HAL
open science

Des humanités aux données. Méthodes en histoire à l'ère du numérique, témoignage d'une e-historienne

Nicole Dufournaud

▶ To cite this version:

Nicole Dufournaud. Des humanités aux données. Méthodes en histoire à l'ère du numérique, témoignage d'une e-historienne. Les Cahiers du numérique, 2014, 10 (3), pp.73-88. hal-02508714

HAL Id: hal-02508714
https://hal.science/hal-02508714v1

Submitted on 17 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DES HUMANITÉS AUX DONNÉES

Méthodes en histoire à l'ère du numérique, témoignage d'une e-historienne

NICOLE DUFOURNAUD

À partir de mes expériences personnelles sur une dizaine de projets en édition numérique des textes, l'article montre comment l'usage de l'ordinateur impose de nouvelles méthodes en histoire et renouvelle nos pratiques. Je plaide pour une collaboration numérique dans les équipes de recherche en faisant le parallèle avec la façon de travailler des humanistes qui ont accompagné les imprimeurs à la Renaissance. Aujourd'hui, de nouvelles compétences sont requises en histoire comme celle de posséder le langage informatique XML qui permet l'encodage des textes. Au-delà d'une simple édition de textes, je montre que la structuration numérique des données rend explicite le sens d'un texte et favorise l'analyse historique. De nouvelles perspectives sociales et scientifiques s'ouvrent qui remettent en question nos connaissances mais également les rapports entre les lecteurs et les équipes de recherche.

1. Introduction

L'informatique impose de nouvelles méthodes de travail en histoire, que ce soit dans la consultation des sources et des livres ou dans le travail scientifique qui résulte d'analyses de textes. Pour Jean-Philippe Genet :

> *... L'informatique ne connaît pas les frontières disciplinaires ! Au contraire, elle nous offre une voie royale vers la pluridisciplinarité, puisqu'elle met à la disposition des historiens les outils qui sont ceux des philologues, des linguistes, des sociologues, des anthropologues, etc. qui introduisent, en même temps à des méthodes, à des points de vue, à des approches diverses qui ne peuvent qu'enrichir l'historien... La machine, ici comme ailleurs, n'est rien sans l'homme. Tout doit être pensé en fonction de la relation entre l'homme et la machine, c'est-à-dire en tenant compte des aptitudes, des compétences et des intentions de l'homme, ici, en l'occurrence de l'historien. [Si l'historien se consacre habituellement au passé, ici, il doit se préoccuper du futur] puisque l'outil informatique modifie et modifiera profondément nos pratiques d'historien, à en juger par les transformations que les débuts de l'informatique ont d'ores et déjà produites.* (Genet, Zorzi, 2011, 4)

En France, Jean-Philippe Genet a été un précurseur et un acteur majeur dans les années 1990 à une époque où l'expression « Humanités numériques » n'existait pas : on parlait alors de la rencontre entre deux disciplines, l'histoire et l'informatique. Que ce soient les rencontres de l'association « Histoire et Informatique » à Orléans en 1997 et à Toulouse en 1998, ou la Xe conférence TeX européenne à Saint-Malo en mars 1998, le groupe de chercheurs en France s'intéressant aux textes numériques était restreint : parmi eux citons Martine Cocaud (Rennes II) et Jacques André (Rennes I et *Les Cahiers Gutemberg*) pour Rennes. Les bases de données et le traitement quantitatif des données, très à la mode dans ces années là, ne répondaient pas à la demande d'historiens soucieux de préserver le texte de leurs sources primaires. En 1999, la revue *Computers*

and the Humanities publie un numéro sur la TEI (*Text Encoding Initiative*)[1] pour célébrer les dix ans de ce nouveau format : Lou Burnard et Laurent Romary entre autres participent à cette publication anglophone.

Ce sont ces premiers travaux qui apportent un nouveau questionnement. Comment l'usage de l'ordinateur affecte non seulement le savoir-faire et les compétences mais aussi la manière de penser les rapports entre les disciplines et ses différents acteurs ? Pourquoi l'informatique change notre vision de certains textes et permet d'une part de nouveaux enrichissements grâce à un nouveau regard, et d'autre part, de renouveler les connaissances grâce à de nouveaux langages et des nouveaux logiciels ? En quoi la structuration des données peut-elle aider à formaliser une pensée ?

Précurseuse en Humanités numériques, dès 1998, j'ai élaboré une méthode numérique avec des outils informatiques artisanaux grâce à l'aide de Jean-Daniel Fekete (École des mines de Nantes puis INRIA) ; en 2007, j'ai déposé ma thèse sur un site « compagnon » à une époque où la mise en ligne n'était pas dans les usages. À partir de mes expériences professionnelles et universitaires menées depuis 1996[2], cet article montre l'apport de méthodes en informatique dans un travail de recherche en histoire. Je m'appuie sur l'édition des sources, leur structuration et leur diffusion, sur les répercussions de la numérisation des textes dans l'organisation du travail – les volumes à traiter nécessitant des collaborations et des traitements informatisés. Enfin, je montre l'impact de l'usage de l'ordinateur en recherche et en quoi l'édition numérique change l'analyse historique des sources et influe

[1]. http://www.tei-c.org/

[2]. Les projets principaux sont chronologiquement : Emma du GSI Cognition de 1996 à 1997 ; mes travaux universitaires de 1998 à 2007 ; Millefeuille (Archives nationales) ACI de 2006 à 2007 ; Epistémon des BVH (CESR Tours) de 2007 à 2010 ; projet européen ENBaCH (EHESS) de 2010 à 2012 ; Archives e-LV (IHPST) en 2013 ; ANR Ampère du Centre Koyré en 2013. Les références à ces projets apparaîtront dans l'article rappelées par leur sigle.

sur les méthodes en histoire. Au préalable, je reviens sur la tension entre la technique et le savoir au cours de la Renaissance entre les imprimeurs et les humanistes. Ce retour vers les précurseurs du livre permet de mettre en évidence la collaboration entre l'humaniste, les gens de l'atelier et le libraire-imprimeur, ainsi que la naissance du concept d'auteur, un parallèle nécessaire pour comprendre la révolution scientifique actuelle[3].

2. Des humanistes de la Renaissance aux humanités numériques

Traditionnellement, une édition savante de textes historiques nécessitait le travail d'une équipe composée d'ingénieurs et de chercheurs avec des compétences diverses par exemple en latin ou en grec ou autres langues, le tout étant remis à un éditeur pour la mise en forme finale. La séparation des tâches était définie au préalable ; néanmoins, les chercheurs se démarquaient des autres : ils apparaissaient comme le « fer de lance ». L'imprégnation du sens du texte se faisait au fur et à mesure des relectures nécessaires à sa correction.

Au XXIe siècle, les nouvelles technologies, et plus particulièrement les systèmes de balisage, posent la question de la relation entre le geste technique et le projet intellectuel dans l'élaboration d'une édition numérique savante. Par exemple, un langage informatique comme XML[4] influe sur l'analyse du texte lors de la constitution de l'édition numérique et formalise une pensée scientifique[5]. Le résultat de l'édition n'est plus figé dans la pierre. Du point de vue du lecteur, l'irruption des éditeurs numériques dans l'éventail des outils informatiques au service de la recherche est sans doute une des grandes nouveautés des dernières années. Sans parler de la diffusion en ligne de

[3]. Sur le sujet, cf. Kuhn (1983).
[4]. XML pour *Extensible Markup Language* est un langage informatique.
[5]. Sur le sujet, cf. Chomsky (2009).

revues et ouvrages scientifiques opérée par les éditeurs, l'édition offre de nouvelles perspectives non seulement pour la diffusion des sources de la recherche, mais aussi pour leur exploitation par les équipes de recherche[6]. Les auteurs sont dessaisis de leur édition au bénéfice des internautes qui dynamiquement récupèrent les données, les interprètent, les ré-analysent et les rediffusent. La question du concept d'auteur et de ses droits se trouve de nouveau posée (projet BVH). Le parallèle avec la Renaissance s'impose alors, période aux rapports étroits entre le savoir-faire et le savoir.

La tension entre la technique et la recherche n'est pas nouvelle. Déjà les imprimeurs de la Renaissance connaissaient les affres d'être relégués à un rôle subalterne ; déjà les écrivains se plaignaient des libertés prises dans les ateliers sur leurs œuvres ; déjà les relations entre les écrivains et les libraires-imprimeurs étaient conflictuelles en raison également des droits donnés par le « privilège » à ces derniers. La révolution du livre était la cause de ces bouleversements car sa fabrication ne ressemblait pas à celle d'un manuscrit : l'apparition de l'imprimerie a bousculé la séparation et la hiérarchisation des tâches telle qu'elle était ordonnée pour l'élaboration d'un manuscrit entre l'écrivain et les membres d'un scriptorium qui fabriquaient l'objet (*Collaboration*, 2003). Une deuxième conséquence est l'apparition du statut d'auteur, statut que certains refusent comme Maurice Scève : pour cet écrivain, il était très difficile de discerner ce que l'on devait aux membres de l'atelier et à lui-même[7].

À la Renaissance, les ateliers des imprimeurs ne sont pas de simples lieux d'exécution : ce sont des lieux d'érudition. En France, les travaux tendent à un but principal : mettre la langue française en règle. Les auteurs demandent parfois dans leur dédicace que les lecteurs fassent part de leurs critiques et suggestions à leur imprimeur. Prenons

[6]. Extrait de la lettre d'information de l'EHESS, Burghart et Dufournaud (2011).
[7]. Sur le sujet, lire le chapitre de Michèle Clément (Keller-Rahbé, 2010, 115-138).

l'exemple de Rabelais longuement étudié par Mireille Huchon : quelle est l'attitude de l'écrivain vis-à-vis de l'impression de son œuvre et comment corrige-t-il les épreuves de ses ouvrages ?[8]

Le plus souvent possible, Rabelais se trouve physiquement dans l'atelier de l'imprimeur l'obligeant à résider en ville pendant de longues durées. En accord avec son imprimeur, Rabelais choisit les caractères et l'orthographe qui lui semblent convenir le mieux à son texte. Il ne se limite pas à opérer des choix : il corrige l'impression au fur et à mesure de l'élaboration de l'ouvrage et surveille les épreuves. Il a également travaillé pour d'autres auteurs : Rabelais corrige les épreuves de Clément Marot chez François Juste son imprimeur où il fréquente Du Saix et Barthélémy Aneau. Les deux activités, auteur et correcteur, sont liées et entremêlées. Les auteurs présents dans les ateliers travaillent sur leurs œuvres mais aussi avec d'autres auteurs, cette ambiance stimulante dynamisant le groupe et donnant les résultats que l'on sait.

Si nous pouvons parler de complicité, voire de symbiose, entre l'auteur et les gens de l'imprimerie qui travaillent ensemble dans l'atelier de l'imprimeur[9], si le travail sur l'orthographe, la ponctuation et la grammaire s'effectue de façon collaborative, les gens de l'atelier ont parfois une vision différente du texte que les auteurs.

> *Les paroles énigmatiques du vieil poète... ont déconcerté les imprimeurs qui ont transformé les vaticinations dont l'obscurité était voulue par Rabelais en un*

---

[8]. Huchon (1981) et particulièrement les pages 20, 21, 50, 77, 95, 352, 361. Le mérite de Mireille Huchon a été de discerner ce qui revenait à l'imprimeur et ce qui appartenait à l'auteur.

[9] L'atelier comprend le compositeur travaillant directement sur le texte manuscrit (ou une version déjà imprimée) mais aussi le correcteur et le chef d'atelier (ou prote). Les différentes séquences de composition et de fabrication du livre sont décrites dans Jimenes (2009, 215-238) : « On aperçoit alors la patronne (Charlotte Guillard), les érudits (Miré, Tilmann, Mangeard), le correcteur et les compositeurs se partager le travail en une longue chaîne de collaboration. » Jimenes (2009, 235).

> *langage plus intelligible, modifiant ponctuation et texte.* (Huchon, 1981, 49)

En fait, les interventions sur le texte se font avec accord ou pas de l'imprimeur, de l'auteur, de correcteurs dans l'atelier, de compositeur ou du prote. Les conflits peuvent être nombreux : ils semblent qu'ils soient souvent le résultat de corrections opérées chez l'imprimeur sans l'approbation de l'auteur éloigné de l'atelier par ses affaires et ses voyages : par exemple, après 1550, Rabelais change d'imprimeur car il est en désaccord avec ses pratiques orthographiques et grammairiennes, et ses corrections. La mésentente entre l'auteur et son imprimeur s'affiche jusque dans le livre incriminé : Rabelais accuse son imprimeur d'être responsable des coquilles ; il parle d'imposture des traducteurs et des imprimeurs car les correcteurs interviennent sur son texte lors de la composition.

Cette mise en accusation de l'imprimeur par l'auteur est fréquente dans ces premiers temps de l'imprimerie : citons Dolet qui se querelle avec son imprimeur Pierre de Tours (Huchon, 1981, 95) et Pierre Boaistuau qui avoue au lecteur, en tête de son *Théâtre du monde* :

> *Au reste, quant à l'orthographe pour n'estre veu ou trop curieux innovateur, ou trop superstitieux imitateur de l'Antiquité, je l'ay laissé à la discretion de l'imprimeur, comme aussi avois-je faict à mes oeuvres precedentes.* (Roudaut, 2003, 28-30)

Par cet aveu, il dégage sa responsabilité quant aux erreurs éventuelles. Quant à Joachim Du Bellay, il est fort critique vis-à-vis de certains imprimeurs : sans les nommer, il expose au lecteur « l'avarice et impudence de certains imprimeurs »[10].

À la Renaissance, l'édition imprimée finale est le produit de plusieurs mains, le fruit d'une collaboration et la rencontre entre les ingénieurs et les humanistes. Faisons le parallèle aujourd'hui en histoire.

---

[10]. Sur le sujet, lire le chapitre de Florence Bonifay dans Keller-Rahbé (2010, 57-69).

3. La structuration des données historiques

Entre utiliser la messagerie pour correspondre avec un collègue ou un étudiant, communiquer sur un carnet de recherche, chercher une information et trouver un outil adéquat pour aider à l'analyse d'un texte, l'éventail informatique est particulièrement large. Ici notre intérêt se porte sur la recherche d'un moyen pour éditer et publier qui permet également une aide à penser grâce aux nouvelles technologies. Si l'historien-ne veut avoir accès aux résultats des traitements informatiques, aujourd'hui encore, le chercheur – ou la chercheuse – est obligé d'interagir avec l'ordinateur[11]. Pour y parvenir, un format existe qui repose sur des balises : l'encodage des textes en XML/TEI. Des recommandations mises au point par des chercheurs depuis les années 1980, permettent maintenant d'échanger des données et de profiter de l'accumulation des apports de la communauté en sciences humaines. Parmi les avantages de ce format, et à la différence des bases de données, saluons celui qui consiste à préserver le document original, à ne pas l'atomiser, ce qui permet de revenir sur la source à tout moment sans avoir à tenter de la reconstruire : une seule saisie permet de multiples traitements informatiques ultérieurs. Si la TEI permet d'éditer et de produire des textes, a priori elle ne semble pas satisfaire la demande d'historiens soucieux de montrer le processus mental de leur travail.

Dans ces mêmes années 1980, les outils de traitement de texte (comme Word) ou des formats (comme PDF) sont conçus pour répondre au besoin de mettre en forme des textes : en conséquence, ils n'accordent aucune place à la modélisation de la structure et du contenu. La saisie de la transcription avec un éditeur graphique permet d'obtenir un simple calque du manuscrit papier : l'ordinateur ne pouvant pas comprendre les textes saisis avec un éditeur graphique comme Word, cette version n'est donc pas exploitable. La

[11]. Ce domaine de l'informatique se nomme interaction humain machine (IHM) dont une définition peut être : simuler l'interaction idéale pour capturer le modèle mental de l'utilisateur.

question est de savoir s'il existe des moyens de rendre compte explicitement d'une analyse du document original « et même de la visualiser par un type de présentation qui ne soit plus une simple reproduction-transcription, mais une édition issue d'un traitement[12] ». Pour y répondre, revenons sur le manuscrit à partir des réflexions des généticiens du texte.

Depuis que le *codex* a remplacé le *volumen*, le texte qui fait sens n'est plus linéaire. La structure fait apparaître une arborescence, du chapitre à la ligne en passant par le paragraphe :

> *...sauf exception, on ne peut se dispenser de réaliser une transcription des manuscrits qu'on étudie. A moins d'un manuscrit sans ratures et ne posant aucun problème de lisibilité... une partie essentielle du travail du chercheur consiste à déchiffrer le manuscrit et à le comprendre, c'est-à-dire à en dégager l'organisation spatiale et à reconstituer à partir de celle-ci la chronologie des opérations d'écriture qui ont donné naissance au manuscrit tel qu'il nous est parvenu. Cette analyse se concrétise dans une transcription... Ce travail n'est pas trivial – il suppose un haut degré de compétence dans l'interprétation des traces laissées par l'écriture.* (Gresillon, 1990, 143)

La transcription est donc déjà le résultat d'une première analyse et doit être faite par le chercheur.

Quant au travail d'encodage, il est également une approche scientifique du document car il révèle, entre autres, le processus d'écriture de l'auteur. La complexité de l'encodage peut aider à révéler un processus d'écriture qui aurait pu jusque-là nous échapper. C'est pourquoi par un choix délibéré, il est important de s'intéresser à l'histoire même de l'écriture du texte. Cette histoire passe par l'étude des différentes strates. Pour l'encodage, la révélation de ces strates se fait notamment par l'étude des ratures et ajouts.

---

[12]. Cette question est posée par Jean-Louis Lebrave (Gresillon, 1990, 145).

Lire les textes à la lettre, « c'est relever la trace de la dynamique langagière » (Gresillon, 1990, 156).

Wioletta Miskiewicz (projet e-LV) ne dit pas autre chose quand elle écrit : « l'encodage génétique […] vise à rendre le temps originaire de la production du contenu intelligible » (Miskiewicz, 2013). Elle précise aussi que l'encodeur – ou l'encodeuse – est « l'accoucheur du texte ». La philosophe met en évidence la théorie des processus effectifs de l'écriture – un produit de la pensée et de la lecture – la perception d'une signification possible de l'écriture. Finalement, elle pose la question :

> *Ce ne sont plus les textes a priori qui déterminent l'encodage, mais l'encodage qui au fur et à mesure de l'avancement de l'encodage donne le texte ?*

L'encodage est en conséquence une activité scientifique et l'encodeur est en quelque sorte un hyperlecteur qui a pour tâche, pour construire le texte, de rendre au mieux les contenus communiqués par le document et de se laisser guider par son objet. Son intentionnalité est celle de tout lecteur possible d'un document et non celle de son créateur.

Donnons un exemple tiré de la génétique textuelle. Pour les spécialistes, il y a deux types d'ajouts appelés « variantes ». La variante d'écriture est celle qui intervient dans le flux même de l'écriture, c'est-à-dire quand le segment du texte est raturé et aussitôt remplacé à droite sur la même ligne. Toute écriture placée ailleurs, qu'elle soit dans l'interligne, en marge, etc., est une variante de lecture (*Cf.* Gresillon, 1990, 30-82). Il est possible ici de parler de substitution au sens génétique :

> *Les substitutions génétiques sont orientées par l'ordre chronologique de leurs éléments. Le remplacement de A par B n'est pas la même substitution que le remplacement de B par A…* [Il faut pouvoir] *reconstituer la chronologie interne de ces substitutions. Ce sont là des contraintes fortes, qui supposent la construction d'hypothèses sur le déroulement chronologique de la genèse et qui impliquent une analyse poussée des documents.* (Gresillon, 1990, 156)

L'exploitation automatique du texte numérique permet de construire un véritable dictionnaire constitué des ajouts et des ratures afin de voir les manières d'écrire. Ce dictionnaire rend visibles les variantes de toutes les éditions et de tous les chapitres d'un même document : une préconisation appelée par les généticiens textuels, une véritable analyse lexicale.

Prenons un autre exemple portant sur l'encodage sémantique (projet ENBACH). Dans les règles typographiques actuelles, le discours oral se repère par des guillemets, les titres de livres par le style italique, etc. ; dans les documents en ligne, la couleur peut être ajoutée. Les éditions de textes structurés reprennent ces règles afin de rendre possible une lecture intuitive du texte en ligne grâce à cet effet visuel. En amont, il s'agit de distinguer les discours et les citations : les paroles et les écrits rapportés. Dans la grammaire de la TEI, des éléments décrivent les écrits rapportés mot pour mot et les discours oraux ou paroles rapportées – « speech » en anglais – de type direct ou indirect dans lesquels les locuteurs sont repérés. Les écrits rapportés peuvent référencer des ouvrages : c'est le « fait de donner le titre d'une œuvre et/ou le nom d'un auteur auxquels on renvoie, qui accompagnent, ou non, une citation » (Compagnon, 1979). Se pose alors souvent le problème de réunir l'ensemble de ces informations ténues, éparpillées et tronquées. La solution se trouve dans la structure à démêler de ces données emboîtées et enchevêtrées[13].

Dora De Lima, doctorante en histoire à l'Université Paris 1, a étudié l'application de l'interintextualité dans le format TEI[14] (projet ENBACH). Reprenons son exemple sur la référence bibliographique extraite du texte *La vie de Marillac* résumé ici. Le premier extrait fait référence à l'*Histoire de la Maison de Bourbon* de Gilbert de Marillac et le second aux *Desseins de professions nobles et publiques*… d'Antoine de Laval en 1605. Dans une lecture attentive du manuscrit, il ressort que pour l'auteur, Nicolas Lefèvre de

[13]. Sur l'hétérogénéïté montrée du récit, cf. Authier-Revuz (1984).
[14]. (De Lima, article en ligne).

Lezeau, l'*Histoire de la Maison de Bourbon* de Gilbert de Marillac prévaut sur les *Desseins de professions nobles et publiques* d'Antoine de Laval. Cependant, l'encodage change la donne : d'une part, l'auteur se trompe sur la référence de l'*Histoire de la Maison de Bourbon* qu'il donne à Gilbert de Marillac alors que cette œuvre est écrite par Guillaume de Marillac ; d'autre part il apparaît que l'auteur possède manifestement l'ouvrage d'Antoine de Laval qu'il référence précisément en pointant l'épître liminaire.

Cet exemple n'est pas isolé et l'encodage rend systématiques ces apports qui peuvent se répéter à l'envi tout au long du travail dans un processus accumulatif : structurer un texte apporte du sens à ce texte.

Au fur et à mesure de ce travail fastidieux d'encodage, la bibliothèque de l'auteur s'édifie alors sous nos yeux (projet ENBACH). En balisant toutes les références bibliographiques peu ou semi-structurées, se dresse la liste des lectures de l'auteur. Si l'ouvrage de Jean Bodin *La République*, est bien cité, les autres références comme l'éditeur, l'imprimeur, le lieu et la date d'impression, sont incomplètes : elles n'apparaissent pas dans le texte mais peuvent faire l'objet d'un développement en tant que métadonnées. À partir de cette liste, le processus d'écriture de l'auteur apparaît tel un véritable catalogue de bibliothèque.

Une autre perspective de cet encodage est l'indexation des références bibliographiques de l'auteur (projet ENBACH). Cette nouvelle forme de classification n'est pas vraiment une nouveauté, ce sont les possibilités qu'offre l'ordinateur qui ouvrent de nouvelles perspectives. À la Renaissance, si la révolution du livre imprimé a favorisé un retentissement intellectuel des œuvres des humanistes par leur diffusion, c'est le nouveau type d'activité intellectuelle que déclenche cette nouvelle technique comme le classement des mots pour préparer les index et l'élaboration des tableaux et des renvois, qui a provoqué un renouveau de la pensée intellectuelle. Quand Elisabeth Eisenstein prend l'exemple de Bernardo Machiavelli qui passa neuf mois en 1475 à établir la liste des villes, des provinces, etc., elle

assure que l'établissement de cet index impliqua à son auteur des recherches particulières inhabituelles. Au-delà de la méthode de travail et des nouvelles techniques, c'est la nature même de l'information scientifique qui change (Revel, 1982).

Éditer un document pour le publier en ligne n'est pas un geste anodin ; il provoque dans les équipes des débats passionnés sur l'aspect heuristique de l'encodage pour la recherche, sur la gestion sociale des projets et sur le processus de l'encodage : les frontières entre les différents métiers de la recherche se délitent. La révolution numérique concerne autant les pratiques, le processus intellectuel que le partage des données, les compétences reconnues des différentes disciplines et les collaborations.

4. Les nouvelles perspectives de recherche

Le travail de l'édition numérique augmente les capacités d'analyses et les résultats scientifiques, ce qui nécessite le passage vers des méthodes nouvelles d'aide à l'analyse facilitée par l'apport d'outils informatiques.

La numérisation massive des livres et des sources aux archives permet de mieux tracer l'information du document informatif au texte final : le lecteur dispose de l'image des sources ce qui lui permet de pouvoir vérifier les données, et de maintenir la chaîne de raisonnement en « cliquant » sur l'ordinateur. Si le processus analytique est connu, les analyses deviennent alors reproductibles. À partir des mêmes sources, si les analyses reproduites ne correspondent pas, les controverses pourront alors être résolues : ici les idées seront corroborées à des faits et les débats seront connectés à des faits. La recherche en histoire reproductible oblige les chercheurs à revisiter les sources. Lors de mes travaux universitaires, j'ai mené une première expérimentation en la matière. J'ai pu en conclure que d'une part l'encodage en XML/TEI permettait d'organiser la réflexion scientifique : construction de l'esprit et formalisation de la pensée ; d'autre part l'encodage lexical puis sémantique était un processus identique à

l'édition d'un livre : lecture du sens, du fond du texte mais aussi sur la forme à des niveaux différents. Grâce à la traçabilité et en gardant le contexte, mon raisonnement était étayé ; grâce à une lecture dynamique sur le site « compagnon », l'internaute pouvait en remonter la provenance.

Si l'encodage semble être une solution pour interagir avec l'ordinateur, le jugement humain reste nécessaire afin d'explorer des données de masse à l'aide de procédés comme la visualisation d'informations, domaine de l'informatique en pleine expansion. Le principe de la visualisation peut se résumer ainsi : lorsqu'on a des données et qu'on cherche des questions intéressantes, on doit explorer. Il s'agit d'utiliser nos capacités visuelles pour détecter des phénomènes inattendus. Aux États-Unis, dans les années 1980, Ben Shneiderman a fondé le laboratoire HCIL (*Human Computer Interaction Lab*) de l'Université du Maryland. L'extraction de connaissances ou *text mining* est une de leur spécialité (Plaisant, 2006) ; ces équipes construisent des interfaces pour visualiser et explorer les textes[15].

La visualisation est une aide à l'analyse (Fekete et Dufournaud, 2006 et 2000). Si la spécialité est ancienne, la nouveauté réside dans l'interactivité qui l'enrichit énormément[16]. Cependant, pour utiliser pleinement ces technologies, les données doivent être structurées. Prenons un exemple, celui de la *timeline*, ou frise chronologique, un outil de visualisation qui fournit un ensemble de points ordonnés dans le temps qui peuvent être liés à des éléments des textes transcrits pour créer un alignement temporel des discours. Pour être explicite, la constitution d'une telle frise est rendue possible dès lors que des éléments de temporalité sont contenus dans un texte notamment les dates et le temps. Puis les évènements sont étiquetés par mots-clés ; au regard de certaines caractéristiques du texte, les créateurs de la frise tout comme l'internaute pourront observer les imbrications et enchevêtrements qui ne doivent

[15]. Pour la France, voir le site de l'équipe AVIZ de INRIA http://www.aviz.fr/

[16]. Les travaux actuels reprennent ceux de Jacques Bertin (1967).

rien au hasard, équilibres et rapports de force pouvant transparaître dans la frise avant même toute lecture du texte. En ligne, la navigation par l'internaute lui permet par ce système d'explorer les textes. Tout ceci confère un rythme et un relief à la frise, en un simple coup d'œil. Enfin, l'affichage de la chronologie des événements permet une entrée innovante dans les textes, navigation impossible à rendre sur le papier.

La navigation et l'exploration, voire même l'extraction des données sont rendues possibles grâce à la construction et le croisement des index, des réseaux sociaux, et de multiples listes : le travail récupéré permet une analyse ultérieure des mêmes sources. Cependant, ces éditions restent trop souvent conçues sans grand effort éditorial : le modèle reste encore le livre. Il s'agit de créer une nouvelle façon de lire ces éditions en ajoutant des vidéos, des photos, des musiques, des textes qui enrichiront l'ensemble. Reconstruire les lieux, montrer les portraits des acteurs, donner à entendre des conférences favorisent une ouverture à ces documents souvent trop obscurs. Le croisement des supports doit être envisagé pour faire de l'édition numérique une nouvelle forme de lecture et de compréhension du texte à multiples versions.

Grâce à l'informatique qui oblige à plus de rigueur dans la réflexion scientifique, l'apport des textes numériques est potentiellement bien plus important qu'un livre, un fac-similé ou une simple édition HTML (projet Ampère). Si la tradition de l'édition des textes – l'ecdotique – est bien respectée, les nouvelles méthodes d'analyses d'exploration informatique de sources sont d'ores et déjà envisagées. L'encodage des sources permet une démultiplication des points de vue sur les sources historiques et favorise les explorations d'hypothèses, une fois le travail initial terminé. De plus, il est alors possible d'explorer et de naviguer facilement dans les données : ces opérations sont rapides, incrémentales et réversibles, et à la fin, on accède directement à la source elle-même. Des outils de visualisation permettent de construire des graphes : des réseaux et des arbres comme les généalogies. Enfin, le

processus de recherche est traçable, de la source (image en ligne) au résultat de la pensée du chercheur qui peut alors être interprétée dynamiquement par le lecteur-internaute.

5. Conclusion

À la Renaissance, faire coïncider les idées à la réalité fut l'œuvre des humanistes-ingénieurs. Aujourd'hui, nous vivons un autre bouleversement au moins de même ampleur : la révolution numérique oblige à des méthodes d'investigations rigoureuses, des analyses vérifiables et reproductibles, au sens de Karl Popper (1973). Le numérique est à la source d'un changement des modes de pensée et d'écriture, passant de la linéarité d'un récit d'autorité à une structuration rendant l'analyse traçable et vérifiable.

Après l'ère de la mise en page, nous sommes entrés dans celle du contenu : structurer un texte rend explicite le sens du texte. En histoire, nous devons utiliser des outils qui permettent l'exploration d'une quantité de données en augmentation constante. L'ordinateur oblige le chercheur à changer le processus méthodologique de l'élaboration des savoirs : il doit apprendre à programmer certains langages comme XML et posséder un bon niveau de culture numérique.

> *Tout le monde se mesle d'escrire l'histoire et n'y cognoist rien. Bref il y en a fort peu qui scachent ce qu'ils font et ce qu'ils devroient faire. Et comme il y en a fort peu qui soient capables de juger de ce qui est escrit ; c'est une maxime tres certaine et qu'il faut tenir pour indubitable, que soli artifices possunt judicare de artibus* [seuls les artistes peuvent juger les arts]*. Il y a trois choses dont personne ne doit juger s'il ne les a faites luy mesme, sçavoir d'escrire, de prescher, de plaider. A entendre un bon advocat qui plaide il semble qu'il n'y a rien si aisé que d'en faire autant ; mais quand on en vient au fait et au prendre, hoc opus, hic*

> *labor est* [voici la difficulté, voilà l'épreuve] *; les plus habiles y sont bien empeschez*[17]*.* (Guy Patin, 1641, 313)

Au XVIe siècle, l'imprimerie a entraîné « une mutation des manières de lire et des systèmes de pensée » (Roudaut, 2003-11) ; elle a autorisé l'accumulation, l'échange et la communication des connaissances. Aujourd'hui, l'édition numérique doit être le fruit de collaborations et de partages, pour permettre la réflexion et le jugement par tous les internautes (chercheurs ou non). À la différence du livre, grâce aux outils informatiques dynamiques, rien n'est figé, la création d'une œuvre n'appartient plus à l'auteur : sur le web, elle se régénère à l'infini.

Remerciements

Je remercie Valérie Gratsac-Legendre (EHESS) pour le soutien qu'elle m'a apporté tout au long du projet ENBACH ainsi que pour avoir relu les ébauches de cet article et m'avoir donné des pistes de réflexion. Les fructueuses collaborations transparaissent dans ce texte : en informatique avec Jean-Daniel Fekete (INRIA, Aviz) ; en littérature avec l'équipe du projet BVH à l'université de Tours sous la direction de Marie-Luce Demonet ; en philosophie avec Wioletta Miskiewicz. Enfin, je remercie Dora De Lima pour son investissement sur les questions d'intertextualité dans le projet ENBACH, connaissances qu'elle a su partager.

Bibliographie

Authier-Revuz J. (1984, mars). Hétérogénéité(s) énonciative(s). *Langages*, n° 73.

Bertin J. (1967). *Sémiologie Graphique. Les diagrammes, les réseaux, les cartes.* La Haye, Mouton, Gauthier-Villars, EHESS, Paris.

Burghart M., Dufournaud N. (2011). Édition électronique de sources : XML et Text Encoding Initiative (TEI) à l'École. *La*

[17]. Les deux textes latin-français ont été traduits par Perrine Mane que je remercie.

lettre d'information de l'EHESS, n° 38, http://lettre.ehess.fr/1475

Chomsky N. (2009). *Le langage et la pensée.* Payot, Paris.

Compagnon A. (1979). *La seconde main ou le travail de la citation.* Seuil, Paris.

De Lima D. (à paraître). En-coder la critique historique d'une source. L'intertextualité dans l'édition numérique de la Vie de Michel de Marillac. http://halshs.archives-ouvertes.fr/halshs-00947220

Fekete J-D., Ogilvie D. (2008). Le projet et la plate-forme Millefeuille recherches et outils informatiques pour de nouveaux usages des almanachs. *Bibliothèque de l'Ecole des Chartes,* vol. 166, n° 1/2008, Paris, p. 89-98.

Fekete J-D., Dufournaud N., (2006). Comparaison d'outils pour la visualisation de sources historiques codées en XML/TEI. *Document numérique, Visualisation pour les bibliothèques numériques,* Hermes-Lavoisier, vol. 9, n° 2/2006, Paris.

Fekete J-D., Dufournaud N. (2000). Compus: visualization and analysis of structured documents for understanding social life in the 16th century. *Proceedings of DL'00: the fifth ACM Conference on Digital Libraries*.

Genet J.-P., Zorzi A. *et al.* (2011). *Les historiens et l'informatique : un métier à réinventer.* Ecole Française de Rome, Rome.

Gresillon A. *et al.* (1990). *Proust à la lettre. Les intermittences de l'écriture.* Du Lérot, Tusson (Charente).

Huchon M. (1981), Rabelais grammairien. De l'histoire du texte aux problèmes d'authenticité, Tome XVI. Etudes Rabelaisiennes, Librairie Droz, Genève.

Jimenes R. (2009). Pratiques d'atelier et corrections typographiques à Paris au XVIe siècle. L'édition des œuvres de Saint Bernard par Charlotte Guillard. Passeurs de textes. Imprimeurs et libraires à l'âge de l'humanisme, *Etudes et rencontres de l'Ecole des Chartes* n° 37, Champion, Paris.

Keller-Rahbé E. *et al.* (2010). *Les arrière-boutiques de la littérature. Auteurs et imprimeurs-libraires aux XVIe et XVIIe siècles.* Presses Universitaires du Mirail, coll. « Essais de littérature Cribles XVIe-XVIIIe », Toulouse.

Kuhn T. S. (1983). *La structure des révolutions scientifiques.* Flammarion (Champs), Paris.

Miskiewicz W. (2013). « Texte » versus « Document ». Sur le platonisme dans les humanités numériques et sur la maïeutique TEI des textes. *TEI Conference and Members Meeting,* October 2-5/2013, Rome (Italy), http://digilab2.let.uniroma1.it/teiconf2013/program/papers/abstracts-paper#C154

Plaisant C., Rose J., Yu B., Auvil L., Kirschenbaum M., Smith M., Clement T., Lord G. (2006, January). Exploring Erotics in Emily Dickinson's Correspondence with Text Mining and Visual Interfaces. *Proceedings of the 6th ACM/IEEE Joint Conference on Digital Libraries,* JCDL 06, p. 141-150.

Popper K. (1973). *La logique de la découverte scientifique.* Payot, Paris.

Revel J. (1982). Elizabeth L. Eisenstein : La culture de l'imprimé. *Le Débat*, 1982/5 n° 22, p. 178-192.

Roudaut F. (2003). *Le livre au XVIe siècle. Eléments de bibliologie matérielle et d'histoire*. Champion, Paris.

(2003). La collaboration dans la production de l'écrit médiéval. *Actes du XIIIe colloque du Comité international de paléographie latine*, École des Chartes, Paris.

Les Papiers de Guy Patin, Vienne, codex 7071 (Hohendorf 133), manuscrit transcrit en ligne, http://patin.ehess.fr/