.

Cottereau, Alain. « Justice et injustice ordinaire sur les lieux de travail d'après les audiences prud'homales (1806-1866) ». *Le Mouvement social*, nº 141 (1987): 25-59. https://doi.org/10.2307/3778206.

—. 2006. « Sens du juste et usages du droit du travail : une évolution contrastée entre la France et la Grande-Bretagne au XIXe siècle ». *Revue d'histoire du XIXe siècle*, nº 33 (décembre) : 101-120. https://doi.org/10.4000/rh19.1148.

Coustaty, Mickaël, Norbert Tsopze, Karell Bertet, Alain Bouju et Georges Louis. 2012. « Traitement des documents anciens à l'aide d'ontologie ». *Les Cahiers du numérique* 8 (3) : 91-118. https://doi.org/10.3166/lcn.8.3.91-118.

Crews, Kenneth. 2010. « Copyright, Museums and Licensing of Art Images ». Rapport de recherche. Kress Foundation. http://www.kressfoundation.org/research/copyright_museums_and_licensing_of_art_images/.

Crofts, Nick, Ifigenia Dionissiadou, Martin Doerr et Matthew Stiff. 1999. « Définition du Modèle Conceptuel de Référence du CIDOC (CRM) ». Version 2.1. ICOM/CIDOC. http://old.cidoc-crm.org/docs/crm_french_version.pdf.

—, Martin Doerr, Tony Gill, Stephen Stead et Matthew Stiff. 2011. « Definition of the CIDOC Conceptual Reference Model ». Version 5.0.4. ICOM/CIDOC. http://cidoc-crm.org/sites/default/files/cidoc_crm_version_5.0.4.pdf.

Cuno, James. 2016. « Introducing Getty Scholars' Workspace, An Open-Source Humanities Research Tool ». *Getty Iris* (blog). 24 février 2016. http://blogs.getty.edu/iris/introducing-getty-scholars-workspace-an-open-source-humanities-research-tool/.

Dacos, Marin. 2011. « Manifeste des *Digital humanities* ». *THATCamp Paris* (blog). 26 mars 2011. https://tcp.hypotheses.org/318.

Daloz, Amélie. 2018. « Vers la représentation terminologique du patrimoine minier ». Dans *Terminologie & Ontologie : Théories et Applications, Actes de la conférence TOTh 2018*. Terminologica. Chambéry, France : Presses universitaires Savoie Mont Blanc.

— et Stéphane Chaudiron. 2019. « Méthodologie de conception d'un thésaurus du domaine minier ». Dans *Actes du 21e Colloque international sur le document numérique (CIDE 21) : La numérisation info-documentaire* : 11-23. Djerba, Tunisie : Europia. https://hal.archives-ouvertes.fr/hal-02568840.

Dalton, Craig, Linnet Taylor et Jim Thatcher. 2016. « Critical Data Studies: A dialog on data and space ». *Big Data & Society* 3 (1). https://doi.org/10.1177/2053951716648346.

De Maeyer, Juliette et Dominique Trudel. 2018. « @franklinfordbot: Remediating Franklin Ford ». *Digital Journalism* 6 (9) : 1270-1287. https://doi.org/10.1080/21670811.2018.1514273.

« Déclaration de Lyon sur l'accès à l'Information et au Développement ». 2014. https://www.lyondeclaration.org/.

« Déclaration de Mexico sur les politiques culturelles ». 1982. Mexico, Mexique : UNESCO. https://www.culture.gouv.fr/Media/Thematiques/Egalite-et-diversite/College-de-la-Diversite/Declaration-de-Mexico.

Del Lungo, Andrea. 2017. « Éditions et représentations de *La* Comédie humaine ». *Genesis. Manuscrits – Recherche – Invention*, nº 44 (mai) : 81-96. https://doi.org/10.4000/genesis.1747.

— et Karolina Suchecka. 2022. « Projet eBalzac : construire une bibliothèque hypertextuelle des sources intellectuelles ». Dans *La Fabrique numérique des corpus en sciences humaines et sociales*. Sous la direction de Clarisse Bardiot, Émilien Ruiz, et Esther Dehoux. Presses universitaires du Septentrion.

Delalande, Nicolas et Julien Vincent. 2011. « Portrait de l'historien.ne en cyborg ». *Revue d'histoire moderne & contemporaine* 58-4bis (5) : 5-29. https://doi.org/10.3917/rhmc.585.0005.

Delmas, Bruno. 2003. « Donner à l'image et au son le statut de l'écrit : pour une critique diplomatique des documents audiovisuels ». *Bibliothèque de l'École des chartes* 161 (2) : 553-601. https://doi.org/10.3406/bec.2003.463630.

Delmotte, Stéphanie. 2009. « Publications scientifiques en sciences humaines ». *Les Cahiers du numérique* 5 (2) : 53-84. https://doi.org/10.3166/lcn.5.2.53-84.

Demars-Sion, Véronique. 2014. « "Heurts" et malheurs d'un fonds : les tribulations des archives du parlement de Flandre ». *Revue du Nord* 407 (4) : 829-858. https://doi.org/10.3917/rdn.407.0829.

DeNardis, Laura. 2014. *The Global War for Internet Governance*. New Haven, États-Unis : Yale University Press. https://doi.org/10.12987/yale/9780300181357.001.0001.

Denoyelle, Martine, Katie Durand, Johanna Daniel et Elli Doulkaridou-Ramantani. 2018. « Droits des images, histoire de l'art et société. » Rapport de recherche. Paris, France : Fondation de France. https://halshs.archives-ouvertes.fr/halshs-02066987.

Détrie, Catherine, Paul Siblot et Bertrand Verine. 2001. *Termes et concepts pour l'analyse du discours : une approche praxématique*. Paris, France : Honoré Champion.

Doerr, Martin. 2003. « The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interope-

rability of Metadata ». *AI Magazine* 24 (3) : 75-92. https://doi.org/10.1609/aimag.v24i3.1720.

—, Chryssoula Bekiari et Patrick Le Bœuf. 2008. « FRBRoo, a Conceptual Model for Performing Arts ». Communication présentée à *Annual Conference of CIDOC 2008*, 16 septembre. https://www.ics.forth.gr/_publications/drfile.2008-06-42.pdf.

—, Stefan Gradmann, Patrick LeBoeuf, Trond Aalberg, Rodolphe Bailly et Marlies Olensky. 2013. « Final Report on EDM – FRBRoo Application Profile Task Force ». Europeana. https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/EDM_FRBRoo/TaskfoApplication%20Profile%20EDM-FRBRoo.pdf.

Domange, Camille. 2013. « Ouverture et partage des données publiques culturelles. Pour une (r)évolution numérique dans le secteur culturel ». Rapport technique 2013-03. Département des programmes numériques, ministère de la Culture et de la Communication. https://www.vie-publique.fr/sites/default/files/rapport/pdf/144000037.pdf.

Dosse, François. 2010. *L'Histoire en miettes : des « Annales » à la « nouvelle histoire »*. Paris, France : La Découverte.

Doueihi, Milad. 2011. *La Grande conversion numérique, suivi de Rêveries d'un promeneur numérique*. Traduit par Paul Chemla. Paris, France : Seuil.

Dougherty, Jack et Kristen Nawrotzki. 2013. *Writing History in the Digital Age*. Ann Arbor, États-Unis : University of Michigan Press. https://doi.org/10.3998/dh.12230987.0001.001.

Drotner, Kirsten, Vince Dziekan, Ross Parry et Kim Christian Schrøder (éd.). 2018. *The Routledge handbook of museums, media and communication*. Routledge international handbooks. Londres, Royaume-Uni : Routledge. https://doi.org/10.4324/9781315560168.

Du Château, Stefan, Danielle Boulanger et Eunika Mercier-Laurent. 2012. « Approche interdisciplinaire du management des connaissances en patrimoine culturel ». *Documentaliste-Sciences de l'Information* 49 (4). https://doi.org/10.3917/docsi.494.0062.

Duclos, Tania. 2013. « Écriture romanesque, intertextualité et genèse chez Balzac : l'exemple de deux romans des années 1838-1839, Béatrix et une fille d'Ève ». Thèse de doctorat, Paris, France : Université Paris-Sorbonne.

Dupont, Yoann. 2017. « Exploration de traits pour la reconnaissance d'entités nommées du Français par apprentissage automatique ». Dans *Recital 2017 : 19e Rencontres jeunes chercheurs en informatique pour le TAL*. Atala. http://taln2017.cnrs.fr/wp-content/uploads/2017/06/actes_RECITAL_2017_2.pdf#page=52.

Duranti, Luciana. 1998. *Diplomatics: New Uses for an Old Science*. Lanham, États-Unis : Scarecrow Press.

—. 2003. « Pour une diplomatique des documents électroniques ». *Bibliothèque de l'École des chartes* 161 (2) : 603-623. https://doi.org/10.3406/bec.2003.463631.

—. 2004. « La conservation à long terme des documents dynamiques et interactifs : InterPARES 2 ». *Document numérique* 8 (2) : 73-86. https://doi.org/10.3166/dn.8.2.73-86.

—. 2010. « From Digital Diplomatics to Digital Records Forensics ». *Archivaria* 68 (janvier) : 39-66.

— et Elizabeth Shaffer (éd.). 2012. « The Memory of the World in the Digital Age: Digitization and Preservation. An international Conference on Permanent Access to Digi-

tal Documentary Heritage ». Dans *Conference Memory of the World 20th Anniversary*. Vancouver, Canada : UNESCO.

« Éditorial ». 1986. *Histoire & Mesure* 1 (1) : 5-6.

Ertzscheid, Olivier. 2012. « Et si on enseignait VRAIMENT le numérique ? » *affordance.info* (blog). 30 avril 2012. https://www.affordance.info/mon_weblog/2012/04/et-si-on-enseignant-vraiment-le-numerique-.html.

Farge, Arlette. 1997. *Le Goût de l'archive*. Paris, France : Seuil.

Febvre, Lucien. 1953. « De 1892 à 1933. Examen de conscience d'une histoire et d'un historien ». Dans *Combats pour l'histoire*, 1re édition : 3-17. Paris, France : Armand Colin. http://classiques.uqac.ca/classiques/febvre_lucien/Combats_pour_lhistoire/Combats_pour_lhistoire.html.

Fenoglio, Irène et Jean-Gabriel Ganascia. 2008. « Le logiciel Medite : approche comparative de documents de genèse ». Dans *L'Édition du manuscrit : de l'archive de création au scriptorium électronique*, édité par Aurèle Crasson : 209-228. Louvain-la-Neuve, Belgique : Bruylant-Academia.

Ferlin, Fabrice. 2008. « D'Alembert et l'optique : l'Encyclopédie comme banc d'essai de recherches originales ». *Recherches sur Diderot et sur l'Encyclopédie* (43 [octobre]) : 127-144. https://doi.org/10.4000/rde.3572.

« Feuille de route stratégique : Métadonnées culturelles et transition Web 3.0 ». 2014. Rapport ministériel. Ministère de la culture et de la communication. https://www.enssib.fr/bibliotheque-numerique/notices/64776-feuille-de-route-strategique-metadonnees-culturelles-et-transition-web-3-0.

Finnemann, Niels Ole. 2015. « Hypertextual Relations in Digital Born Materials: Hypertext and Time: Towards a Genre Analysis of Heterogeneous Digital Materials ». Dans *Web Archives as Scholarly Sources: Issues, Practices, Perspectives*. Aarhus, Danemark.

Fort, Karen, Maud Ehrmann et Adeline Nazarenko. 2009. « Vers une méthodologie d'annotation des entités nommées en corpus ? » Dans *Traitement Automatique des Langues Naturelles 2009*. Senlis, France. https://hal.archives-ouvertes.fr/hal-00402321.

Francony, Jean-Marc. 2018. « L'éditorialisation des données aux bornes des API : enjeux et perspectives pour une analyse empirique ». *Les enjeux de l'information et de la communication* 19 (2) : 69-79. https://doi.org/10.3917/enic.025.0069.

Fréger, Laurie. 2009. « Les épices au parlement de Flandre : pratiques singulières ? » *Revue du Nord* 382 (4) : 847-866. https://doi.org/10.3917/rdn.382.0847.

Gaillard, Claire-Lise. 2018. « Feuilleter la presse ancienne par Giga Octets ». *Le Goût de l'archive à l'ère numérique* (blog). 4 juin 2018. http://www.gout-numerique.net/table-of-contents/feuilleter-la-presse-ancienne-par-giga-octets.

Galleron, Ioana, Marie-Luce Demonet, Cécile Meynard, Idmhand Fatiha, Elena Pierazzo, Geoffrey Williams, Julia Roger, et Pierre-Yves Buard. 2018. « Les publications numériques de corpus d'auteurs – Guide de travail, grille d'analyse et recommandations ». Rapport de recherche. Huma-Num. Consulté le 5 mars 2022. https://halshs.archives-ouvertes.fr/halshs-01932519.

Galvez-Behar, Gabriel. 2017. « En finir avec le triple échec de l'Université : Réhabiliter la pédagogie universitaire ». *Gabriel Galvez-Behar* (blog). 14 mai 2017. https://ggb.ouvaton.org/spip.php?article73.

Ganascia, Jean-Gabriel. 2011. « Medite: A Unilingual Text Aligner for Humanities. Applications to Textual Genetics and to the Edition of Text Variants ». Communication présentée à *Supporting Digital Humanities (SDH 2011)*, novembre. http://www-poleia.lip6.fr/~ganascia/Publications?action=AttachFile&do=view&target=SDH2011.pdf.

— et Jean Bourdaillet. 2006. « Alignements unilingues avec Medite ». Dans *JADT'06 : 8e Journées internationales d'Analyse statistique des Données Textuelles*, édité par Jean-Marie Viprey, 1 : 427-437. Besançon, France : Presses universitaires de Franche-Comté.

—, Pierre Glaudes et Andrea Del Lungo. 2014. « Automatic Detection of Reuses and Citations in Literary Texts ». *Literary and Linguistic Computing* 29 (3) : 412-421. https://doi.org/10.1093/llc/fqu020.

Giglietto, Fabio et Yenn Lee. 2015. « To Be or Not to Be Charlie: Twitter Hashtags as a Discourse and Counter-discourse in the Aftermath of the 2015 *Charlie Hebdo* Shooting in France ». Dans *#Microposts2015*, 1395 : 33-37. http://ceur-ws.org/Vol-1395/.

Giovacchini, Julie. 2018. « De la source à l'image : y a-t-il une philologie numérique ? » *Le Goût de l'archive à l'ère numérique* (blog). 6 juillet 2018. http://www.gout-numerique.net/table-of-contents/de-la-source-a-limage-y-a-t-il-une-philologie-numerique.

Gitelman, Lisa. 2006. *Always Already New: Media, History and the Data of Culture*. Cambridge, États-Unis : MIT Press.

Glorieux, Frédéric. 2015. « Le traitement de textes (odt) pour produire des documents structurés (XML/TEI) – Odette ». *J'attends des résultats. Fouille de documents, expériences réussies et ratées*. https://resultats.hypotheses.org/267.

Gnoli, Claudio. 2012. « Des métadonnées représentant quoi ? Établissement d'une distinction entre les dimensions ontiques, épistémologiques et documentaires dans l'organisation des connaissances ». Dans *L'Organisation des connaissances : dynamisme et stabilité*, édité par Widad Mustafa El Hadi : 51-66. Paris, France : Hermès science.

Gomes, Daniel, João Miranda et Miguel Costa. 2011. « A Survey on Web Archiving Initiatives ». Dans *Research and Advanced Technology for Digital Libraries*, édité par Stefan Gradmann, Francesca Borri, Carlo Meghini et Heiko Schuldt, 6966 : 408-420. Berlin, Allemagne : Springer. https://doi.org/10.1007/978-3-642-24469-8_41.

Goyet, Samuel. 2017. « De briques et de blocs. La fonction éditoriale des interfaces de programmation (api) web : entre science combinatoire et industrie du texte ». Thèse de doctorat, Paris, France : Université Paris-Sorbonne. http://www.theses.fr/2017PA040188.

Grandi, Elisa et Émilien Ruiz. 2012. « Ce que le numérique fait à l'historien.ne. Entretien avec Claire Lemercier ». *Diacronie. Studi di Storia Contemporanea* 10 (2). https://doi.org/10.4000/diacronie.2780.

Guaresi, Magali. 2018. *Parler au féminin. Les professions de foi des député.e.s sous la Cinquième République (1958-2007)*. Humanités numériques. Paris, France : L'Harmattan. https://www.editions-harmattan.fr/index_harmattan.asp?navig=catalogue&obj=livre&no=59698.

Guilbaud, Alexandre. 2017. « L'ENCCRE, édition numérique collaborative et critique de l'Encyclopédie ». *Recherches sur Diderot et sur l'Encyclopédie* (52 [décembre]) : 5-22. https://doi.org/10.4000/rde.5488.

Guilhaumou, Jacques. 2006. *Discours et événement : l'histoire langagière des concepts*. Besançon, France : Presses universitaires de Franche-Comté.

Guiraud, Pierre. 1954. *Les Caractères statistiques du vocabulaire : essai de méthodologie*. Paris, France : Presses universitaires de France.

Guldi, Jo et David Armitage. 2014. *The History Manifesto*. Cambridge, Royaume-Uni : Cambridge University Press.

Gupta, Maya R., Nathaniel P. Jacobson et Eric K. Garcia. 2007. « OCR Binarization and Image Pre-processing for Searching Historical Documents ». *Pattern Recognition* 40 (2) : 389-397. https://doi.org/10.1016/j.patcog.2006.04.043.

Guyon, Céline. 2018. « La fabrique de l'archive : le rituel de la collecte des archives ». *Le Goût de l'archive à l'ère numérique* (blog). 15 juillet 2018. http://www.gout-numerique.net/table-of-contents/la-fabrique-de-larchive-le-rituel-de-la-collecte-des-archives.

Guyotjeannin, Olivier. 2011. « Notions de diplomatique ». Cours en ligne. École nationale des chartes : Theleme (Techniques pour l'Historien en Ligne : Études, Manuels, Exercices, Bibliographies). http://theleme.enc.sorbonne.fr/cours/diplomatique.

Haber, Peter. 2011. *Digital Past: Geschichtswissenschaft im digitalen Zeitalter*. Munich, Allemagne : Oldenbourg Verlag.

Habert, Benoît, Adeline Nazarenko et André Salem. 1997. *Les Linguistiques de corpus*. Paris, France : Armand Colin.

Halliday, Michael Alexander Kirkwood et Ruqaiya Hasan. 1976. *Cohesion in English*. Londres, Royaume-Uni : Longman.

Hamilton, Gill et Fred Saunderson (éd.). 2017. *Open licensing for cultural heritage*. Londres, Royaume-Uni : Facet Publishing. https://doi.org/10.29085/9781783302505.

Hathout, Nabil et Franck Sajous. 2016. « Wiktionnaire's Wikicode GLAWIfied: a Workable French Machine-Readable Dictionary ». Dans *LREC 2016: 10th International Conference on Language Resources and Evaluation*, 1369-1376. Portoroz, Slovénie : European Language Resources Association (ELRA). https://www.aclweb.org/anthology/L16-1218.

Hayles, Nathalie Katherine. 2016. *Lire et penser en milieux numériques : Attention, récits, technogenèse*. Grenoble, France : UGA Éditions. https://doi.org/10.4000/books.ugaeditions.379.

Hedges, Mark, Heike Neuroth, Kathleen Marie Smith, Tobias Blanke, Laurent Romary, Marc Küster et Malcolm Illingworth. 2013. « TextGrid, TEXTvre, and DARIAH: Sustainability of Infrastructures for Textual Scholarship ». *Journal of the Text Encoding Initiative* (5 [avril]). https://doi.org/10.4000/jtei.774.

Heerma Van Voss, Lex, Els Hiemstra-Kuperus et Elise Van Nederveen Meerkerk (éd.). 2010. *The Ashgate Companion to the History of Textile Workers, 1650-2000*. Farnham, Royaume-Uni : Ashgate Publishing.

Heiden, Serge. 2010. « The TXM Platform: Building Open-source Textual Analysis Software Compatible with the TEI Encoding Scheme ». Dans *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation* : 389-398. Tokyo, Japon : Institute of Digital Enhancement of Cognitive Processing, Waseda University. https://www.aclweb.org/anthology/Y10-1044.

Heimburger, Franziska et Émilien Ruiz. 2011. « Faire de l'histoire à l'ère numérique : retours d'expériences ». *Revue d'histoire moderne & contemporaine* 58-4bis (5) : 70-89. https://doi.org/10.3917/rhmc.585.0070.

Henneton, Thibault. 2017. « Les tâcherons du clic ». *Manière de voir* 156 (12) : 32-33.

Hodge, Gail. 2000. *Systems of Knowledge Organization for Digital Libraries: Beyond Traditional Authority Files*. Washington, États-Unis : Digital Library Federation, Council on Library and Information Resources. https://eric.ed.gov/?id=ED440657.

Holley, Rose. 2009. « How Good Can It Get? : Analysing and Improving OCR Accuracy in Large Scale Historic Newspaper Digitisation Programs ». *D-Lib Magazine* 15 (3/4). https://doi.org/10.1045/march2009-holley.

Horrell, Sara et Jane Humphries. 1992. « Old Questions, New Data, and Alternative Perspectives: Families' Living Standards in the Industrial Revolution ». *The Journal of Economic History* 52 (4) : 849-880. https://doi.org/10.1017/S0022050700011931.

Horton, Russell, Mark Olsen et Glenn Roe. 2011. « Something Borrowed: Sequence Alignment and the Identification of Similar Passages in Large Text Collections ». *Digital Studies/Le champ numérique* 2 (1). https://doi.org/10.16995/dscn.258.

Humphries, Jane. 2010. « The First Industrial Nation and the First "Modern" Family ». Dans *Gender Inequalities, Households and the Production of Well-being in Modern Europe*, édité par Tindara Addabbo : 41-58. Farnham, Royaume-Uni : Ashgate Publishing.

Humphries, Jane et Carmen Sarasúa. 2012. « Off the Record: Reconstructing Women's Labor Force Participation in the European Past ». *Feminist Economics* 18 (4) : 39-67. https://doi.org/10.1080/13545701.2012.746465.

Hurez, Jean-Pierre. 2015. *Des fibres et des hommes : promenade au cœur de collections textiles*. Proscitec. Lille, France : La Voix du Nord.

Iliadis, Andrew et Federica Russo. 2016. « Critical data studies: An introduction ». *Big Data & Society* 3 (2). https://doi.org/10.1177/2053951716674238.

Ingarao, Maud et Samantha Saïdi. 2011. « Guide pratique pour la production de corpus numérique ». Guide pratique. Mutec. http://mutec.huma-num.fr/sites/www.mutec-shs.fr/files/Guide%20pratique%20pour%20la%20production%20de%20corpus%20num%C3%A9rique_0.pdf.

Isaac, Antoine et Valentine Charles. 2015. « Enhancing the Europeana Data Model (EDM) ». Livre blanc. Europeana. https://pro.europeana.eu/files/Europeana_Professional/Publications/EDM_WhitePaper_17062015.pdf.

Jacob, Christian. 2019. « Pour une herméneutique numérique en sciences historiques ». *Lieux de savoir* (blog). 27 avril 2019. https://lieuxdesavoir.hypotheses.org/1549.

Jeanneney, Jean-Noël. 2006. *Quand Google défie l'Europe : plaidoyer pour un sursaut*. 2e éd. Essai. Paris, France : Mille et une nuits.

Kahle, Brewster. 2014. « Please Help Protect Net Neutrality ». *Internet Archive Blogs* (blog). 10 septembre 2014. https://blog.archive.org/2014/09/10/please-help-protect-net-neutrality/.

—. 2016. « Help Us Keep the Archive Free, Accessible, and Reader Private ». *Internet Archive Blogs* (blog). 29 novembre 2016. https://blog.archive.org/2016/11/29/help-us-keep-the-archive-free-accessible-and-private/.

Kennedy, Graeme. 1998. *An Introduction to Corpus Linguistics*. Londres, Royaume-Uni : Longman.

Kergosien, Éric, Bernard Jacquemin, Marta Severo et Stéphane Chaudiron. 2015. « The Tectoniq Project for Valorization of Digital Textile Industrial Heritage in North of France ». Dans *Industrial Heritage in the Twenty-first Century, New Challenges*. Lille, France : The International Committee for the Conservation of the Industrial Heritage (TICCIH). https://hal.archives-ouvertes.fr/hal-01358478.

—, Marta Severo et Marie-Aimée Berthelot. 2019. « Cartographier les acteurs d'un territoire : une approche appliquée au patrimoine industriel textile des Hauts-De-France ». Dans *Demande(s) territoriale(s)*, édité par Romain Lajarge, Laurent Cailly, Anne Ruas et Guy Saez : 66-72. Paris, France : Éditions Karthala. https://halshs.archives-ouvertes.fr/CIST2016/hal-01353660.

— et Mathilde Wybo. 2017. « DENIM : Données numériques, langages et représentations du patrimoine textile en région Hauts-de-France, quelles compréhensions réciproques ? » Rapport de recherche. Université Lille 3. https://reccits.hypotheses.org/files/2018/01/DENIM-Rapport2017-Def.pdf.

Kessous, Emmanuel. 2012. *L'Attention au monde : sociologie des données personnelles à l'ère numérique*. Paris, France : Armand Colin.

Kheraj, Sean. 2014. « Canada's Historical Newspaper Digitization Problem, Part 2 ». *Active History* (blog). 13 février 2014. http://activehistory.ca/2014/02/historical-newspaper-digitization-problem/.

Kirschenbaum, Matthew G., Richard Ovenden, Gabriela Redwine et Rachel Donahue. 2010. *Digital Forensics and Born-digital Content in Cultural Heritage Collections*. Washington, États-Unis : Council on Library and Information Resources.

Koyré, Alexandre. 1988. *Du monde clos à l'univers infini*. Traduit par Raissa Tarr. Réédition. Paris, France : Gallimard.

La Barre, Kathryn. 2010. « A Semantic (Faceted) Web? » *Les Cahiers du numérique* 6 (3) : 103-131.

Lainé, Brigitte. 2006. *Le conseil de prud'hommes du département de la Seine : 1844-1940 (1762-1971). Répertoire des fonds conservés aux Archives de Paris*. Paris, France : Archives de Paris.

« La longue durée en débat ». 2015. *Annales. Histoire, Sciences Sociales* 70 (2) : 285-287. https://doi.org/10.3917/anna.702.0285.

Laks, Bernard. 2008. « Pour une phonologie de corpus ». *Journal of French Language Studies* 18 (1) : 3-32. https://doi.org/10.1017/S0959269507003146.

Landragin, Frederic et Céline Poudat. 2017. *Explorer des données textuelles : Méthodes, pratiques, outils*. 1re édition. Paris, France : De Boeck Supérieur.

Langlais, Pierre-Carl. 2017. « Identifier les rubriques de presse ancienne avec du topic modeling ». *Numapresse* (blog). 31 mars 2017. https://numapresse.hypotheses.org/11.

—. 2018. « Avancement du projet Numapresse ». *Numapresse* (blog). 20 janvier 2018. http://www.numapresse.org/2018/01/20/avancement-du-projet-numapresse/.

—. 2019. « Les avancées de Numapresse : pour une approche contextuelle du Text Mining ». *Numapresse* (blog). 22 janvier 2019. http://www.numapresse.org/2019/01/22/les-avancees-de-numapresse-pour-une-approche-contextuelle-du-text-mining/.

Langlois, Charles-Victor et Charles Seignobos. 1992. *Introduction aux études historiques*. Paris, France : Éditions Kimé. http://classiques.uqac.ca/classiques/langlois_charles_victor/intro_etudes_historiques/intro_etudes_historiques.html.

Lanthaler, Markus et Christian Gütl. 2012. « On using JSON-LD to Create Evolvable RESTful Services ». Dans *Proceedings of the 3rd International Workshop on RESTful Design* : 25-32. Lyon, France : Association for Computing Machinery. https://doi.org/10.1145/2307819.2307827.

Lascar, Alex. 2010. « Balzac et Sue : échanges à feuilletons mouchetés ». *L'Année balzacienne* 11 (1) : 201-221. https://doi.org/10.3917/balz.011.0201.

Le Bœuf, Patrick. 2013. *Functional Requirements for Bibliographic Records (FRBR): Hype or Cure-all?* New York, États-Unis : Routledge.

Le Deuff, Olivier. 2012. « Humanisme numériques et littératies ». *Semen. Revue de sémio-linguistique des textes et discours* (34 [novembre]). https://doi.org/10.4000/semen.9752.

—. 2018. *Les Humanités digitales : historique et développements*. Vol. 5. Londres, Royaume-Uni : ISTE éditions.

Le Fourner, Victoria. 2019. « Étude de la structuration automatique et de l'éditorialisation d'un corpus hétérogène ». Thèse de master, École nationale des chartes.

Lebart, Ludovic, Bénédicte Pincemin et Céline Poudat. 2019. *Analyse des données textuelles*. Montréal, Canada : Presses de l'université du Québec.

Lécossais, Sarah et Nelly Quemener. 2018. *En quête d'archives. Bricolages méthodologiques en terrains médiatiques*. Bry-sur-Marne, France : INA Publications.

Legallois, Dominique. 2006. « Présentation générale. Le texte et le problème de son et ses unités : propositions pour une déclinaison ». *Langages* 163 (3) : 3-9. https://doi.org/10.3917/lang.163.0003.

Lemercier, Claire. 2007. « Juges du commerce et conseillers prud'hommes face à l'ordre judiciaire (1800-1880). La constitution de frontières judiciaires ». Dans *La Justice au risque des profanes*, édité par Hélène Michel et Laurent Willemez : 11-27. Paris, France : Presses universitaires de France. https://halshs.archives-ouvertes.fr/halshs-00461826.

—. 2014. « Les grand corpus en ligne changeront-ils la boîte à outils de l'historien.ne ». Communication présentée à *Data, digital methods and mapping social complexity. A research seminar on mapping social and semantic dynamics in the social sciences*, Paris, France, 27 mars. https://www.dailymotion.com/video/x1p1zvs.

— et Claire Zalc. 2008. *Méthodes quantitatives pour l'historien*. Repères. Paris, France : La Découverte. https://www.cairn.info/methodes-quantitatives-pour-l-historien--9782707153401.htm.

Lepetit, Bernard. 1989. « L'histoire quantitative : deux ou trois choses que je sais d'elle ». *Histoire & Mesure* 4 (3-4) : 191-199. https://doi.org/10.3406/hism.1989.1355.

Lepetit, Bernard. 1990. « Propositions pour une pratique restreinte de l'interdisciplinarité ». *Revue de synthèse* 111 (3) : 331-338. https://doi.org/10.1007/BF03181048.

—. 1993. « Pour une nouvelle histoire sociale : Journées de réflexion du CRH. Paris, 14-15-16 octobre 1993 ». *Les Cahiers du Centre de recherches historiques* (11 [octobre]). https://doi.org/10.4000/ccrh.2775.

—. 1999. *Carnet de croquis : Sur la connaissance historique*. Paris, France : Albin Michel.

—. 2013. *Les Formes de l'expérience une autre histoire sociale*. Nouvelle édition augmentée. Paris, France : Albin Michel.

Les Études sociales : Les monographies de familles de l'École de Le Play (1855-1930). 2000. Vol. 131-132. Paris, France : Société d'économie et de science sociales. http://lascienceso-ciale.org/20001-2-(n%C2%B0-131-132).html.

Lesage, Frederik et Simone Natale. 2019. « Rethinking the Distinctions between Old and New Media: Introduction ». *Convergence* 25 (4) : 575-589. https://doi.org/10.1177/1354856519863364.

Lewi, Hannah, Wally Smith, Dirk vom Lehn et Steven Cooke. 2019. *The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites*. Routledge international handbooks. Londres, Royaume-Uni : Routledege. https://doi.org/10.4324/9780429506765.

Limelette, Renaud. 2009. « À la recherche de son juge dans le ressort du parlement de Flandre ». *C@hiers du CRHIDI. Histoire, droit, institutions, société* 31 : 29-46.

—. 2018. « Recherche sur le conseiller-commissaire au parlement de Flandre ». *C@hiers du CRHIDI. Histoire, droit, institutions, société* 41 (décembre). https://popups.uliege.be:443/1370-2262/index.php?id=595.

—. 2020. « La gouvernance du service de santé des hôpitaux militaires, de la réforme de 1747 à 1789 ». Dans *Gouvernance, justice et santé* : 145-174. Lille, France : Centre d'histoire judiciaire de Lille. https://hal.archives-ouvertes.fr/hal-01688606.

— et Sabrina Michel. 2014. « L'affaire est dans la base ! L'exemple du contentieux du parlement de Flandre (1668-1790) ». Dans *L'Affaire est dans le sac ! Dossiers de procès d'Ancien Régime et perspectives de recherche historique*, édité par Harald Deceulaer, Sébastien Dubois et Laetizia Puccio, 148 : 131-152. Bruxelles, Belgique : Archives générales du Royaume. https://doi.org/10.1108/JD-06-2017-0095.

Longrée, Dominique et Sylvie Mellet. 2013. « Le motif : une unité phraséologique englobante ? Étendre le champ de la phraséologie de la langue au discours ». *Langages* 189 (1) : 65-79. https://doi.org/10.3917/lang.189.0065.

López, Susana. 1993. « The cultural policy of the European community and its influence on museums ». *Museum Management and Curatorship* 12 (2) : 143-157. https://doi.org/10.1080/09647779309515353.

Loriou, Céline. 2018. « Faire de l'histoire, un casque sur les oreilles : le goût de l'archive radiophonique ». *Le Goût de l'archive à l'ère numérique* (blog). 27 mars 2018. http://www.gout-numerique.net/table-of-contents/faire-de-lhistoire-un-casque-sur-les-oreilles-le-gout-de-larchive-radiophonique.

Lough, John. 1968. *Essays on the Encyclopédie of Diderot and D'Alembert*. Londres, Royaume-Uni : Oxford University Press.

MacDonald, George. 1987. « The future of museums in the Global Village ». *Museum International* 39 (3) : 209-216. https://doi.org/10.1111/j.1468-0033.1987.tb00695.x.

— et Stephen Alsford. 1989. « Museums as bridges to the Global Village ». Dans *A Different Drummer: Readings in Anthropology with a Canadian Perspective*, édité par Bruce Alden Cox, Jacques Chevalier et Valda Blundell : 41-48. McGill-Queen's University Press. https://www.jstor.org/stable/j.ctt7zt44m.

— et Stephen Alsford. 1991. « The museum as information utility ». *Museum Management and Curatorship* 10 (3) : 305-311. https://doi.org/10.1080/09647779109515282.

Magallon, Thibault, Frédéric Béchet et Benoit Favre. 2018. « Détection d'erreurs dans des transcriptions OCR de documents historiques par réseaux de neurones récurrents multi-niveau ». Dans *TALN 2018 : 25e conférence sur le Traitement Automatique des Langues Naturelles*. Vol. 1. Atala. https://hal.archives-ouvertes.fr/hal-01905258/document.

Magnani, Eliana. 2017. « Qu'est-ce qu'un corpus ? » *Les carnets de l'IRHT* (blog). 2 octobre 2017. https://irht.hypotheses.org/3187.

Maguire, Rob. 2016. « What's the Future of Canada's Museums? » *Canadian Art*, 8 mars 2016. https://canadianart.ca/features/whats-the-future-of-canadas-museums/.

Maingueneau, Dominique. 1996. *Les Termes clés de l'analyse du discours*. Paris, France : Seuil.

Mak, Bonnie. 2014. « Archaeology of a Digitization ». *Journal of the Association for Information Science and Technology* 65 (8) : 1515-1526. https://doi.org/10.1002/asi.23061.

Malraux, André. 1977. *L'Homme précaire et la Littérature*. Blanche. Paris : Gallimard.

Malrieu, Denise et François Rastier. 2001. « Genres et variations morphosyntaxiques ». *Traitement Automatique des langues* 42 (2) : 548-577.

Maron, Deborah et Melanie Feinberg. 2018. « What Does It Mean to Adopt a Metadata Standard? A Case Study of Omeka and the Dublin Core ». *Journal of Documentation* 74 (4) : 674-691. https://doi.org/10.1108/JD-06-2017-0095.

Martini, Manuela. 2021. « Pratiques de la réclamation du prix du travail : différends autour des rémunérations des tisseurs et des tisseuses en soie de Lyon au début des années 1830 ». *Parlement[s], Revue d'histoire politique* 33 (1) : 61-78. https://doi.org/10.3917/parl2.033.0061.

— et Anaïs Albert. 2021. « Les mutations du textile et de ses mains-d'œuvre (fin XVIIIe siècle-années 1930) ». *Le Mouvement Social* 276 (3) : 3-15. https://doi.org/10.3917/lms1.276.0003.

Marty, Paul et Katherine Burton Jones (éd.). 2007. *Museum informatics. People, information, and technology in museums*. Routledge studies in library and information science 2. New York, États-Unis : Routledge. https://doi.org/10.4324/9780203939147.

Mayaffre, Damon. 2002. « Les corpus réflexifs : entre architextualité et hypertextualité ». *Corpus* (1) : 51-69.

—. 2007a. « L'analyse de données textuelles aujourd'hui : du corpus comme une urne au corpus comme un plan ». *Lexicometrica* (n° Spécial) : 1-12.

—. 2007b. « Philologie et/ou herméneutique numérique : nouveaux concepts pour de nouvelles pratiques ? » Dans

Corpus en Lettres et Sciences sociales. Des documents numériques à l'interprétation, édité par François Rastier et Michel Ballabriga : 15-25. Toulouse, France : Centre pluridisciplinaire de sémiolinguistique textuelle. https://hal.archives-ouvertes.fr/hal-00551477.

—. 2008a. « De l'occurrence à l'isotopie : Les co-occurrences en lexicométrie ». *Syntaxe et sémantique* 9 (1) : 53-72. https://doi.org/10.3917/ss.009.0053.

—. 2008b. « L'entrelacement lexical des textes. Cooccurrences et lexicométrie ». Dans *Texte et corpus : Actes des Journées de la linguistique de Corpus 2007*, 3 : 91-102. Lorient, France : Groupe de recherche en lexicographie, corpus et ressources numériques (Licorn). https://hal.archives-ouvertes.fr/hal-00553808.

—. 2008c. « Quand "travail", "famille", "patrie" co-occurrent dans le discours de Nicolas Sarkozy. Étude de cas et réflexion théorique sur la co-occurrence ». Dans *JADT 2008 : 9e Journées d'Analyse statistique des Données Textuelles*, édité par Serge Heiden et Bénédicte Pincemin, 2 : 811-822. Lyon, France : Presses universitaires de Lyon. https://hal.archives-ouvertes.fr/hal-00551300.

—. 2010. « Vers une herméneutique matérielle numérique. Corpus textuels, logométrie et langage politique ». HDR, Nice, France : Université Nice Sophia Antipolis. https://tel.archives-ouvertes.fr/tel-00655380.

—. 2014. « Plaidoyer en faveur de l'Analyse de Données co(n) Textuelles. Parcours cooccurrentiels dans le discours présidentiel français (1958-2014) ». Dans *JADT 2014 : 12e Journées internationales d'Analyse statistique des Données Textuelles*, édité par Emilie Née, Jean-Michel Daube, Mathieu Valette et Serge Fleury : 15-32. Vincennes, France. https://hal.archives-ouvertes.fr/hal-01181337.

—. 2021. *Macron par l'intelligence artificielle*. La Tour-d'Aigues, France : Éditions de l'Aube.

—, Camille Bouzereau, Mélanie Ducoffe, Magali Guaresi, Frédéric Precioso et Laurent Vanni. 2017. « Les mots des candidats, de "allons" à "vertu" ». Dans *Le Vote disruptif. Les élections présidentielle et législatives de 2017*, édité par Pascal Perrineau : 129-152. Paris, France : Presses de Sciences Po. https://hal.archives-ouvertes.fr/hal-01635941.

—, Camille Bouzereau, Magali Guaresi, Frédéric Precioso et Laurent Vanni. 2020. « Du texte à l'intertexte. Le palimpseste Macron au révélateur de l'Intelligence artificielle ». Dans *SHS Web of Conferences : 7e Congrès Mondial de Linguistique Française*. Les Ulis, France : EDP Sciences. https://hal.archives-ouvertes.fr/hal-02520224.

—, Bénédicte Pincemin et Céline Poudat. 2019. « Explorer, mesurer, contextualiser. Quelques apports de la textométrie à l'analyse de discours ». *Langue française* 203 (3) : 101-115. https://doi.org/10.3917/lf.203.0101.

— et Laurent Vanni (éd.). 2021. *L'Intelligence artificielle des textes : Des algorithmes à l'interprétation*. Paris, France : Honoré Champion.

— et Jean-Marie Viprey. 2012. « La cooccurrence. Du fait statistique au fait textuel : présentation ». *Corpus* (11 [janvier]). http://journals.openedition.org/corpus/2200.

Mazeau, Guillaume. 2016. « Au-delà du cours magistral : quelques pistes de pédagogie universitaire ». *Aggiornamento hist-geo* (blog). 15 octobre 2016. https://aggiornamento.hypotheses.org/3553.

Mazzone, Jason. 2015. « Copyfraud ». *New York University Law Review* 81 (août) : 1026-1100.

Mellet, Sylvie. 2002. « Corpus et recherches linguistiques. Introduction ». *Corpus* (1 [novembre]). https://doi.org/10.4000/corpus.7.

— et Dominique Longrée. 2009. *New Approaches in Text Linguistics*. Amsterdam, Pays-Bas : John Benjamins Publishing Company.

Merzeau, Louise. 2012. « Faire mémoire de nos traces numériques ». *E-dossier de l'audiovisuel*, juin. https://halshs.archives-ouvertes.fr/halshs-00727308/.

—. 2014. « Vers un Web temporel ? Constituer des corpus pour la recherche contemporaine : de l'archivage du Web à son analyse ». Communication présentée à *Conférence du consortium international pour la préservation de l'internet (IIPC)*, Paris, France.

Messaoudi, Tommy. 2017. « Proposition d'une ontologie de domaine dédiée à l'annotation d'images spatialisées pour le suivi de la conservation du patrimoine culturel bâti ». Thèse de doctorat, Paris, France : École nationale supérieure d'arts et métiers. http://www.theses.fr/2017ENAM0021.

Metwally, Heba. 2017. « Les thèmes et le temps dans *Le Monde diplomatique* (1990-2008) ». Thèse de doctorat, Nice, France : Université Côte d'Azur. http://www.theses.fr/2017AZUR2042.

Michel, Jean-Baptiste, Yuan Kui Shen, Aviva Presser Aiden, Adrian Veres, Matthew Gray, Joseph Pickett, Dan Clancy, *et al.* 2011. « Quantitative Analysis of Culture Using Millions of Digitized Books ». *Science* 331 (6014) : 176-182. https://doi.org/10.1126/science.1199644.

Milligan, Ian. 2013. « Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997-2010 ». *Canadian Historical Review* 94 (4) : 540-569. https://doi.org/10.3138/chr.694.

—, Scott Weingart et Shawn Graham. 2015. *Exploring Big Historical Data: The Historian's Macroscope*. Londres, Royaume-Uni : Imperial College Press. https://doi.org/10.1142/p981.

Morardo, Mikaël et Éric Villemonte de la Clergerie. 2014. « Towards an Environment for the Production and the Validation of Lexical Semantic Resources ». Dans *LREC 2014: 9th International Conference on Language Resources and Evaluation*, édité par Nicoletta Calzolari. European Language Resources Association (ELRA). https://hal.inria.fr/hal-01005464.

Moretti, Franco. 2005. *Graphs, Maps, Trees: Abstract Models for Literary History*. Londres, Royaume-Uni : Verso.

—. 2013. *Distant Reading*. Londres, Royaume-Uni : Verso.

Moulin, Claudine. 2011. « Vom mittelalterlichen Griffel zum Computer-Tagging. Zur sprach- und kulturgeschichtlichen Bedeutung der Annotation ». Dans *Akademie der Wissenschaften und der Literatur, Mainz. Jahrbuch 2010*, 61 : 84-89. Stuttgart, Allemagne : Steiner.

Mounier, Pierre. 2015. « Une "utopie politique" pour les humanités numériques ? : Modèles de communication savante et de gestion de la recherche en transformation ». *Socio. La nouvelle revue des sciences sociales* (4 [avril]) : 97-112. https://doi.org/10.4000/socio.1451.

Mullen, Abby. 2016. « Untangling the Mess: Researchers' Photo Practices ». *Tropy* (blog). 1 novembre 2016. http://tropy.org/blog/untangling-the-mess-researchers-photo-practices/.

Muller, Caroline. 2017. « Autour d'une machine à café virtuelle. Twitter et les historien.nes ». *Le Goût de l'archive à l'ère numérique* (blog). 7 octobre 2017. http://www.gout-numerique.net/table-of-contents/autour-dune-machine-a-cafe.

—. 2018. « Le cours "de numérique" est un cours comme les autres ». *Acquis de conscience* (blog). 12 juillet 2018. https://consciences.hypotheses.org/1417.

Muller, Charles. 1967. *Étude de statistique lexicale : Le vocabulaire du théâtre de Pierre Corneille*. 1 vol. Paris, France : Larousse.

—. 1977. *Principes et méthodes de statistique lexicale*. Paris, France : Hachette.

Musen, Mark Alan, Karen Wieckert, Erika Thickman Miller, Keith Campbell et Lawrence Fagan. 1995. « Development of a Controlled Medical Terminology: Knowledge Acquisition and Knowledge Representation ». *Methods of Information in Medicine* 34 (1-2) : 85-95. https://doi.org/10.1055/s-0038-1634576.

Musiani, Francesca et Valérie Schafer. 2019. « Science and Technology Studies Approaches to Web History ». Dans *The SAGE Handbook of Web History*, par Niels Brügger et Ian Milligan : 73-85. Londres, Royaume-Uni : SAGE. https://halshs.archives-ouvertes.fr/halshs-02320717.

Musso, Pierre. 2002. « L'économie symbolique de la société d'information ». *Revue européenne des sciences sociales* 40 (123) : 91-113. https://doi.org/10.4000/ress.618.

Née, Émilie (éd.). 2017. *Méthodes et outils informatiques pour l'analyse des discours*. Didact Méthodes. Rennes, France : Presses universitaires de Rennes.

Nelzin-Santos, Anthony. 2019. « L'archive à l'épreuve de la mise en ligne, le chercheur à l'épreuve de l'archive numérisée. L'exemple des recensements canadiens en 1871 à 1916 ». Dans *Dans les dédales du web : Historiens en territoires numériques*, par Stéphane Lamassé et Gaëtan Bonnot : 47-55. Paris, France : Éditions de la Sorbonne.

Neuroth, Heike, Felix Lohmeier et Kathleen Marie Smith. 2011. « TextGrid. Virtual Research Environment for the Humanities ». *International Journal of Digital Curation* 6 (2) : 222-231. https://doi.org/10.2218/ijdc.v6i2.198.

Niu, Jinfang. 2012. « An Overview of Web Archiving ». *D-Lib Magazine* 18 (3/4). https://doi.org/10.1045/march2012-niu1.

Nora, Simon et Alain Minc. 1978. *L'informatisation de la société : Rapport à M. le président de la République*. Points politique 92. Paris, France : Seuil.

Olsen, Mark. 2008. « From Words to Works: Machine Learning and Text Mining at ARTFL ». Communication présentée à *Technological Innovation and Cooperation for Foreign Information Access (TICFIA) Annual Conference*, Chicago, États-Unis, 30 avril.

Paganoni, Maria Cristina. 2015. « Cultural Heritage in the Discourse of European Institutions ». *Languages Cultures Mediation* 2 (2) : 117-130. https://doi.org/10.7358/lcm-2015-002-paga.

Paloque-Bergès, Camille. 2014. « Le rôle des communautés patrimoniales d'Internet dans la constitution d'un patrimoine numérique : des mobilisations diverses autour de l'auto-médiation ». Dans *Heritage and Digital Humanities: How should training practices evolve?*, par Bernadette Nadia Saou-Dufrêne et Benjamin Barbier : 277-290. Zurich, Suisse : LIT.

—, Camille. 2017. « Usenet as a Web Archive. Multi-layered Archives of Computer-mediated Communication ». Dans *Web 25: Histories from the First 25 Years of the World Wide Web*, par Niels Brügger, 112 : 227-250. New York, États-Unis : Peter Lang.

—, Camille. 2018. *Qu'est-ce qu'un forum internet ? : Une généalogie historique au prisme des cultures savantes numériques*. OpenEdition Press. https://doi.org/10.4000/books.oep.1843.

Parikka, Jussi. 2018. *Qu'est-ce que l'archéologie des médias ?* Grenoble, France : UGA Éditions.

Parry, Ross. 2007. *Recoding the museum: Digital heritage and the technologies of change*. Museum meanings. Londres, Royaume-Uni : Routledge.

— (éd.). 2009. *Museums in a digital age*. Leicester readers in museum studies. Londres, Royaume-Uni : Routledge. https://doi.org/10.4324/9780203716083.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam, Pays-Bas : John Benjamins Publishing Company.

—, John Morley et Louann Haarman (éd.). 2004. *Corpora and Discourse*. Berlin, Allemagne : Peter Lang.

Passini, Michela. 2017. *L'Œil et l'archive une histoire de l'histoire de l'art*. Paris, France : La Découverte.

Pauli, Julien et Guillaume Ponçon. 2008. *Zend Framework. Bien développer en PHP.* Paris, France : Eyrolles.

Peccatte, Patrick. 2009. « Pour une diplomatique des images numériques ». *Du bruit au signal (et inversement)* (blog). 2 septembre 2009. http://blog.tuquoque.com/post/2009/09/02/Pour-une-diplomatique-des-images-num%C3%A9riques.

Petermann, Damien. 2018. « Réutiliser les images numériques des collections : enjeux, questions pratiques ». Communication présentée à *15e rencontres professionnelles de la Fédération des écomusées et musées de société (FEMS) : Le musée digital*, Saint-Sulpice-sur-Risle, France, 23 mars. https://hal.archives-ouvertes.fr/hal-01763577.

Peters, John Durham. 2008. « History as a Communication Problem ». Dans *Explorations in Communication and History*, édité par Barbie Zelizer : 19-34. Londres, Royaume-Uni : Routledge.

—. 2015. *The Marvelous Clouds: Toward a Philosophy of Elemental Media*. Chicago, États-Unis : University of Chicago Press.

Pierazzo, Elena. 2015. *Digital Scholarly Editing: Theories, Models and Methods*. Londres, Royaume-Uni : Routledge.

—. 2019. « How Subjective Is Your Model? » Dans *The Shape of Data in the Digital Humanities: Modeling Texts and Text-based Resources*, édité par Julia Flanders et Fotis Jannidis : 117-132. Londres, Royaume-Uni : Routledge.

Pinter, Andrej. 2003. « Thought News a Quest for Democratic Communication Technology ». *Javnost: The Public* 10 (2) : 93-104. https://doi.org/10.1080/13183222.2003.11008830.

Pirez-Huart, Stéphanie. 2018. « Du parchemin à l'octet : quelles pratiques de l'archive médiévale à l'ère des humanités numériques ? » *Le Goût de l'archive à l'ère numérique* (blog). 26 novembre 2018. http://www.gout-numerique.net/table-of-contents/du-parchemin-a-loctet-quelles-pratiques-de-larchive-medievale-a-lere-des-humanites-numeriques.

Poibeau, Thierry. 2005. « Sur le statut référentiel des entités nommées ». Dans *Conférence Traitement Automatique des Langues 2005*, édité par Michele Jardino, 173-183. Dourdan, France : ATALA. https://hal.archives-ouvertes.fr/hal-00009448.

Possompès, Julien. 2017. « La norme ISO 21127 au regard de l'offre logicielle d'A&A Partners : analyse prospective des besoins des musées et propositions d'évolution des solutions WebMuseo ». Mémoire de master Sciences de l'information et de la communication, CNAM-INTD. https://memsic.ccsd.cnrs.fr/mem_01723673.

Poublanc, Sébastien. 2018. « Les jeunes historiens rêvent-ils d'archives numériques ? » *Le Goût de l'archive à l'ère numérique* (blog). 5 avril 2018. http://www.gout-numerique.net/table-of-contents/les-historiens-numeriques-revent-ils-darchives-electroniques.

—. 2019. « Faire du "numérique" dans un TD d'histoire ». *Devenir historien-ne* (blog). 21 janvier 2019. https://devhist.hypotheses.org/3635.

Poupeau, Gautier. 2010. « XML vs RDF : logique structurelle contre logique des données ». *Les Petites cases* (blog). 29 août 2010. https://www.lespetitescases.net/xml-vs-rdf.

Powell, Walter et Kaisa Snellman. 2004. « The Knowledge Economy ». *Annual Review of Sociology* 30 (1) : 199-220. https://doi.org/10.1146/annurev.soc.29.010202.100037.

Prime-Claverie, Camille et Annaïg Mahé. 2017. « Le défi de l'interopérabilité entre plates-formes pour la construction de savoirs augmentés en sciences humaines et social ». Dans *Écrilecture augmentée dans les communautés scientifiques : humanités numériques et construction des savoirs*, édité par Gérald Kembellec et Évelyne Broudoux. Londres, Royaume-Uni : ISTE éditions.

Prost, Antoine. 1974. *Vocabulaire des proclamations électorales de 1881, 1885 et 1889*. Paris, France : Presses universitaires de France.

— et Christian Rosenzveig. 1971. « La Chambre des députés (1881-1885). Analyse factorielle des scrutins ». *Revue française de science politique* 21 (1) : 5-50. https://doi.org/10.3406/rfsp.1971.393277.

Proust, Jacques. 1995. *Diderot et l'Encyclopédie*. Réédition. Paris, France : Albin Michel.

Putnam, Lara. 2016. « The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast ». *The American Historical Review* 121 (2) : 377-402. https://doi.org/10.1093/ahr/121.2.377.

Rageot, Laurence. 2014. « Quel droit d'auteur pour les éditeurs de source ? » *Consortium CAHIER*. https://cahier.hypotheses.org/1090.

Rahtz, Sebastian et Lou Burnard. 2013. « Reviewing the TEI ODD system ». Dans *Proceedings of the 2013 ACM symposium on Document engineering*, 193-196. DocEng '13. New York, USA : Association for Computing Machinery. https://doi.org/10.1145/2494266.2494321.

Rastier, François. 2001. *Arts et sciences du texte*. 1re édition. Paris, France : Presses universitaires de France.

—. 2005. « Enjeux épistémologiques de la linguistique de corpus ». Dans *La linguistique de corpus*, édité par Geoffrey Williams : 31-45. Rennes, France : Presses universitaires de Rennes. http://www.revue-texto.net/Inedits/Rastier/Rastier_Enjeux.html.

—. 2011. *La Mesure et le grain. Sémantique de corpus*. Paris, France : Honoré Champion.

« Report of the Working Group on Internet Governance ». 2005. Bogis-Bossey, Suisse : Working Group on Internet Governance. https://www.wgig.org/docs/WGIGREPORT.pdf.

Robineau, Régis. 2016. « Comprendre IIIF et l'interopérabilité des bibliothèques numériques ». *Insula* (blog). 8 novembre 2016. https://insula.univ-lille3.fr/2016/11/comprendre-iiif-interoperabilite-bibliotheques-numeriques/.

Ruiz, Émilien. 2018. « Historien.ne.s numériques : gare au SSPQ ! » *La Boîte à outils des historien.ne.s* (blog). 15 mars 2018. https://www.boiteaoutils.info/2018/03/sspq/.

—. 2019. « #DHIHA8 Nous sommes à la croisée des chemins ! » *Devenir historien-ne* (blog). 9 juin 2019. https://devhist.hypotheses.org/3692.

Rutner, Jennifer et Roger Schonfeld. 2015. « Supporting the Changing Research Practices of Historians ». Livre blanc. Roy Rosenzweig Center for History and New Media. https://doi.org/10.18665/sr.22532.

Rygiel, Philippe. 2004. « Les sources de l'historien à l'heure d'Internet ». *Hypothèses* 7 (1) : 341-354. https://doi.org/10.3917/hyp.031.0341.

—. 2017. *Historien à l'âge numérique : Essai*. Villeurbanne, France : Presses de l'ENSSIB.

Salvador, Xavier-Laurent. 2019a. « La France va-t-elle en finir avec les humanités classiques ? » *Le Figaro*, 20 août 2019, sect. FigaroVox. https://www.lefigaro.fr/vox/culture/pourquoi-la-notion-d-humanites-numeriques-est-absurde-20190820.

—. 2019b. « Dans la réforme du bac, la définition ratée des Humanités Numériques ». *Le Huffington Post*, 29 août 2019, sect. Les Blogs. https://www.huffingtonpost.fr/entry/dans-la-reforme-du-bac-la-definition-ratee-des-humanites-numeriques_fr_5d6678fde4b063c341f8a3b1.

Sassatelli, Monica. 2002. « Imagined Europe: The Shaping of a European Cultural Identity Through EU Cultural Policy ». *European Journal of Social Theory* 5 (4) : 435-451. https://doi.org/10.1177/1368431022760513848.

Schafer, Valérie. 2015. « En construction : la fabrique française d'Internet et du Web dans les années 1990 ». HDR, Paris, France : Université Paris-Sorbonne.

—. 2018. « De la Wayback Machine à la bibliothèque : les différentes saveurs de l'archive du Web… ». *Le Goût de l'archive à l'ère numérique* (blog). 15 janvier 2018. http://www.gout-numerique.net/table-of-contents/de-la-wayback-machine-a-la-bibliotheque-les-differentes-saveurs-de-larchive-du-web.

—, Francesca Musiani et Marguerite Borelli. 2016. « Negotiating the Web of the Past: Web Archiving, Governance and STS ». *French Journal For Media Research* 6. http://frenchjournalformediaresearch.com/lodel/index.php?id=952.

— et Benjamin Thierry. 2015. « L'ogre et la toile. Le rendez-vous de l'histoire et des archives du web ». *Socio. La nouvelle revue des sciences sociales* (4 [avril]) : 75-95. https://doi.org/10.4000/socio.1337.

Scholliers, Peter. 1996. *Wages, Manufacturers, and Workers in the Nineteenth-century Factory: The Voortman Cotton Mill in Ghent*. Oxford, Royaume-Uni : Berg Publishers.

Schudson, Michael. 1981. *Discovering the News: A Social History of American Newspapers*. New York, États-Unis : Basic Books.

Schuh, Julien. 2017. « La réimpression dans la presse francophone du 19e siècle (G. Pinson, J. Schuh avec P.-C. Langlais) ». Billet. *Numapresse* (blog). 29 mars 2017. http://numapresse.hypotheses.org/40.

Schwab, Richard Nahum. 1984. *Inventory of Diderot's Encyclopédie: Plates*. Vol. 7. Oxford, Royaume-Uni : Voltaire Foundation.

—, Walter Rex et John Lough. 1971. *Inventory of the Encyclopédie*. Vol. 80, 83, 85, 91-93. 6 vol. Oxford, Royaume-Uni : Voltaire Foundation.

Schwerzmann, Katia. 2018. « Pour une réforme des humanités. La théorie des média selon N. Katherine Hayles ». *Acta Fabula* 19 (9). https://www.fabula.org:443/revue/document11620.php.

Seguin, Benoit. 2018a. « Making large art historical photo archives searchable ». Thèse de doctorat, Lausanne, Suisse : École polytechnique fédérale de Lausanne. https://infoscience.epfl.ch/record/261212?ln=fr.

—. 2018b. « The Replica Project: Building a visual search engine for art historians ». *XRDS* 24 (3) : 24–29. https://doi.org/10.1145/3186653.

Shannon, Claude. 1948. « A Mathematical Theory of Communication ». *The Bell System Technical Journal* 27 (4) : 623-656. https://doi.org/10.1002/j.1538-7305.1948.tb00917.x.

Shoemaker, Tyler. 2019. « Error Aligned ». *Textual Cultures* 12 (1) : 155-182. https://doi.org/10.14434/textual.v12i1.27153.

Shore, Cris. 2000. *Building Europe: The cultural politics of European integration*. Londres, Royaume-Uni : Routledge.

Sinatra, Michaël E., et Marcello Vitali-Rosati. 2014. « Chapitre 3. Histoire des humanités numériques ». Dans *Pratiques de l'édition numérique*. Sous la direction de Michael E. Sinatra, 49-60. Parcours numérique. Montréal : Presses de l'Université de Montréal. http://books.openedition.org/pum/317.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford, Royaume-Uni : Oxford University Press.

Sire, Guillaume. 2015. « Inclusion exclue : le code est un contrat léonin. Enquête sur la valeur technique et juridique du protocole robots.txt ». *Réseaux* 189 (1). https://doi.org/10.3917/res.189.0187.

Smit, Eefke, Jeffrey Van Der Hoeven et David Giaretta. 2011. « Avoiding a Digital Dark Age for Data: Why Publishers Should Care about Digital Preservation ». *Learned Publishing* 24 (1) : 35-49. https://doi.org/10.1087/20110107.

Smith, David et Ryan Cordell. 2018. « A Research Agenda for Historical and Multilingual Optical Character Recognition ». Northeastern University. https://repository.library.northeastern.edu/downloads/neu:m043p093w?datastream_id=content.

—, Ryan Cordell et Abby Mullen. 2015. « Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers ». *American Literary History* 27 (3) : E1-E15. https://doi.org/10.1093/alh/ajv029.

Sobréro, Aurélie. 2014. *Développer des services REST en Java. Échanger des données au format JSON*. Saint-Herblain, France : Éditions ENI.

Strauß, Tobias, Max Weidemann, Johannes Michael, Gundram Leifert, Tobias Grüning et Roger Labahn. 2018. « System Description of Citlab's Recognition & Retrieval

Engine for ICDAR2017 Competition on Information Extraction in Historical Handwritten Records ». Communication présentée à *ICDAR2017: Historical Handwritten Script Analysis*, Kyoto, Japon, 26 avril. http://arxiv.org/abs/1804.09943.

Suchecka, Karolina, et Victoria Le Fourner. 2021. « Pérenniser les méthodes et données d'un projet d'édition numérique. Exemple du projet ANR Phœbus eBalzac ». Article de colloque présenté à *DHNord 2021. Publier, partager, réutiliser les données de la recherche. Les data papers et leurs enjeux*, Lille. https://hal.archives-ouvertes.fr/hal-03576242.

Szabados, Anne-Violaine et Rosemonde Letricot. 2014. « L'ontologie CIDOC CRM appliquée aux objets du patrimoine antique ». Dans *Actes des 3e Journées d'Informatique et Archéologie de Paris (JIAP 2012)*, édité par Laurent Jacques Costa, François Djindjian et François Giligny. Archeologia e calcolatori. Borgo San Lorenzo, Italie : All'Insegna del Giglio. https://halshs.archives-ouvertes.fr/halshs-00752996.

Tessier, Georges. 1961. « La diplomatique ». Dans *L'Histoire et ses méthodes*, par Charles Samaran, 11 : 633-676. L'Encyclopédie de la Pléiade. Paris, France : Gallimard. https://www.cairn.info/l-histoire-et-ses-methodes--9782070104093-page-616.htm.

« The Berlin declaration on Open Access to Scientific Knowledge ». 2003. Max Planck Open Access. http://oa.mpg.de/lang/en-uk/berlin-prozess/berliner-erklarung/.

Tkacz, Nathaniel. 2015. *Wikipedia and the politics of openness*. Chicago, États-Unis : University of Chicago Press.

Tognini-Bonelli, Elena. 2001. *Corpus Linguistics at Work*. Amsterdam, Pays-Bas : John Benjamins Publishing Company.

Trudel, Dominique et Juliette De Maeyer. 2017. « L'unité de l'enquête et le pipeline de la connaissance : alliances entre journalistes et universitaires au prisme de la comparaison historique ». *Sur le journalisme, About journalism, Sobre jornalismo* 6 (2) : 42-55.

Valette, Mathieu. 2008. « À quoi servent les lexiques sémantiques généralistes ? Discussion et propositions ». *Cahiers du Cental* (5) : 43-58.

Valtysson, Bjarki. 2012. « Europeana. The digital construction of Europe's collective memory ». *Information, Communication & Society* 15 (2) : 151-170. https://doi.org/10.1080/1369118X.2011.586433.

Van Nederveen Meerkerk, Elise. 2006. « Segmentation in the Pre-industrial Labour Market: Women's Work in the Dutch Textile Industry, 1581-1810 ». *International Review of Social History* 51 (2) : 189-216. https://doi.org/10.1017/S0020859006002422.

—, Elise et Ariadne Schmidt. 2008. « Between Wage Labor and Vocation: Child Labor in Dutch Urban Industry, 1600-1800 ». *Journal of Social History* 41 (3) : 717-736. https://doi.org/10.1353/jsh.2008.0041.

Villemonte de la Clergerie, Éric, Benoît Sagot, Lionel Nicolas et Marie-Laure Guénot. 2009. « FRMG : évolutions d'un analyseur syntaxique TAG du français ». Dans *Journée de l'Atala : Quels analyseurs syntaxiques pour le français ?* Atala. https://hal.inria.fr/inria-00553260.

Villey, Michel. 1995. *Réflexions sur la philosophie et le droit : Les Carnets de Michel Villey*. Édité par Marie-Anne Fri-

son-Roche et Christophe Jamin. 1re édition. Paris, France : Presses universitaires de France.

Viprey, Jean-Marie. 1997. *Dynamique du vocabulaire des Fleurs du mal*. Paris, France : Honoré Champion.

—. 2004. « Analyse séquencée de la micro-distribution lexicale ». Dans *JADT04 : Le poids des mots*, 2 : 1165-1175. Louvain-la-Neuve, Belgique : Presses universitaires de Louvain.

—. 2005a. « Corpus et sémantique discursive : éléments de méthode pour la lecture des corpus ». Dans *Sémantique et corpus*, par Anne Condamines : 245-276. Paris, France : Lavoisier.

—. 2005b. « Philologie numérique et herméneutique intégrative ». Dans *Sciences du texte et analyse de discours*, par Jean-Michel Adam et Ute Heidmann : 51-68. Genève, Suisse : Slatkine.

—. 2006. « Structure non-séquentielle des textes ». *Langages* 40 (163) : 71-85. https://doi.org/10.3406/lgge.2006.2684.

Vitali-Rosati, Marcello. 2018. « Les chercheurs en SHS savent-ils écrire ? » *The Conversation*, 11 mars 2018, sect. Culture numérique : Pour une philosophie du numérique. http://theconversation.com/les-chercheurs-en-shs-savent-ils-ecrire-93024.

Vlachou, Maria. 2018. « That's mine too! » *ICOM (International Committee for Documentation)* (blog). Septembre 2018. http://network.icom.museum/cidoc/blog/maria-vlachou/.

Welger-Barboza, Corinne. 2001. *Du musée virtuel au musée médiathèque : le patrimoine à l'ère du document numérique*. Patrimoines et sociétés. Paris, France : L'Harmattan.

Whaling, Richard. 2010. « High Performance XML: Anatomy of a Search Engine ». Communication présentée à *Computation Institute*, Chicago, États-Unis, juin.

Williams, Geoffrey (éd.). 2005. *La Linguistique de corpus*. Rennes, France : Presses universitaires de Rennes.

Winner, Langdon. 1980. « Do Artifacts Have Politics? » *Daedalus* 109 (1) : 121-136.

Witcomb, Andrea. 2006. « Interactivity. Thinking Beyond ». Dans *A Companion to Museum Studies*, édité par Sharon Macdonald : 353-361. Blackwell Publishing. https://doi.org/10.1002/9780470996836.ch21.

Wybo, Mathilde. 2017. « Recherche collaborative pour une cité régionale de l'histoire des gens du textile (RECCITS) ». Rapport de recherche. Institut de Recherches Historiques du Septentrion. https://reccits.hypotheses.org/541.

Zacklad, Manuel. 2010. « Évaluation des systèmes d'organisation des connaissances ». *Les Cahiers du numérique* 6 (3) : 133-166. https://doi.org/10.3166/lcn.6.3.133-166.

Biographies

1. Ingénieure d'études CNRS, **Céline Alazard** est archiviste et responsable de la plateforme technologique Archives-Documentation-Numérisation (ADN) de la MSH de Dijon.

2. **Anaïs Albert** est maîtresse de conférences en histoire contemporaine à l'université de Paris-Cité et membre du laboratoire de recherche Identités, Cultures, Territoires (ICT). Elle est spécialiste d'histoire économique et sociale, ainsi que d'histoire du genre et des classes populaires dans la France du XIX^e^ siècle. Après avoir travaillé sur la consommation et le crédit, ses travaux portent aujourd'hui sur les rapports économiques au sein des familles populaires et sur l'économie domestique. Elle a publié notamment *La vie à crédit. La consommation des classes populaires à Paris (1880-1920)* aux Éditions de la Sorbonne en 2021.

3. **Clarisse Bardiot** est professeur à l'Université Rennes 2 en études théâtrales. Ses axes de recherche concernent les humanités numériques, l'histoire et l'esthétique des digital performances, les relations art/science/technologie, les traces numériques des arts de la scène et les processus de création. Elle développe des environnements numériques, dont l'application web d'annotation vidéo MemoRekall, et codirige avec Émilien Ruiz la collection Humanités numériques et science ouverte aux PUS.

4. Archiviste et diplomatiste, **Marie-Anne Chabin**, exerce comme professeure associée à l'université Paris 8. Diplômée de l'École nationale des chartes, elle a été successivement conservatrice à la Direction des archives de France (DAF), consultante en gestion électronique de documents et responsable de la vidéothèque d'actualités de l'Institut national de l'audiovisuel (INA). Elle a créé son propre cabinet de conseil, Archive 17, et exerce aujourd'hui ses activités de conseil dans le domaine de l'archivage et de la gouvernance de l'information.

5. **Alix Chagué** est ingénieure de recherche et développement à l'INRIA au sein de l'équipe projet ALMANACH et coordinatrice du master Documentation et humanités numériques de l'École nationale des chartes. Ses travaux de recherche portent sur le traitement automatique des documents manuscrits, la patrimonialisation des données et sur l'archivistique computationnelle. Elle est issue d'une formation en histoire de l'art contemporain et en *gender studies*, et a suivi par la suite des études en humanités numériques à l'École nationale des chartes.

6. **Emmanuel Château-Dutier** est historien de l'architecture et professeur adjoint en muséologie numérique à l'université de Montréal. Ses travaux s'inscrivent notamment dans le champ de l'histoire de l'art numérique et plus largement des humanités numériques. Ses recherches portent sur l'administration de l'architecture

publique en France au XIXe siècle, la profession d'architecte et les relations entre la maîtrise d'ouvrage et la maîtrise d'œuvre, l'architecture des jardins zoologiques, l'édition et le livre d'architecture.

7. Professeur assistant en histoire contemporaine, **Frédéric Clavert** est chercheur au Center for Contemporary and Digital History (C²DH) de l'Université du Luxembourg. Ses travaux portent sur les sources de l'historien à l'ère numérique ainsi que sur les relations entre mémoire collective et réseaux sociaux numériques. Avec Caroline Muller, il codirige le livre en ligne *Le goût de l'archive à l'ère numérique*. Il est par ailleurs *managing editor* du *Journal of Digital History*.

8. Maître de conférences associé à l'université Paris-Nanterre et rattaché au laboratoire DICEN, **Antoine Courtin** a rejoint depuis septembre 2021 la direction de la conservation et des collections au sein de l'Établissement public des musées d'Orsay et de l'Orangerie – Valéry Giscard d'Estaing pour la mise en œuvre de la stratégie et du développement numérique du futur Centre de ressources et de recherche. De 2015 à 2021, il fut responsable à l'Institut national d'histoire de l'art de la réalisation et de la gestion des systèmes d'information documentaires, puis de la mise en œuvre du service numérique de la recherche dont il devint le chef de service en 2019. Ses travaux portent ainsi depuis plus de 10 ans sur l'articulation entre les nouvelles technologies et le monde du patrimoine, au sein d'institutions telles que les services d'archives, les musées ou encore les bibliothèques.

9. Doctorante en sciences de l'information et de la communication au laboratoire GERIICO de l'université de Lille, **Amélie Daloz** participe au projet ANR MémoMines dédié à la valorisation et à la médiation numérique de la mémoire minière dans les Hauts-de-France. Elle travaille dans ce cadre sur la conception et l'analyse de systèmes d'organisation des connaissances pour le Web sémantique, et a notamment participé à la construction du thésaurus du patrimoine minier ThesoMines.

10. Diplômée de l'École du Louvre ainsi que de l'École nationale des chartes, **Johanna Daniel** est chargée d'études et de recherche à l'Institut national d'histoire de l'art (INHA) depuis octobre 2019. Elle travaille au sein du Service numérique de la recherche sous la direction d'Antoine Courtin. Rattachée au laboratoire de recherches historiques Rhône-Alpes (LARHRA), elle est doctorante à l'université Lumière-Lyon 2 et prépare une thèse sur les vues d'optique sous la direction de Sophie Raux et d'Emmanuel Château-Dutier.

11. **Andrea Del Lungo** est professeur de littérature française et humanités numériques à Sorbonne Université et membre de l'Institut universitaire de France (IUF). Ses domaines de recherche sont la théorie de la littérature ; les relations entre littérature et savoir dans une perspective croisant épistémologie, sémiologie et sociologie ; ainsi que la poétique et l'histoire du roman

moderne. Il dirige le projet Phœbus-eBalzac, financé par l'Agence nationale de la recherche (ANR), d'édition électronique et hypertextuelle de l'œuvre d'Honoré de Balzac.

12. **Odile Gaultier-Voituriez** est responsable du département archives de la Direction des ressources et de l'information scientifique de Sciences Po ainsi qu'enseignante à Sciences Po. Elle est également chercheuse associée au Centre de recherches politiques de Sciences Po (CEVIPOF). Membre du Conseil supérieur des archives, elle participe au comité de suivi du portail FranceArchives, au conseil scientifique de l'Institut Marc Sangnier et anime le partenariat de Sciences Po avec les Archives nationales. Elle contribue en outre à plusieurs groupes de travail transversaux au sein de Sciences Po.

13. Ancien chef de projets informatiques et documentaires pour diverses entreprises et institutions en France (AJLSM, INALCO, École nationale des chartes, ENS-Cachan), **Frédéric Glorieux** travaille depuis 2013 en tant qu'ingénieur de recherche en humanités numériques pour le Labex OBVIL, spécialisé dans la numérisation de corpus littéraires. Expert en philologie numérique, depuis la conception de schémas et modèles de documents jusqu'à l'extraction et l'exploitation des données pour un objectif scientifique, il conduit plusieurs projets de logiciels libres pour la production et la publication des corpus.

14. Maître de conférences à Sorbonne Université, **Alexandre Guilbaud** est membre de l'Institut de mathématiques de Jussieu-Paris Rive Gauche (CNRS, Sorbonne Université, Université de Paris). Une partie de ses travaux de recherche porte sur l'histoire des sciences mathématiques et physico-mathématiques et leurs interactions au XVIII^e siècle ; l'autre s'inscrit dans le champ des humanités numériques. Il rejoint en 2005 le groupe d'édition des *Œuvres complètes* de D'Alembert et participe depuis 2011 au développement puis à la coordination du projet d'*Édition numérique collaborative et critique de l'*Encyclopédie *de Diderot et D'Alembert*.

15. **Solenn Huitric** est maîtresse de conférences en sciences de l'éducation et de la formation à l'université Lumière Lyon 2 et membre du laboratoire ECP. Elle est porteuse du projet de Bibliothèque historique de l'éducation (BHE) ainsi que membre du comité d'histoire du ministère de l'Éducation nationale. Ses recherches portent sur la prise en charge par l'État de l'enseignement au XIX^e siècle ainsi que sur le rôle des chefs d'établissements. Elle a été notamment chargée d'études au Laboratoire de recherches historiques Rhône-Alpes (LARHRA) ainsi qu'à l'Institut français de l'éducation (IFE).

16. Maître de conférences en sciences de l'information et de la communication à l'université de Lille, **Eric Kergosien** est membre du laboratoire de recherche GERIICO. Titulaire d'un doctorat en informatique, ses travaux s'inscrivent dans le domaine de l'analyse des pratiques informationnelles, l'appropriation des dispo-

sitifs numériques et l'organisation des connaissances. Il dirige ou participe à plusieurs projets de recherche sur la valorisation du patrimoine, la représentation du territoire et la gestion de l'information scientifique et technique (IST).

17. Ingénieure d'études au CNRS, **Victoria Le Fourner** est chargée du traitement des données scientifiques à la MESHS. Elle est issue d'une formation en humanités numérique à l'École nationale des chartes. Elle a notamment participé au projet *TIME-US* où elle a développé une partie des traitements de structuration et d'annotation sémantique du corpus.

18. Ingénieur d'études en analyses de sources, **Renaud Limelette** a poursuivi des études juridiques au sein de la Faculté des sciences juridiques, politiques et sociales de Lille, où il a reçu une formation d'historien du droit. Il s'est intéressé plus particulièrement aux sources juridiques d'Ancien Régime ainsi qu'à l'histoire du droit militaire. Il conduit et coordonne les activités du pôle numérique au sein du Centre d'histoire judiciaire (UMR 8025, Université de Lille, CNRS) en guidant les chercheurs du laboratoire à travers les recommandations du Plan national pour la science ouverte.

19. Professeure d'histoire contemporaine à l'université Lyon 2, **Manuela Martini** est membre du laboratoire de recherche historique Rhône-Alpes (LARHRA) et coordinatrice du programme de recherche ANR *TIME-US*. Ses travaux portent notamment sur l'histoire des migrations européennes du XIX^e^ au XX^e^ siècle, le travail indépendant et les petites entreprises en France au XX^e^ siècle ainsi que la famille et les rapports intergénérationnels et de genre en Europe à l'époque contemporaine. Manuela Martini est également membre depuis 2020 de l'Institut universitaire de France (IUF).

20. Historien de formation, **Damon Mayaffre** est chargé de recherche en linguistique au CNRS et chargé de cours à l'Université Côte d'Azur. Il s'est spécialisé dans l'analyse du discours politique contemporain grâce à la logométrie et l'intelligence artificielle. Ses travaux portent ainsi sur le discours présidentiel, la linguistique textuelle, la construction du sens en corpus ainsi que la philologie et l'herméneutique numériques. En 2021, il publie son cinquième ouvrage, *Macron ou le mystère du verbe. Ses discours décryptés par la machine* (éd. de l'Aube).

21. Diplômée d'un doctorat en information et communication, **Juliette De Maeyer** est aujourd'hui professeure agrégée au département de communication de l'université de Montréal. Ses travaux de recherche s'inscrivent en partie dans le champ de l'épistémologie du journalisme et portent sur l'intersection entre journalisme et nouvelles technologies, la matérialité des processus de production journalistique ainsi que sur les discours métajournalistiques, souvent dans des approches historiques. Elle est membre du Centre de recherche interuniversitaire sur les humanités numériques (CRIHN) et est cofondatrice du collectif interdisciplinaire Paperology.

22. Agrégée d'histoire, **Caroline Muller** est maîtresse de conférences en histoire contemporaine à l'université Rennes 2 et spécialisée en histoire du genre et histoire du catholicisme au XIXe siècle en France. Elle tient un carnet de recherche en ligne intitulé *Acquis de conscience – Histoire(s) du XIXe siècle* dans lequel elle développe des réflexions à partir de ses thèmes de recherche ainsi que sur les pratiques du numérique dans les sciences humaines. Avec Frédéric Clavert, elle codirige le livre en ligne *Le goût de l'archive à l'ère numérique*.

23. Chargée de recherche au CNRS, **Francesca Musiani** est cofondatrice et directrice adjointe du Centre Internet et société (CIS). Ses travaux portent sur la gouvernance d'Internet et des archives du Web ainsi que le développement et usages des technologies de chiffrement dans les outils de messagerie. Docteure en socioéconomie de l'innovation à MINES ParisTech, elle est également chercheuse associée au Centre de sociologie de l'innovation. Ses recherches théoriques explorent ainsi les approches STS (science and technology studies) à la gouvernance d'Internet, avec une attention particulière aux controverses sociotechniques et à la gouvernance « par l'architecture » et « par l'infrastructure ».

24. **Émilien Ruiz** est professeur assistant à Sciences Po (CHSP), en détachement de l'université de Lille (IRHiS). Ses travaux portent principalement sur les relations entre savoirs et pouvoirs depuis la fin du XIXe siècle. Il s'intéresse également à l'histoire des sciences sociales, notamment aux transformations numériques et mutations de l'écriture de l'histoire. Membre du comité de rédaction et de direction des revues *Le Mouvement social* et *Humanités numériques*, il coanime avec Franziska Heimburger le blog *La boîte à outils des historien.ne.s* et codirige avec Clarisse Bardiot la collection *Humanités numériques et science ouverte* aux PUS.

25. Archiviste paléographe, **Agathe Sanjuan** est conservatrice archiviste de la Bibliothèque-Musée de la Comédie-Française. Elle participe régulièrement aux publications et travaux de recherche portant sur l'histoire du théâtre, ainsi qu'au projet en humanités numériques sur les registres de la Comédie-Française associant plusieurs universités françaises, américaines et canadiennes.

26. **Karolina Suchecka** est doctorante contractuelle en littérature comparée au laboratoire ALITHILA de l'université de Lille, travaillant sur la réécriture et l'édition intermédiale en explorant les possibilités des outils informatiques et de traitement automatique des langues. Chargée de recherche en informatique éditoriale au sein du projet ANR Phœbus-eBalzac, elle a participé au développement d'outils de repérage intertextuel qui visent à mettre en résonance l'ensemble de l'œuvre balzacienne avec un corpus d'écrits contemporains.

27. Professeur d'histoire contemporaine à l'université de Bourgogne–Franche-Comté (Dijon), **Jean Vigreux** est agrégé, docteur en histoire et a soutenu en 2007 une habilitation à diriger des recherches. Il travaille au sein du LIR3S (Laboratoire interdisciplinaire de recherche « So-

ciétés, Sensibilités, Soin ») sur le communisme, l'histoire du temps présent et la prise en compte de la ruralité. Il est également directeur de la MSH de Dijon et membre du bureau du RnMSH. Il a dirigé l'ANR Paprik@2F.

28. Chercheur à l'INRIA et membre de l'équipe ALMANACH, **Éric Villemonte De La Clergerie** a été responsable du projet de recherche ATOLL (Atelier d'outils logiciels pour le langage naturel) et membre d'ALPAGE (Analyse linguistique profonde à grande échelle). Ses travaux s'inscrivent dans le domaine du traitement automatique des langues (TAL) et portent notamment sur l'analyse syntaxique (symbolique, statistique et neuronale), l'ingénierie grammaticale ainsi que l'extraction d'informations et l'acquisition de connaissances.

29. **Serge Wolikow** est licencié de philosophie, agrégé d'histoire, docteur d'état, et aujourd'hui professeur émérite d'histoire contemporaine à l'université de Bourgogne. Spécialiste de l'histoire politique du mouvement ouvrier, il a effectué diverses missions internationales comme chercheur et expert français concernant les archives du communisme. Il a dirigé la Maison des sciences de l'homme (MSH) de Dijon dont il a été l'un des initiateurs, et a été par la suite président du réseau national des MSH. Il a été le coordinateur scientifique du Consortium Archives des mondes contemporains de la TGIR Huma-Num. Depuis 2018 il est chargé de mission pour les plateformes technologiques du réseau national des MSH.

30. **Mathilde Wybo** est ingénieure d'études en communication scientifique à la Maison européenne des sciences de l'Homme et de la société et membre associée au laboratoire IRHIS (Institut de recherches historiques du Septentrion). Elle a notamment coordonné les projets RECCITS (Recherche collaborative pour une cité régionale de l'histoire des gens du textile) et DENIM (Données numériques, langages et représentations du patrimoine textile en région Hauts-de-France).

Glossaire

ALTO : *Analysed layout and text object*

1. ALTO est un standard XML permettant de rendre compte de la mise en page physique et de la structure logique d'un texte transcrit par reconnaissance optique de caractères (OCR).

API : *Application programming interface*

2. Le rôle d'une API (en français, une interface de programmation applicative) est de permettre à une entité informatique d'agir avec ou sur un système tiers. Cette interaction se fait de manière normalisée en respectant les contraintes d'accès définies par le système tiers, grâce à une bibliothèque d'outils (mettant à disposition des fonctions, des programmes, etc.).

ARK : *Archival ressource key*

3. Créé en 2001 par la California Digital Library, le système ARK est un format d'identifiants pérennes basé sur la norme URI garantissant l'identification d'une ressource sur le long terme. Les ARK peuvent désigner des ressources de tous types : aussi bien des objets physiques, comme des livres ou des statues ; que des objets numériques, comme des objets textuels, des images, logiciels, des sites web ; voire même des concepts immatériels.

CER : *Character error rate*

4. Le taux d'erreur de caractères mesure le nombre de caractères mal transcrits dans un texte par rapport à une version correcte de la transcription.

CMS : *Content management system*

5. Un système de gestion de contenu est un logiciel permettant la conception et la mise à jour dynamique de sites web (par exemple : Wordpress, Drupal, Omeka).

Computer vision

6. La vision par ordinateur, ou vision artificielle, est un champ d'étude et d'ingénierie en intelligence artificielle visant à permettre l'acquisition, le traitement ainsi que l'analyse automatique d'images numériques. L'OCR, ou reconnaissance optique de caractères, fait notamment partie des champs d'application de la vision par ordinateur.

CQL : *Contextual query langage*

7. Le CQL est un langage formel pour la récupération d'informations permettant d'exprimer des requêtes à destination de systèmes de recherche. Ces requêtes s'expriment à travers des expressions construites sur la base de chaîne de caractères représentant un motif linguistique défini en fonction de formes graphiques, lemmatisées ou de catégories grammaticales.

Crawling

8. L'exploration automatique sur Internet est un processus réalisé par un robot d'indexation, conçu pour collecter des ressources (pages web, images, vidéos, etc.) et les classer selon leur pertinence afin de permettre à un moteur de recherche de les indexer.

Crowdsourcing

9. Le *crowdsourcing* – ou production participative – est une forme d'externalisation d'une activité vers un grand nombre d'acteurs anonymes. Son essor est fortement lié au développement des nouvelles technologies du Web 2.0, qui facilite la mise en relation d'un grand nombre d'acteurs dispersés. Il se caractérise par un appel ouvert, non discriminatoire, auquel une foule d'individus plus ou moins hétérogène vient répondre.

CSV : *Comma-separated values*

10. Le CSV est un format de stockage de données tabulaires sous forme de texte brut. Les valeurs y sont ainsi séparées par des virgules (d'où le nom de ce format). Caractérisé par une facilité de partage, le CSV est lisible par la plupart des logiciels de tableur, mais il ne permet pas l'enrichissement typographique.

Deep learning

11. L'apprentissage profond est une méthode d'apprentissage automatique (cf. machine learning) fondée sur l'apprentissage de modèles de données.

Dépôt légal

12. Le dépôt légal est une disposition légale impliquant la remise aux institutions nationales de conservation du patrimoine que sont la Bibliothèque nationale de France, l'Institut national de l'audiovisuel et le Centre national de la cinématographie, d'exemplaires de toute production littéraire ou artistique, textuelle, graphique, sonore et audiovisuelle.

Distant reading

13. La lecture distante correspond à l'analyse portant sur un corpus de plusieurs centaines de textes dans le but d'identifier des *patterns* (ou motifs – cf. expressions régulières) au sein de ce corpus. Cette approche s'oppose à une lecture dite « traditionnelle » des sources, appelée *close reading* ou lecture rapprochée.

Dublin Core

14. Le Dublin Core est un modèle de description de tout type de ressource numérique, issu d'un consensus international et multidisciplinaire et développé par la Dublin Core Metadata Initiative (DCMI). Il a pour objectif de décrire des documents de manière simple et standardisée en fournissant un socle commun d'éléments descriptifs suffisamment structuré pour permettre une interopérabilité minimale entre des systèmes conçus indépendamment les uns des autres.

EAD : *Encoded archival description*

15. La description archivistique encodée est un standard de description archivistique qui permet d'encoder en XML un inventaire d'archives ainsi que des instruments de recherche archivistiques.

Entité nommée

16. Empruntée à la linguistique, une entité nommée désigne les termes à valeur de référence dans un texte, comme les noms propres ou les noms d'institutions par exemple.

Entrepôt OAI

17. Un entrepôt OAI est une base de données qui supporte le protocole OAI-PMH (pour Open Archives Initiative Protocol for Metada Harvesting) et dans laquelle des fournisseurs de données vont pouvoir y déposer leurs métadonnées en attendant qu'un robot vienne les moissonner, dans le but de les intégrer à son propre catalogue. Il contient alors des métadonnées qui sont disponibles dans différents formats afin de répondre à différents types de demandes.

Expressions régulières

18. Appelées également *regex* (pour *regular expression*) les expressions régulières sont des expressions rationnelles, des motifs constitués d'une chaîne de caractères et spécifiant des conditions à remplir lors d'une recherche dans un éditeur de texte, grâce à l'utilisation de caractères spéciaux (ou métacaractères) qui vont avoir une fonction d'opérateur. Elles servent à tester la présence ou l'absence d'un motif dans une chaîne de caractères et sont simples à utiliser, concises et puissantes. Elles sont

présentes dans de nombreux logiciels (tels que Word, Oxygen, etc.) et langages informatiques (Python, R, etc.).

FAIR : Facile à trouver, accessible, interopérable et réutilisable

19. Les principes FAIR s'inscrivent dans des démarches de science ouverte ainsi que d'ouverture des données de la recherche, et recouvrent des manières de construire, stocker, présenter ou publier des données afin qu'elles puissent être « faciles à trouver », « accessibles », « interopérables » et « réutilisables ». Chaque principe FAIR se décline ainsi en un ensemble de caractéristiques que doivent présenter les données et les métadonnées pour faciliter leur découverte et leur utilisation par les hommes ainsi que par les machines.

FRBR : Functional Requirements for Bibliographic Records

20. Les spécifications fonctionnelles des notices bibliographiques sont un modèle conceptuel de description bibliographique utilisé en bibliothèque. Les FRBR ne sont ni une norme de description bibliographique ni un format d'encodage. Elles se contentent de décrire les informations d'une notice bibliographique d'un point de vue logique, en décomposant la notice comme un ensemble d'informations correspondant à 4 niveaux d'analyse, du plus général au plus particulier : « Œuvre », « Expression », « Manifestation » et « Item ».

HTR : *Handwritten text recognition*

21. La reconnaissance de texte manuscrit est une technologie de reconnaissance automatique d'écriture manuscrite, un traitement informatique qui a pour but de traduire un texte écrit en un texte codé numériquement.

Instrument de recherche

22. Un instrument de recherche désigne tout outil papier ou informatisé énumérant ou décrivant un ensemble de documents d'archives de manière à les faire connaître aux utilisateurs. Quel que soit leur support, quel que soit le niveau de description auquel ils se situent, les instruments de recherche doivent respecter des principes communs. Un instrument de recherche doit mettre en évidence la structure du ou des fonds qu'il décrit, c'est-à-dire la hiérarchie de ses différents composants.

Interopérabilité

23. L'interopérabilité renvoie à la possibilité de communication entre deux ou plusieurs systèmes, appareils ou éléments informatiques. Pour les bases de données, il

s'agit de garantir le fait qu'elles puissent être ouvertes ou interrogées avec plusieurs outils.

JSON : *JavaScript object notation*

24. Le JSON est un format de données textuelles, dérivé du Javascript et permettant de représenter de l'information structurée.

KWIC : *Key word in context*

25. Le « mot-clé en contexte » désigne une méthode d'indexation qui consiste à repérer et lister des mots-clés par ordre alphabétique et dans leur contexte d'utilisation.

Layout analysis

26. L'analyse de la mise en page du document est un processus d'identification et de catégorisation des différentes régions d'intérêt au sein de l'image numérisée d'un document textuel, utilisé dans le cadre de la vision par ordinateur ainsi qu'en traitement automatique des langues (TAL).

Licence Creative Commons

27. Les licences Creative Commons constituent des autorisations non exclusives de reproduction, distribution et communication d'une œuvre en ligne à titre gratuit selon des conditions spécifiées. Élaborées par l'organisation Creative Commons, elles permettent ainsi de faire apparaître clairement au public les conditions de réutilisation d'une œuvre.

Machine learning

28. Le *machine learning* ou apprentissage automatique est un champ d'étude de l'intelligence artificielle qui se fonde sur des approches mathématiques et statistiques pour façonner et entraîner des algorithmes d'apprentissage incrémental. Ces algorithmes vont apprendre à partir de données afin de résoudre des tâches pour lesquels ils n'étaient pas explicitement programmés.

Métadonnée

29. Une métadonnée est une caractéristique formelle, normalisée et structurée, utilisée pour la description et le traitement des contenus des ressources numériques. Elles servent ainsi à référencer, identifier et partager correctement un document et sont à la base des techniques du Web sémantique.

N-gramme

30. En traitement automatique des langues (TAL), les n-grammes sont des séquences de n mots consécutifs. On parle aussi d'unigramme, de bigramme, de trigramme, etc. en fonction du nombre de mots composant les séquences. La recherche par n-gramme permet d'aller plus loin que l'analyse de fréquence par unigramme car elle permet de rattacher un terme à son contexte d'utilisation. De nombreux outils en TAL utilisent les n-grammes pour détecter les contenus dupliqués, les répétitions ainsi que pour donner la probabilité d'apparition d'un mot suivant.

OAI-PMH : Open Archives Initiative – Protocol for Metadata Harvesting

31. L'OAI-PMH est un protocole informatique développé par l'Open Archives Initiative pour la collecte et l'échange de métadonnées entre plusieurs institutions afin de multiplier les accès aux documents numériques. Il permet de constituer et de mettre à jour automatiquement des entrepôts centralisés (ou entrepôt OAI) où les métadonnées de sources diverses peuvent être interrogées simultanément et de construire ainsi des portails thématiques avec uniquement le résultat de requêtes spécifiques opérées sur ces entrepôts.

OCR : *Optical character recognition*

32. La reconnaissance optique de caractères désigne l'ensemble des techniques de transcription automatique de texte, en particulier imprimé, permettant de transformer l'image d'un texte numérisé en un document textuel et de le sauvegarder dans un fichier pouvant être exploité grâce un traitement de texte.

Ontologie

33. Une ontologie est une spécification formelle d'une conceptualisation partagée, c'est-à-dire un ensemble structuré des termes et concepts permettant de représenter le sens d'un champ d'informations. Elle constitue ainsi un modèle de données représentatif à la fois d'un ensemble de concepts propres à un domaine, ainsi que des relations qui articulent ces concepts entre eux.

Open access

34. Le libre accès ou accès ouvert est la mise à disposition en ligne de contenus numériques sous un régime de propriété intellectuelle ou de licences d'utilisation dites libres (telles que les licences Creative Commons). On parle alors dans ce cas d'*Open content* ou contenus libres, à la suite des logiciels libres. Il porte sur toute œuvre de l'esprit, littéraire ou artistique. L'*Open access* concerne notamment les publications scientifiques, soit sous la

forme d'articles publiés dans des revues ouvertes ou de dépôts dans des archives ouvertes, mais s'étend également aux données de la recherche ainsi qu'aux données publiques dans le cadre des mouvements d'*Open data* et de science ouverte.

OWL : *Web ontology langage*

35. Le langage OWL est un langage de représentation des connaissances permettant de définir des ontologies web structurées.

Parser

36. Un *parser* est un programme informatique utilisé en traitement automatique des langues (TAL) permettant de réaliser une analyse syntaxique d'un texte, c'est-à-dire d'analyser automatiquement une chaîne de mots constituant une phrase dans le but de mettre en évidence les relations coexistant entre ces mots et ainsi d'identifier la structure d'un texte.

Python

37. Python est un langage de programmation particulièrement utilisé comme langage de script pour automatiser des tâches simples mais répétitives.

RDF : *Resource description framework*

38. Le modèle RDF est un modèle informatique permettant l'échange de données sur le Web développé par le W3C (World Wide Web Consortium). Il s'agit du langage de base du Web sémantique car il permet la description de ressources web, de leurs métadonnées ainsi que des relations qui existent entre elles, amenant alors à la création de liens entre ces ressources selon la valeur des relations décrites.

Segmentation

39. En linguistique, la segmentation est le fait de transformer une chaîne de caractères en mots ou éléments sémantiques dans un texte afin de permettre son traitement automatique et son indexation.

SOC : Système d'organisation des connaissances

40. Un système d'organisation des connaissances (SOC) est un modèle permettant de représenter, d'organiser et de gérer des connaissances. Le terme est emprunté au monde des bibliothèques et désigne tout langage documentaire structuré : les schémas de classification des connaissances telles que les classifications bibliographiques, les répertoires vedettes-matière, les fichiers d'autorité, les thésaurus, les ontologies, etc.

TAL : Traitement automatique des langues

41. Le TAL est un domaine de recherche multidisciplinaire impliquant à la fois la linguistique, l'informatique et l'intelligence artificielle. Il cherche à modéliser le langage humain dans un but d'automatisation et afin de créer des outils pouvant être utilisés pour diverses applications telles que la traduction automatique ou la recherche d'informations multilingues.

TEI : *Text Encoding Initiative*

42. La TEI (ou « Initiative pour l'encodage du texte ») est une communauté académique internationale fondée pour normaliser l'utilisation du XML dans l'encodage sémantique de documents textuels historiques et littéraires. Par extension, on appelle TEI l'ensemble des balises et leurs règles d'application telles que définies et régulièrement mises à jour par le consortium.

Terminologie

43. Une terminologie désigne un ensemble de dénominations utilisées dans un domaine du savoir. Il regroupe alors des termes rigoureusement définis, relatifs à un système notionnel élaboré par des constructions théoriques, par des classements ou des structurations de matériaux observés, de pratiques sociales ou d'ensembles culturels.

Thésaurus

44. Langage documentaire et d'indexation, le thésaurus est une liste organisée de termes contrôlés et normalisés représentant les concepts d'un domaine de la connaissance. Ces termes obéissent à des règles terminologiques propres et sont reliés entre eux par des relations sémantiques. Le thésaurus sert alors à traduire en langage documentaire des notions exprimées en langage naturel.

Token

45. En analyse lexicale, un *token* est une unité lexicale issue de la conversion d'une chaîne de texte en mots par segmentation.

URI : *Uniform resource identifier*

46. Un URI est un identifiant uniforme de ressource, c'est-à-dire une chaîne de caractères permettant d'identifier de façon unique et pérenne une ressource sur un réseau (par exemple une ressource web), et ce même si cette ressource est déplacée ou supprimée. L'URN (*Uniform resource name*), qui permet d'identifier une ressource, et l'URL (*Uniform resource locator*), qui permet d'identifier la localisation d'une ressource, sont des spécialisations d'URI.

Web scraping

47. Le *web scraping* désigne une technique d'extraction du contenu de sites web (tel que des métadonnées par exemple), via un script ou un programme (appelé *scraper*). L'intérêt réside alors dans la transformation de ce contenu en vue de le réutiliser dans un autre contexte.

Web sémantique

48. Le Web sémantique est une vision du Web reposant sur le partage structuré et intelligent des données, en généralisant un système de métadonnées. Appelé également « Web de données » ou Web 3.0, il s'appuie sur des standards tels que l'URI (*Uniform resource identifier*), qui identifie une ressource, et le modèle de données RDF qui permet quant à lui de décrire, représenter et relier des données.

Word embedding

49. Le *word embedding* – vectorisation de mots ou plongement lexical – est une technique issue du TAL permettant de représenter des mots par un vecteur de nombres réels afin de rendre compte des contextes dans lesquels ces mots apparaissent au sein d'un texte.

XML : *Extensible markup langage*

50. Langage de balisage extensible, le XML est un métalangage informatique dont la liste des balises n'est pas limitée (extensible) et permet alors de structurer son langage selon des besoins variés. Cette structure ouverte vise à faciliter l'interopérabilité, c'est-à-dire l'échange automatisé de contenus numériques complexes entre systèmes d'informations hétérogènes. Le XML est devenu ainsi le langage de référence pour la représentation et l'échange d'informations contenues dans une ressource numérique.

Résumés

Traverser les corpus de presse numériques : un travail d'artisan ?

Juliette De Maeyer

1. Ce chapitre interroge ce que les études médiatiques ont à apporter aux humanités numériques, et réciproquement. Deux projets de recherche récents, l'un portant sur l'évolution du copier-coller dans la presse, l'autre sur Franklin Ford, un obscur journaliste et théoricien des médias américains de la fin du XIX^e^ siècle, ont permis d'expérimenter différentes façons d'aborder, de traverser et de travailler des corpus de presse nativement numériques ou numérisés.
2. Ce parcours révèle des tensions productives entre lecture distante et lecture rapprochée, entre anciens et nouveaux médias, entre l'aléatoire et le linéaire. Au gré des erreurs d'OCR, des corpus aux frontières indisciplinées, et d'un dialogue parfois surprenant avec un *bot*, ce chapitre propose de mettre en évidence des enjeux liés à la matérialité des archives, aux différentes couches de remédiation que celles-ci traversent inévitablement, et à la possibilité d'inclure le désordre dans nos démarches méthodologiques.

Le corpus de tous les livres depuis les débuts de l'imprimerie, tous comptes faits…

Frédéric Glorieux

3. Lancé en 2010 par Google, Ngram Viewer est un outil linguistique permettant aux utilisateurs d'observer l'évolution de la fréquence d'un mot ou d'un groupe de mots à travers le temps parmi les sources imprimées de Google Books. Si le projet est aujourd'hui à l'arrêt, son ambition s'appuyait jusqu'en 2013 sur une importante campagne de numérisation d'un très grand nombre d'ouvrages en proposant par-dessus un service d'analyse en statistique lexicale.
4. Afin de rendre compte de l'ampleur du projet et de le confronter à cette ambition, il est nécessaire de revenir sur le travail d'océrisation et de documentation effectué. Or, entre les erreurs d'OCR et les problèmes de catalogage, Ngram Viewer met à disposition un corpus fortement bruité, interrogeable avec des outils ne nous renseignant pas sur la fiabilité des données textuelles consultées. Dans ce cadre, il convient de s'interroger sur la capacité réelle de cet outil à « faire corpus ».

Archelec, les archives électorales françaises de la V^e^ République, du papier au numérique : reflet fidèle ou distorsion ?

Odile Gaultier-Voituriez

5. Les archives électorales françaises de la V^e^ République du Centre de recherche politique de Sciences Po (CEVIPOF) ont commencé à être numérisées en partenariat avec la bibliothèque de Sciences Po à partir de 2013. Constitué dans un but de recherche en interne par les politologues et les sociologues du CEVIPOF, le fonds d'archives papier contient des profes-

sions de foi, bulletins de vote, tracts, affiches mais aussi des résultats électoraux, de la presse et des travaux de chercheurs effectués à partir de ces documents.

6. La variété de ce fonds a soulevé diverses questions quant aux choix à effectuer en matière de numérisation et de risque juridique. Les droits s'appliquant aux différents types de documents du fonds peuvent varier, et les modalités de réutilisation à envisager également. De fait, le corpus mis en ligne au final n'est pas le reflet fidèle du fonds papier de départ.

7. Mais quels sont les enjeux qui apparaissent dans cette distorsion entre ce fonds papier d'origine et son miroir mis en ligne ? Est-il utilisable en toute connaissance de cause et d'une manière éthique par les chercheurs universitaires en science politique, histoire ou sociologie, les particuliers, généalogistes, passionnés de politique ou simples curieux ? Quels usages peuvent être envisagés ?

Le traitement numérique des sources : la construction des corpus et des instruments de recherche comme enjeu pour la mise à disposition des données

Céline Alazard, Jean Vigreux et Serge Wolikow

8. Depuis l'émergence de pratiques informelles historiennes liées à l'usage de la photographie numérique dans la collecte d'archives jusqu'à la collaboration entre laboratoires et dépôts d'archives, différentes expériences menées ces dernières années ont permis de montrer la nécessité de produire des guides de bonnes pratiques et des « boîtes à outils ». Dans une démarche méthodique, la MSH de Dijon a travaillé à la structuration et à la normalisation de bases de données et instruments de recherche électroniques notamment autour de trois thèmes : mouvements sociaux, vigne et vin, et archives de la recherche.

9. L'ouverture des archives russes après 1990 a rendu consultables les archives du communisme français restées longtemps inaccessibles. La coopération entre chercheurs, archivistes et informaticiens dans une entreprise de guide des fonds, collecte des archives numérisées, organisation des données, stockage et mise en ligne a permis d'offrir à la communauté scientifique l'accès aux archives via un outil performant, visant à l'exhaustivité des inventaires indexés, pour une histoire renouvelée de la galaxie communiste. Quant à la constitution d'instruments de recherche consacrés aux ressources sur la vigne et le vin et aux archives de la recherche sous la forme notamment de bases de données interopérables, elle a permis d'envisager des corpus plus vastes en croisant des sources de natures diverses : archives, imprimés – brochures et revues –, archives électroniques natives, etc.

Les corpus textuels numériques (re)spécifiés

Damon Mayaffre

10. Si le corpus a acquis une place centrale en linguistique à partir des années 2000, il s'est vu revendiqué par tous en SHS. Ce triomphe n'est pas sans poser problème pour la linguistique de corpus qui a vu son objet banalisé et ses contours s'estomper, au point que le corpus semble aujourd'hui avoir perdu de ses spécificités.

11. Cette contribution se propose alors d'interroger à nouveau l'objet *corpus*, que ce soit dans sa dimension textuelle comme dans sa dimension numérique. Composé de textes, le corpus

en emprunte-t-il les propriétés ? Est-il une ressource soumise à interrogatoire ou un objet qui nous dirige dans le questionnement ? De quel co(n)texte est-il le nom ? En quoi peut-on considérer un corpus comme un texte ? Et quel rôle joue le numérique dans cette redéfinition du corpus ?

12. Au-delà des enjeux de constitution empirique et d'établissement théorique du corpus, des questions relatives au traitement ainsi qu'aux méthodes et logiciels à disposition pour le chercheur demeurent. Le traitement des corpus textuels numériques par l'intelligence artificielle et le *deep learning*, par exemple, participent ainsi de cette actualisation de la définition du corpus.

La méthode diplomatique face à l'information numérique

Marie-Anne Chabin

13. La diplomatique est une discipline tricentenaire élaborée pour déterminer méthodiquement, scientifiquement, si un acte est authentique ou si c'est un faux, c'est-à-dire si ce document est bien ce qu'il prétend être, s'il a été fabriqué de toute pièce ou falsifié. La méthode propose une sorte de quadrillage du document avec un vocabulaire dédié à la description des différentes zones et valeurs d'information, ainsi qu'une démarche de critique de l'écrit comparé à d'autres écrits comparables.

14. Autrement dit, au-delà de la critique du contenu, c'est-à-dire du message exprimé par l'auteur, la diplomatique s'intéresse à la forme, c'est-à-dire à tout ce que l'on peut apprendre par l'examen de l'apparence de l'écrit mais aussi en analysant l'agencement des données et les références de production ou de transmission. Si la diplomatique est pertinente pour identifier les faux millénaires, pourquoi ne le serait-elle pas pour établir le degré de véracité des écrits d'aujourd'hui (mails, articles, posts, notes, etc.) mais aussi pour participer à la détection des faux administratifs (en recrudescence) voire des fake news ?

15. La diplomatique numérique consiste alors en cette transposition de la démarche traditionnelle aux traces numériques des actes et gestes pouvant engager la responsabilité des individus dans leurs relations administratives, contractuelles, sociales ou privées. Cette démarche reste ainsi originale et efficace dans l'évaluation de la fiabilité des données, tant dans une dimension juridique que patrimoniale. Toutefois, elle doit pour ce faire reformuler son champ d'application au XXIe siècle, moderniser son vocabulaire au sein des sciences humaines et sociales, et élargir les recherches.

Le goût de l'archive à l'ère numérique : gestes et récits historiens, du document au corpus

Caroline Muller et Frédéric Clavert

16. *Le Goût de l'archive à l'ère numérique* est un projet co-dirigé par Caroline Muller et Frédéric Clavert. Livre en ligne écrit de manière collaborative, il entend interroger les routines numériques « discrètes » des historiens face à l'archive, y compris au moment de la constitution du corpus, et examiner les conséquences possibles pour l'écriture de l'histoire de cette introduction, tant logicielle que matérielle, de l'informatique dans les pratiques historiennes.

17. Les différentes contributions en ligne qui alimentent ce livre « vivant » traitent tout à la fois de la collecte des archives et de la constitution des corpus, du gigantisme de ces cor-

pus numérisés ou nativement numériques et des relations à l'archive ainsi qu'à la salle de lecture. Ces témoignages à teneur ethnographique permettent ainsi de dégager quelques conclusions intermédiaires sur les pratiques informatiques et numériques exposées.

18. Quelle place occupe l'outillage de l'historien dans l'écriture de l'histoire ? Quelle diversité de pratiques numériques se cachent ainsi derrière ce *Goût de l'archive à l'ère numérique* ? Quels contrastes entre les disciplines ces pratiques reflètent-elles ? Ce retour d'expérience de publication collective permet à la fois d'esquisser des réponses ainsi que d'offrir des perspectives d'étude sur cette nouvelle relation à l'archive.

Enrichir un corpus de sources numérisé en histoire de l'éducation. Le cas du *Bulletin administratif de l'instruction publique*

Solenn Huitric

19. Publié tous les mois entre 1850 et 1932, le *Bulletin administratif de l'instruction publique (BAIP)* contient l'ensemble des actes pris par le ministère de l'Instruction publique sur cette période. Il a fait l'objet d'une numérisation dans le cadre du projet de *Bibliothèque historique de l'éducation (BHE)* visant à proposer une version numérisée et enrichie de corpus en lien avec l'histoire de l'éducation. Cependant, le passage de ce corpus sous un régime numérique n'est pas sans poser des problèmes dont les caractéristiques ne tiennent pas seulement de la nature du corpus en lui-même.

20. À travers ce retour d'expérience, la question de la position que les historiens peuvent adopter dans cette configuration apparaît centrale. La numérisation ne peut simplement correspondre à une transposition numérique du *BAIP* et il devient alors nécessaire d'outiller l'accès aux documents d'archives. Ces possibilités impliquent alors un changement de rapport à la source, un rapport enrichi mais dont les usages différenciés peuvent contraindre cet enrichissement. Dans ce contexte, ce retour d'expérience vient mettre en lumière les différentes étapes de définition du cadre de la numérisation à travers une double lecture à la fois du processus ainsi que des enjeux, en adoptant le regard de l'historien.

Deux siècles de sources disparates sur l'industrie textile en France : comment automatiser les traitements d'un corpus non uniforme ?

Alix Chagué, Victoria Le Fourner, Manuela Martini et Éric Villemonte de la Clergerie

21. Le projet ANR *TIME-US* ambitionne de créer un corpus permettant d'analyser les rémunérations et budgets-temps des hommes et femmes travaillant dans l'industrie du textile dans les régions de Lille, Paris, Lyon et Marseille. Il s'agit de collecter et analyser des données couvrant une période longue, allant de la fin du XVII^e^ au début du XX^e^ siècle.

22. Pour mener cette recherche, le projet rassemble les expertises d'historiens, de sociologues, de spécialistes du TAL et du traitement numérique des documents historiques. Outre l'analyse classique des sources, le projet entend créer des séries comparables de données sur les rémunérations et le temps alloué à l'emploi des travailleurs du textile. Le traitement du corpus tire profit de la variété des méthodologies en jeu dans cette approche pluridisciplinaire tout en visant à correspondre aux attentes de chacun.

23. En deux ans, de nombreux fonds d'archives ont été dépouillés et numérisés, aboutissant à la création d'un corpus disparate de 18 000 images qui mêlent imprimés et manuscrits. La diversité de ces documents a conduit à l'élaboration de plusieurs stratégies pour traiter et unifier le corpus ; celles-ci se sont avérées, jusqu'à un certain point, généralisables entre type de documents. Ce retour d'expérience entend alors présenter les stratégies mises en œuvre pour l'acquisition des doubles numériques, l'extraction du texte et des données, ainsi que leurs transversalités et leurs limites.

Un océan d'images : établir un catalogue raisonné d'estampes à l'ère du numérique

Johanna Daniel

24. En s'appuyant sur une expérience en cours – la constitution d'un catalogue raisonné numérique de plusieurs milliers d'estampes dans le cadre d'une thèse – ce chapitre porte sur les interactions entre les historiens de l'art et les institutions patrimoniales détentrices du matériau mobilisé dans le cadre de la recherche. La mise en ligne des collections patrimoniales permet aujourd'hui une récupération et un traitement automatisés d'importantes quantités de données et d'images numérisées.

25. Bien que s'inscrivant dans une démarche de recherche scientifique, ces nouvelles façons de constituer un corpus à l'aide d'outils numériques peuvent susciter des inquiétudes chez les agents des institutions propriétaires des originaux (« avez-vous le droit de récupérer nos données ? », « qu'allez-vous faire de nos images ? »). Des blocages peuvent alors survenir compliquant un traitement computationnel du matériau.

26. Pour le chercheur, travailler à partir de données produites par des professionnels de l'information (ici en bibliothèques et en musées) nécessite de connaître et de comprendre leurs pratiques métiers spécifiques, parfois éloignées de celles du monde universitaire. Comment travailler ensemble sur un matériau commun à l'heure du numérique ? Comment concilier des pratiques métiers parfois divergentes ? Comment rendre interopérables les données produites et enrichies de part et d'autre ? Ce chapitre esquissera quelques pistes possibles.

Du corpus archivistique au corpus numérique : les soubassements du Web sémantique. L'exemple des sources relatives au parlement de Flandre

Renaud Limelette

27. Archive et numérique sont deux mots dont la conjonction est devenue presque un topique. Nous mesurons, au fil du temps, que le public se déplace moins dans les dépôts d'archives, la consultation d'un document se faisant à travers une application numérique. Pourtant la donnée numérique n'est pas l'archive, elle n'est qu'une représentation approchante. Autrement dit, la donnée numérique n'est qu'un artefact de la donnée archivistique.

28. L'appréhension de ce lien distendu entre document archivistique et document numérique se manifeste à travers la « granularité des données ». D'un autre côté, la donnée archivistique est par nature beaucoup plus statique, car elle se trouve dans un entrepôt ou une bibliothèque, alors que, sans être

volatile, la donnée numérique est interopérable, grâce aux métadonnées qui la caractérisent.

29. À travers un cas concret d'archives judiciaires, ce chapitre se propose d'apporter des éléments de réflexion sur la granularité des données et les métadonnées. Car face à l'ampleur des corpus archivistiques, l'informatique offre aujourd'hui des moyens de structuration des données à même de faciliter leur échange et leur compréhension entre machines.

Les technologies du Web pour la valorisation d'un patrimoine industriel textile en mouvement dans les Hauts-de-France

Éric Kergosien et Mathilde Wybo

30. Une question sociale importante dans le domaine du patrimoine culturel est liée à la collecte, l'analyse, la publication et la mise en valeur de la mémoire des acteurs du domaine, soit parlée ou écrite. La formalisation de l'information sur le patrimoine culturel constitue un véritable défi. Le volume et la diversité des ressources posent de nombreux problèmes tels que l'indexation des données, leur structuration et leur valorisation au sein d'une même base de connaissances. La plupart des tentatives de résolution des problèmes d'interopérabilité sémantique se concentrent sur la standardisation et le développement de structures communes telles que FRBR, FRBRoo, CIDOC CRM, etc. Parmi ces modèles, le CIDOC est une référence conceptuelle, modèle spécialement conçu pour la modélisation du patrimoine culturel. Ce modèle offre en effet un schéma commun de métadonnées rendant les concepts compréhensibles et interopérables.

31. Afin d'aider les experts du domaine à produire et fournir des contenus numériques, une méthodologie en trois étapes est adoptée afin de permettre la construction semi-automatique d'une représentation sémantique d'un domaine étudié à partir de documents hétérogènes. Cette méthodologie passe par un recueil et la formalisation d'un historique par le biais d'entretiens avec les acteurs du domaine, pour permettre par la suite l'identification et l'extraction d'informations relatives au patrimoine culturel industriel à partir de milliers de documents collectés auprès de ces acteurs (interviews, articles de journaux, etc.). L'approche proposée combine la projection lexicale avec des méthodes de fouille de textes pour améliorer l'identification de l'information pertinente et aboutit alors à une première version d'une ontologie construite automatiquement au format OWL, en utilisant le modèle CIDOC CRM comme base conceptuelle pour fusionner toutes les informations extraites. Les expériences sont menées sur un corpus relatif au patrimoine industriel textile collecté grâce au projet *DENIM*.

Méthodologie de validation et d'enrichissement d'une ontologie minière fondée sur le CIDOC CRM

Amélie Daloz

32. Dans le contexte du projet ANR *Mémo-Mines*, qui participe à la sauvegarde numérique du patrimoine minier du Nord–Pas-de-Calais, ce chapitre présente la méthodologie de validation et d'enrichissement du modèle ontologique CIDOC CRM. Il s'agit d'un modèle de représentation de données qui doit permettre l'interopérabilité des référencements des objets de musées puis, par extension, de tout objet du patrimoine

culturel matériel ou immatériel, selon la définition proposée par l'UNESCO.

33. Pour définir le modèle ontologique et pour le peupler, la méthodologie articule une approche ascendante et descendante et s'appuie sur la constitution et l'analyse de deux types de corpus : un corpus presse centré sur le domaine minier et un corpus audiovisuel constitué de témoignages d'anciens mineurs. Dans quelle mesure alors la construction d'un modèle ontologique CIDOC CRM pour le domaine minier contribue-t-elle ainsi à formaliser la notion de patrimoine minier ? Les résultats illustrent l'approche et la discutent à partir d'un exemple précis issu de la tradition minière.

Le programme des registres de la Comédie-Française : un corpus numérique en extension

Agathe Sanjuan

34. Le programme des registres de la Comédie-Française associe depuis 2008 plusieurs universités françaises, américaines, canadiennes et un établissement culturel – la Comédie-Française – autour d'un ensemble d'archives physiques d'une très grande richesse nous renseignant sur l'histoire du théâtre depuis la fin du XVII^e^ siècle. Par le numérique, ce corpus de documents s'est transformé en corpus de données toujours plus vaste, s'élargissant au fil des années et des contributions de chercheurs, apportant leur pierre à la réflexion sur les méthodes, les outils et l'exploitation des données.

35. Néanmoins, le chercheur est confronté à une nouvelle difficulté : le raisonnement s'appuie sur un corpus en extension. Au fil des mois, des financements et des projets se rajoutent des corpus, des tables, des outils de visualisation. Alors que la consultation des archives papier est nécessairement limitée à l'unité de communication en salle de lecture, l'archive numérique n'est peut-être pas, quant à elle, réellement « définitive ». Si la pérennité est assurée par l'interopérabilité des systèmes et l'archivage du site sur des plateformes d'État, l'accroissement des corpus et la mobilité des outils (visualisations, design) posent de nouveaux défis à la recherche et à la démarche scientifique.

Le projet *eBalzac* : construire une bibliothèque hypertextuelle des sources intellectuelles

Andrea Del Lungo et Karolina Suchecka

36. Ce chapitre expose les résultats du projet *Phœbus-eBalzac* (*Projet d'hypertexte de l'œuvre de Balzac reposant sur l'utilisation des similarités*), financé par l'Agence nationale pour la recherche (ANR) pour la période 2015-2019 et présente un développement encore inédit du projet sur la détection automatique et la visualisation de l'hypertexte balzacien. Porté par les équipes CELLF et LIP6 de Sorbonne Université et par l'équipe ALITHILA de l'université de Lille, ce projet consiste à mettre en résonance l'ensemble de l'œuvre balzacienne avec un vaste corpus d'écrits contemporains qui ont pu la nourrir (œuvres romanesques, littérature panoramique, ouvrages scientifiques).

37. L'objectif est de permettre des recherches et des comparaisons intertextuelles élaborées afin de détecter des citations, des reprises ou des plagiats éventuels, et de constituer ainsi une cartographie de l'univers intellectuel de Balzac à partir des traces que d'autres textes ont laissées dans l'œuvre. Par son ampleur, mais aussi par le caractère hétérogène de ses

sources, *La Comédie humaine* constitue alors un objet idéal pour ce type d'édition expérimentale qui pourra prendre valeur de paradigme. Car en effet, le modèle ainsi constitué vise à être opératoire pour d'autres auteurs chez qui l'usage d'une intertextualité abondante et éclectique est avéré.

L'*Édition numérique collaborative et critique de l'Encyclopédie de Diderot et D'Alembert* (*ENCCRE*), comme prototype d'un laboratoire virtuel de recherches sur l'*Encyclopédie* et les Lumières

Alexandre Guilbaud

38. Librement accessible à l'adresse http://enccre.academie-sciences.fr depuis le 19 octobre 2017, l'*Édition numérique collaborative et critique de l'Encyclopédie* (1751-1772), l'*ENCCRE*, est un projet numérique réalisé par une équipe internationale et pluridisciplinaire de 130 membres. Parmi cette équipe figurent plus d'une centaine de chercheurs spécialistes de l'*Encyclopédie* de Diderot et D'Alembert, des chercheurs en informatique, des ingénieurs, des étudiants et des bénévoles.

39. L'édition réalisée dans ce cadre repose sur une ambitieuse politique d'annotation fondée sur plusieurs niveaux d'articulation entre la matérialité de l'exemplaire original sur lequel s'appuient l'édition, la représentation numérique de l'œuvre et l'apparat critique que nous sommes en mesure de constituer. L'*ENCCRE* s'appuie en effet sur une interface de consultation, complétée, en amont, par une interface collaborative d'édition munie de nombreux outils à disposition des chercheurs pour décrire, annoter et effectuer des recherches sur l'œuvre.

40. Nous examinerons les possibilités et limites actuelles de cet espace numérique au regard de ce qu'il tend et aspire progressivement à devenir : un laboratoire virtuel de recherche sur l'*Encyclopédie*, où l'étude collective de l'œuvre et de son contexte doit permettre aux diverses facettes de l'histoire des idées, des sciences et des techniques au siècle des Lumières de se rencontrer, de s'enrichir et d'avancer de concert.

Pour un regard à 360 degrés sur les corpus visuels : pratiques de mise à disposition et de réutilisation

Antoine Courtin

41. Qu'est-ce qu'un corpus visuel numérique en histoire de l'art ? Cette question se pose à tout acteur confronté à un projet de mise en ligne d'un corpus visuel, que ce soit à des fins de diffusion, de valorisation d'un travail de recherche voire par obligation (dans le cadre d'un financement par exemple). Ce chapitre se propose alors d'évoquer les principaux enjeux actuels en matière de constitution de tels corpus ainsi que le champ des possibles ouvert par leur mise à disposition, notamment grâce aux avancées du *deep learning*, en adoptant un double point de vue à la fois d'acteur institutionnel et de réutilisateur de contenus.

42. Il sera question d'étudier comment les corpus visuels permettent de constituer ainsi un « millefeuille informationnel » sur les artefacts culturels par les différents acteurs qui constituent, manipulent, enrichissent, publient ou encore analysent ces corpus visuels. L'étude de cette problématique sera alimentée par des travaux inspirants et témoignant des pratiques actuelles, à l'intersection de la recherche en histoire de l'art et des GLAM (*Galleries, Libraries, Archives and Museums*).

Le musée comme service d'information. Pour une politique des interfaces muséales

Emmanuel Château-Dutier

43. L'ouverture des collections muséales – ou Open GLAM – suppose que les musées envisagent tant dans leurs politiques que dans leur organisation les nouvelles modalités qu'implique la publicisation de ces collections. Il ne s'agit pas tant ici d'évoquer les catalogues en ligne, leurs applications pour terminaux mobiles, ou les déploiements d'expositions virtuelles, mais plutôt des formes plus spécifiques à l'informatique comme la mise à disposition de jeux de données, ou la création d'interfaces programmables qui peuvent directement intéresser les historiens d'art. Comment ces diverses formes de publication offrent-elles alors de nouvelles interfaces destinées à la création de services numériques et à des usages nouveaux des collections ?

44. Dans son ouvrage publié en 2012 intitulé *The Interface Effect*, le philosophe Alexander Galloway invite non seulement à définir l'interface mais également à l'interpréter. Les interfaces ne sont pas des objets simples ou des points de contact, mais constituent selon lui des zones autonomes d'activité. Ce ne sont pas des choses mais plutôt des processus qui effectuent un résultat, et qui racontent les forces plus larges qui les ont engendrées. Il s'agit pour lui d'une allégorie du contrôle. À travers ce chapitre, nous voudrions aborder les enjeux politiques qui se posent actuellement dans la mise en place de ce type d'accès qui oblige à penser le musée comme service d'information et proposer l'esquisse d'une politique des interfaces.

Archivage du Web, un enjeu de gouvernance (d'Internet)

Francesca Musiani

45. En 1980, le philosophe et sociologue Langdon Winner se demandait dans un article qui a fait école : « Est-ce que les artefacts sont politiques ? » (« *Do artifacts have politics?* »). Si l'on souhaite appliquer cette hypothèse aux archives du Web, il s'agit de comprendre en quoi, dans l'archivage du Web, existent des formes spécifiques d'autorité et de pouvoir qui dessinent une sorte de microcosme de la gouvernance d'Internet. Quels sont les différents modèles appliqués en matière d'archivage du Web ? Quels sont les contours d'une archive Web ? Quel régime particulier est appliqué pour les réseaux sociaux numériques ? Quels sont les freins à l'archivage ? Et quels enjeux de gouvernance y retrouve-t-on ?

46. Nous verrons dans ce chapitre que l'archivage du Web repose sur un modèle multi-parties prenantes, avec une grande variété d'acteurs : des fondations comme Internet Archive, des organisations transnationales (à commencer par l'International Internet Preservation Consortium, ou IIPC), la société civile et enfin le secteur privé. S'il n'échappe pas à des tensions ayant trait à la standardisation – un des enjeux traditionnellement les plus vifs de la gouvernance d'Internet – et à des visions et imaginaires divergents, des communs aux formats propriétaires, l'archivage du Web révèle également la présence de tensions géopolitiques où s'y retrouvent des dynamiques qui rappellent le problème de la fracture numérique. À cela s'ajoute enfin une relation complexe entre différentes pratiques et sources d'autorité ou de normativité : de la technologie au marché, de la concertation transnationale et internationale aux standards et aux droits.

Former « au numérique » en sciences humaines et sociales ? Propositions d'un historien

Émilien Ruiz

47. Après une dizaine d'années de développement des humanités numériques en France, nos disciplines sont à la croisée des chemins. Pourquoi former au numérique des étudiants en SHS ? Comment les initier à la critique et à l'exploitation de corpus et d'archives à « l'ère numérique » ? Qui doit se charger de tels enseignements ? Si des masters professionnels de haut niveau existent, ces questions fondamentales restent ouvertes pour les formations initiales généralistes.

48. Tandis que le monde qui nous entoure poursuit sa conversion, en SHS les réticences perdurent chez certains étudiants comme parmi leurs enseignants. Nos disciplines sont ainsi encore souvent perçues comme quasiment étrangères à ces enjeux. Dès lors, la tentation d'un repli sur soi des partisans des humanités numériques peut être grande. Pourtant les injonctions à « faire du numérique » se multiplient. Souvent assimilées à des contraintes, elles sont, en réalité, une chance pour les SHS.

49. Intégrer pleinement le numérique dans nos licences et nos masters permettra bien sûr de mieux former les apprentis chercheurs. Mais c'est également une occasion rêvée de repenser et d'élargir l'éventail des débouchés professionnels offerts à celles et ceux qui ne souhaiteraient (ou ne pourraient) pas rejoindre le monde académique au terme de leurs études. Pour cela, au moins une condition semble indispensable : cesser de considérer le numérique comme une spécialité pour le placer au cœur de nos pratiques pédagogiques.

Structurer automatiquement un corpus homogène issu de la reconnaissance d'écriture manuscrite : les jugements du Conseil des prud'hommes des tissus parisiens

Victoria Le Fourner, Alix Chagué, Manuela Martini et Anaïs Albert

50. Cet article, rédigé dans le cadre de la publication des données du présent ouvrage, fait suite à un premier retour d'expérience dans lequel la méthodologie établie pour la constitution du corpus de textes numériques ainsi que des premiers résultats avaient été présentés. Le *data paper* que nous proposons revient sur la structuration d'une des sources utilisées dans le cadre du projet *TIME-US* : les minutes du Conseil des prud'hommes de tissus parisiens. La question de recherche au centre de ce traitement des données était celle de l'utilisation conjointe des principes de la TEI et du traitement automatique des langues (TAL). Nous souhaitons ici mettre en lumière la recherche et l'application de règles communes pour la structuration automatique d'un corpus homogène. Après avoir rappelé le contexte de production de ces documents, nous détaillons les stratégies de constitution du jeu de données mis en place ainsi que la chaîne de traitements. Nous présentons ensuite l'organisation des données déposées dans l'entrepôt Zenodo. Enfin, nous évoquons les possibles réutilisations des données.

Phœbus e-Balzac : édition numérique exhaustive d'un monument littéraire

Karolina Suchecka, Victoria Le Fourner et Andrea Del Lungo

51. Nous présentons les données du corpus principal du projet e-Balzac sous la forme d'un *data paper*. Ce corpus des

95 textes de *La Comédie humaine* propose en plusieurs versions des éditions parues du vivant de Balzac, principalement celles dites « Furne » et « Furne corrigé ». Nous introduisons dans un premier temps le protocole de production des données, suivant la chaîne du traitement Odette/Teinte mise en place par Frédéric Glorieux. Nous présentons ensuite le jeu de données déposé dans Nakala et son organisation. Nous proposons enfin une réflexion sur la réutilisation des ressources mises à disposition.

52. Rédigé dans le cadre de la collection humanités numériques et science ouverte, destinée à favoriser la publication des données scientifiques, ce *data paper* est en lien avec le chapitre « Le projet e-Balzac : construire une bibliothèque hypertextuelle des sources intellectuelles ». Il a été présenté et discuté lors du colloque *DHNord 2021. Publier, partager, réutiliser les données de la recherche. Les data papers et leurs enjeux*.

La collection Humanités numériques et science ouverte (HNSO), co-dirigée par Clarisse Bardiot et Émilien Ruiz, est financée par le Fonds national pour la science ouverte et portée par la Maison Européenne des Sciences de l'Homme et de la Société (MESHS) et les Presses universitaires du Septentrion (PUS).

Elle a pour objectif de publier en *open access* des monographies et des ouvrages collectifs ainsi que les données associées. Contribuant ainsi à l'ouverture et à la diffusion des données, la collection se veut aussi un terrain d'expérimentation et de réflexion en pratique sur ce que la science ouverte fait aux SHS.
Défendant une conception pluraliste des humanités numériques, cette collection s'adresse aux spécialistes des diverses disciplines des sciences humaines et sociales qui inscrivent leurs travaux dans une démarche empirique et accordent une attention particulière à la constitution, la structuration, l'exploitation et à la visualisation de leurs données ; sans exclusive concernant les types de sources, les méthodes employées ou les tailles de corpus mobilisés.

Cet ouvrage a été financé par le Fonds national pour la science ouverte.

ISBN : 978-2-7574-3610-3
ISSN : en cours

Ouvrage composé par
Jonas Mazot, Chloé Gaillard & Sarah Bouchez

Ouvrage réalisé avec
La chaîne d'édition XML-TEI Métopes
Méthodes et outils pour l'édition structurée

Dépôt légal
décembre 2022

2 108e volume édité par
les Presses universitaires du Septentrion
Villeneuve d'Ascq – France