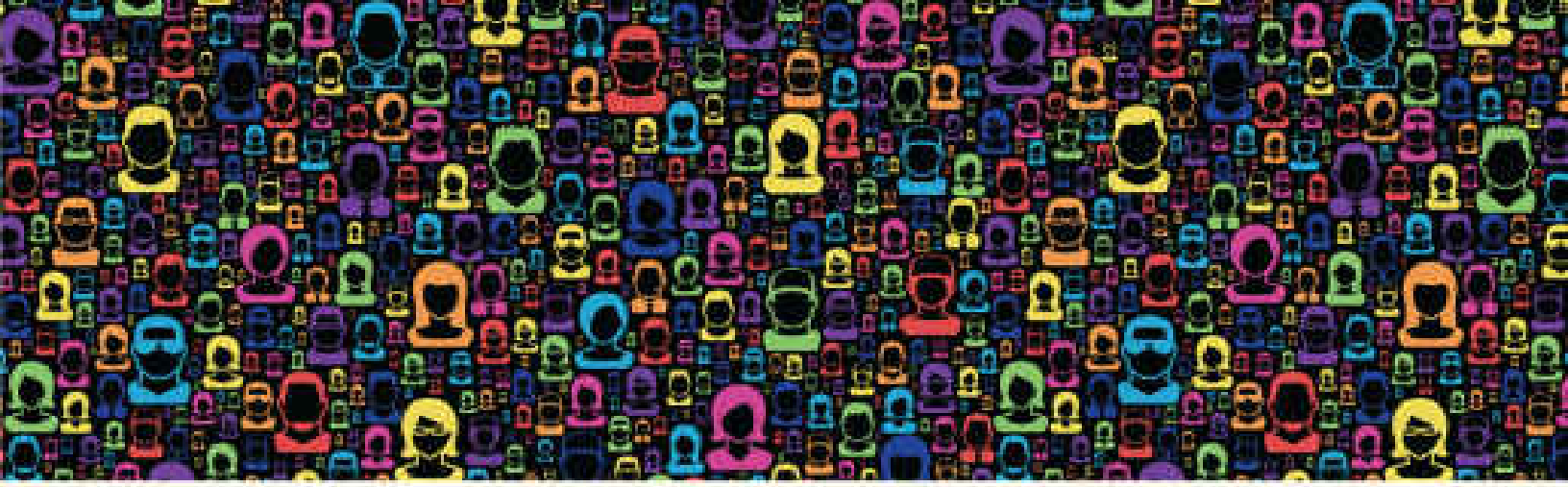

Expérimenter les humanités numériques

Sous la direction d'Étienne Cavalié, Frédéric Clavert, Olivier Legendre et Dana Martin

La collection « Parcours numériques » est accessible gratuitement en édition augmentée sur parcoursnumeriques-pum.ca.

Univers en perpétuelle expansion et au foisonnement chaotique, Internet offre un nombre incalculable d'outils, dont l'exploration paraît parfois hors de portée. Dans le paysage des sciences humaines, les blogs, les logiciels bibliographiques, les bases de données, les éditions en ligne et les wikis, tous ces objets qui éveillaient notre curiosité il y a une décennie, sont devenus aussi anodins qu'omniprésents. Mais comment bien s'en servir ?

Les appréhensions face à ces outils – et leur simple mais robuste méconnaissance – sont encore largement répandues. Or on ne peut plus ignorer leur intérêt, voire leur nécessité, et les chercheurs qui s'y essaient ne savent souvent pas par quel bout attraper ces logiciels nouveaux. C'est à cela que cet ouvrage veut les aider, de façon simple et précise, et il entend le faire sans en cacher les difficultés, mais sans dissimuler non plus qu'elles sont désormais connues, donc surmontables, et que, dans la majorité des cas, le résultat vaut tous les efforts à consentir.

Étienne Cavalié est conservateur à la Bibliothèque nationale de France. **Frédéric Clavert** est maître assistant à l'Université de Lausanne. **Olivier Legendre** est conservateur à la Bibliothèque Clermont Université. **Dana Martin** est maître de conférences en allemand à l'Université Blaise Pascal de Clermont-Ferrand.

Expérimenter les humanités numériques

Des outils individuels aux projets collectifs

Sous la direction de
Étienne Cavalié,
Frédéric Clavert,
Olivier Legendre
et Dana Martin

La collection «Parcours numériques» est accessible gratuitement en édition augmentée sur parcoursnumeriques-pum.ca. La version enrichie comprend une bibliographie, des notes ainsi que de nombreux contenus complémentaires (vidéos, illustrations, etc.). La collection est dirigée par Michaël E. Sinatra et Marcello Vitali-Rosati.

Merci au Centre de recherche interuniversitaire sur les humanités numériques (CRIHN) ainsi qu'à la Chaire de recherche du Canada sur les écritures numériques de soutenir la publication des livres de la collection «Parcours numériques».
Merci également aux partenaires qui ont soutenu la préparation de ce livre: la Bibliothèque Clermont Université, le Center for Contemporary and Digital History (C2DH, Université du Luxembourg), l'Institut historique allemand Paris, le laboratoire Communication et Sociétés de l'Université Clermont Auvergne, la Maison des sciences de l'homme Clermont-Ferrand et la Bibliothèque nationale de France.

Couverture: ©bgblue/iStockphoto.com

Mise en pages: Yolande Martel

ePub: Folio infographie

Catalogage avant publication de Bibliothèque et Archives nationales du Québec et Bibliothèque et Archives Canada

Cavalié, Étienne, 1979-

Expérimenter les humanités numériques: des outils individuels aux projets collectifs

(Parcours numériques)

Comprend des références bibliographiques.

Publié en formats imprimé(s) et électronique(s).

ISBN 978-2-7606-3802-0

ISBN 978-2-7606-3803-7 (PDF)

ISBN 978-2-7606-3804-4 (EPUB)

1. Sciences humaines numériques. 2. Sciences humaines – Recherche – Information électronique – Guides, manuels, etc. 3. Sciences humaines – Recherche – Ressources Internet – Guides, manuels, etc. I. Clavert, Frédéric. II. Legendre, Olivier, 1976- . III. Martin, Dana, 1975- . IV. Titre.

V. Collection: Parcours numériques.

AZ105.c38 2017 001.30285 C2017-941257-4

C2017-941258-2

Dépôt légal: 3e trimestre 2017

Bibliothèque et Archives nationales du Québec

www.pum.umontreal.ca

Les Presses de l'Université de Montréal remercient de leur soutien financier le Conseil des arts du Canada et la Société de développement des entreprises culturelles du Québec (SODEC).

Financé par le gouvernement du Canada | Canada

Table des matières

Note: L’intégralité des chapitres 3, 4, 5, 11 et 12 se retrouve sur
www.parcoursnumeriques-pum.ca/experimenterleshumanitesnumeriques

Introduction

Ce livre est né de l'envie de combler un manque. Pas un manque de réflexion: les humanités numériques sont omniprésentes dans le discours scientifique. Les discussions sur leur définition, leur périmètre, leurs ruptures et leurs continuités abondent. Pas un manque technique, non plus: non seulement les humanités numériques offrent aux chercheurs pléthore d'outils, mais si l'on considère ceux dont il sera question ici, les documents techniques permettant leur prise en main et leur utilisation (tutoriels, procédures en ligne, etc.) sont nombreux.

Cette abondance laisse toutefois une lacune béante. Sur un terrain que l'on n'ose plus dire neuf, mais où bien des chercheurs hésitent encore à s'avancer, sur un terrain que la technicité, le savoir-faire et l'entreprise scientifique explorent de conserve et à tâtons, il s'avère qu'une seule chose est irremplaçable: l'exemple. Ce qui, par-dessus tout, pousse un chercheur à empoigner un outil inconnu, c'est l'expérience partagée d'un devancier. Le récit d'une expérience réussie – ou d'un échec dépassé! – lui permet d'éviter les écueils les plus redoutables; de gagner le temps que les pionniers ont accepté de perdre; de se convaincre en les écoutant que le jeu (scientifique) en vaut la chandelle (numérique); bref de saisir avec confiance l'outil peu familier, en sachant déjà que ses efforts seront récompensés, et peut-être même de se prendre lui-même au jeu de l'expérimentation.

Une fois franchie cette première étape, une réflexion peut s'engager sur l'élaboration des outils et sur leurs apports, occasionnant souvent – les expériences rapportées le montrent – un réexamen des méthodes. Pour certains de ces chercheurs, il est possible qu'arrive même une seconde

étape où, passée la maîtrise des outils existants, on en vienne à en percevoir les limites au point de contribuer à l'élaboration de nouveaux outils et de nouvelles méthodes. L'enjeu est bien là, dans ce goût de l'essai et dans ce face-à-face tantôt passionnant, tantôt inquiétant avec la technique. Tout chercheur est désormais censé utiliser quotidiennement les ressources numériques, et non plus seulement comme lecteur, mais bien comme auteur. Dans le paysage scientifique des sciences humaines, les blogs, les logiciels bibliographiques, les bases de données, les éditions en ligne et les wikis, tous ces objets qui éveillaient notre curiosité il y a une décennie, sont devenus aussi anodins qu'omniprésents. Par ailleurs, la France se défend plutôt bien en ce domaine, des réalisations comme le portail de publication OpenEdition, avec les plates-formes hypotheses.org et http://www.revues.org/, étant saluées par nos confrères en Europe.

Et pourtant, en jetant un regard sur nos propres pratiques, sur les chercheurs qui travaillent autour de nous, nous voyons bien que les appréhensions face aux outils numériques (ou leur simple mais robuste méconnaissance) sont encore largement répandues. Les sept décennies écoulées depuis les premiers travaux de l'école des Annales ou de Roberto Busa n'y font rien: le numérique continue d'intimider les sciences humaines et sociales. À leur décharge, il faut reconnaître que l'informatique, plus encore depuis l'avènement d'Internet, est un univers en perpétuelle expansion, dont l'exploration paraît parfois hors de portée et le foisonnement chaotique, imprévisible et potentiellement inconfortable.

Notons aussi que, par un effet de génération ou encore par le penchant classique de bien des chercheurs en sciences humaines, l'*alphabétisation numérique*, si l'on peut ainsi désigner la familiarité avec la chose informatique, reste balbutiante chez tous ceux – les plus nombreux pour l'instant – qui n'ont pas grandi avec un mobile connecté

dans la poche. Du reste, supposer que l'usage quotidien de tels appareils confère la compétence nécessaire à l'usage scientifique du numérique est une naïveté dont scientifiques, pédagogues et informaticiens sont revenus depuis longtemps. Certains chercheurs préfèrent plutôt parler de «Facebook natives» (Clavert, 2011), peu au fait du fonctionnement des outils qu'ils utilisent: avoir un mobile dans la poche est ainsi plutôt un facteur d'analphabétisme numérique.

Quoi qu'il en soit, le paradoxe est là, bien installé: les réalisations en sciences humaines étayées par des méthodes numériques sont désormais trop présentes, et leur fécondité trop évidente, pour que l'on puisse feindre d'ignorer leur intérêt. Certaines de ces réalisations ne sont rendues possibles que par l'existence du numérique. Or de nombreux chercheurs, quand bien même ils souhaitent le faire, ne savent pas encore par quel bout attraper ces logiciels nouveaux que, de partout, font surgir les humanités numériques. C'est à cela que ce recueil veut les aider, de façon simple et précise. Il entend le faire sans cacher les difficultés qui attendent l'apprenti numérique, mais sans dissimuler non plus qu'elles sont désormais connues, donc dépassables, et que, dans la majorité des cas, le résultat vaut tous les efforts consentis.

Chacun des chapitres proposés ici se veut une initiation à un ou plusieurs des logiciels et des méthodes actuellement disponibles, qui seront présentés, évalués et critiqués à la lumière de l'expérience personnelle des auteurs. Le statut d'expert de ceux-ci repose précisément sur cette double compétence technique et didactique, qui permet de transmettre un savoir-faire théorique *et* son application pratique – condition *sine qua non* pour être non seulement audible mais crédible. L'ensemble des sujets traités constitue un tour d'horizon accessible aux novices comme aux experts. Il s'agit d'un instantané kaléidoscopique d'expériences réellement vécues, dont nous espérons qu'il

sera à la fois représentatif et instructif. Le choix des contributions a été effectué par les coordinateurs à la suite d'un appel à contributions (http://comsol.univ-bpclermont.fr/article212.html).

Expérimenter les humanités numériques: le titre le dit, l'ouvrage a été conçu comme un recueil de retours d'expériences concrètes et utiles; *Des outils individuels aux projets collectifs:* il s'inscrit dans une double perspective, en réunissant des expériences individuelles et des expériences collectives et, pour offrir une palette diversifiée, il s'organise en trois parties. «Les outils personnels» concernent les cartes mentales et les blogs personnels, l'annotation vidéo, le logiciel bibliographique Zotero, un exemple de base de données ou encore les réseaux sociaux numériques. «L'outillage collectif» présente le système de gestion de contenu Omeka, le wiki comme outil de gestion d'un projet scientifique, ainsi que l'édition électronique. Enfin, la dernière partie, «La gestion de projet», est consacrée à des projets vus sous l'angle de l'organisation: la gestion des données, les blogs communs et surtout les bases de données collectives, qui occupent actuellement une place centrale dans les réflexions et les stratégies des équipes de recherche.

Tout considéré, les exemples ici réunis offrent, par facettes, une définition des humanités numériques qui nous paraît à la fois simple et convaincante: pour nous, ce sont des outils numériques appliqués aux sciences humaines et sociales. Par leur ancrage dans la réalité de projets aboutis ou non, toujours imparfaits mais bel et bien menés, ces textes réfutent efficacement l'idée que le numérique révolutionnerait, de fond en comble, la pratique scientifique. Le numérique ne permet nullement d'inventer une science nouvelle, fondée sur des structures et des concepts purement informatiques – et désespérément inaccessibles aux non-spécialistes.

Bien au contraire: les projets retenus se situent aux antipodes de tout fantasme idéaliste ou alarmiste. Leurs artisans considèrent, à juste titre, l'informatique comme un moyen. Un moyen puissant de collecte, d'analyse et de diffusion du savoir, assez puissant pour marquer la démarche scientifique, mais pas au point d'en dénaturer les fondamentaux. Quels que soient les moyens mis à son service, l'entreprise de recherche, dans sa complexité et son exigence, s'efforce de suivre une même méthodologie. Formuler des hypothèses; collecter des données adéquates; confronter ses hypothèses initiales avec le matériau disponible, en le complétant si nécessaire; interpréter et soumettre son interprétation à la critique informelle, anonyme ou publique de ses pairs; publier et diffuser ses données et ses résultats aussi largement que possible, au-delà du cercle des seuls spécialistes, et également hors de son propre pays. Les humanités numériques font tout cela, en s'appuyant sur des instruments aujourd'hui disponibles, dont la caractéristique commune est d'être massivement informatiques. C'est tout, et déjà beaucoup, car les résultats sont remarquables.

Toutefois, même si l'on considère que le cours général du travail de recherche n'est pas altéré, les modalités de mise en place de chacune de ces opérations successives se trouvent transformées, notamment par la capacité de traitement tout à fait nouvelle qui s'offre au chercheur. Bien plus, les réalisations concrètes des humanités numériques produisent leur lot de remises en question épistémologiques. Le numérique permet de traiter une masse d'information colossale et sans cesse croissante. Ce sont de nouvelles sources qui s'offrent et qui imposent de nouveaux modes de traitement à la recherche. Continuité intellectuelle et scientifique d'un côté, mais aussi renouvellement des objets d'études et des méthodes de travail de l'autre, ces deux traits caractérisent indissociablement, pour nous, les humanités numériques.

Une autre constante doit être soulignée, dont les chapitres de cet ouvrage collectif, d'ailleurs cosignés pour la plupart, témoignent par leur contenu: dans les humanités numériques, les projets s'épanouissent dans la coopération, non seulement entre chercheurs de disciplines identiques ou différentes, mais aussi entre corps de métier complémentaires. L'ampleur de la tâche et des ressources à mobiliser, les vastes compétences nécessaires pour maîtriser l'ensemble des enjeux techniques et scientifiques, ou encore la masse des données à traiter, rendent cette coopération vitale pour la majorité des projets. Comme beaucoup de travaux collectifs ou interdisciplinaires, les projets d'humanités numériques ont pour caractéristique commune de rassembler, presque inévitablement, des professionnels de différents métiers, notamment les enseignants-chercheurs, les bibliothécaires et les ingénieurs d'études et de recherche (en informatique, en gestion de projet, etc.). Rares sont les professionnels qui peuvent se targuer de réunir les compétences de tous ces métiers, pourtant indispensables pour concevoir et mener à bien un projet scientifique et technique pertinent, original et d'envergure. Pour autant, les humanités numériques n'exigent pas que quiconque apprenne un autre métier que le sien. En particulier, pour répondre à une crainte souvent exprimée et toujours sous-jacente, les chercheurs n'ont pas à devenir des apprentis informaticiens.

En revanche, ce qu'il leur faut à tout prix, c'est un «dialogue des rigueurs». L'informatique est souvent perçue par les chercheurs comme contraignante, du fait de son caractère systématique. Or «[i]l ne s'agit pas d'opposer la rigueur du numérique à l'amateurisme foutraque du terrain, mais bel et bien de faire dialoguer deux rigueurs: celle de l'outillage informatique, qui de fait impose sa structuration, gage d'efficacité mais poids au quotidien; et celle de l'épistémologie propre à chaque démarche de chercheur, et

adaptée à chaque terrain, chaque problème» (Boulaire et Carabelli, ci-dessous).

Cette réciprocité dans la coopération, qui va de pair avec une dose généreuse d'entraide et de formation mutuelle sur le tas, est l'une des leçons majeures des projets aboutis et réussis. Il est vrai que l'informatique impose la contrainte de sa rigidité technique. Notamment quand il s'agit de concevoir et de construire des bases de données, elle exige d'emblée, de la part du chercheur, un esprit de système (dans la constitution d'un corpus, par exemple), qui n'est pas aisément atteint lorsqu'un projet démarre. Or, la rigueur et la cohérence du cheminement ne sont pas nées avec l'informatique: elles font partie intégrante de la démarche scientifique. L'informaticien les renforce en exigeant du chercheur qu'il systématise d'emblée son approche, qu'il segmente ses données et les standardise, à un stade très précoce de son travail. Le bibliothécaire, s'il est invité dans un projet, renchérit généralement sur cette exigence, avec son insatiable besoin de normalisation et d'indexation, de recours aux nomenclatures, aux listes d'autorité et aux formats structurés de métadonnées. Il n'empêche qu'informaticiens et bibliothécaires peuvent (et doivent) adapter leurs outils aux besoins des chercheurs. Non pour fragiliser leurs infrastructures respectives, mais d'une part, pour en faciliter l'appropriation et l'usage, et d'autre part, pour économiser aux chercheurs le temps d'un fastidieux apprentissage qui n'apporte rien à la recherche. Lorsque le dialogue s'instaure réellement, l'informatique et le traitement documentaire se révèlent pour ce qu'ils sont, de prodigieux dispositifs de soutien au travail scientifique, notamment pour collecter, analyser et gérer les données dans le temps.

L'ultime enseignement qui peut sortir des expériences réunies, c'est celui-là même qui nous a poussés à concevoir un recueil de contributions, de préférence à toute autre formule: tous les auteurs *ont appris en faisant*. Certains des

projets décrits ont été conçus d’emblée comme expérimentaux (Idmhand, Riffard et Walter; Chagnoux et Humbert), d’autres l’ont été par nécessité, parce qu’ils empruntaient des chemins non balisés (Boulaire/Carabelli). Tous ont fait face, au long de leur travail, à la double nécessité d’adapter un logiciel informatique à leur besoin précis, et – inversement – d’adapter leur démarche aux possibilités et aux contraintes du logiciel choisi. Tous ont été amenés à se saisir d’outils numériques nouveaux, à les prendre en main, à les transformer, et finalement à les mettre au service de leur projet scientifique en sciences humaines et sociales. Aucun ne prétend avoir déniché, ou élaboré, *l’outil parfait*. Tous concluent à la richesse de l’exploration.

PREMIÈRE PARTIE

LES OUTILS PERSONNELS

CHAPITRE 1

Les outils d’annotation vidéo pour la recherche

LAURENT TESSIER ET MICHAËL BOURGATTE

La pratique de l’analyse filmique dans le cadre de la recherche en Sciences humaines et sociales (SHS) n’est pas nouvelle. En histoire par exemple, elle est répandue et même institutionnalisée depuis des dizaines d’années. En France, les travaux de Marc Ferro (1974) et d’Henri Rousso (1987) ont, parmi d’autres, permis de montrer les bénéfices de cette pratique, qu’il s’agisse de l’étude de films d’actualité, de films de propagande et même de films de fiction. Face aux images, le travail du chercheur consiste à définir un corpus, puis à cerner les thèmes récurrents, la manière dont ils sont traités, les types de discours, les personnages représentés, etc.

Pour autant, avec la production de plus en plus massive de vidéos et leur circulation facilitée sur Internet, ce sont de nouveaux terrains d’investigation qui émergent aujourd’hui pour la recherche: YouTube, Dailymotion, Archive.org, Canal-U.tv, l’Université ouverte des humanités ou encore les Archives audiovisuelles de la recherche sont des espaces dédiés à la vidéo numérique qui offrent des quantités immenses de matériaux et d’occasions intellectuelles. Cette situation peut même susciter un vertige chez les chercheurs qui ne sont pas bien outillés pour s’emparer de ces corpus vidéo. Comment peuvent-ils investiguer, analyser, traiter ou catégoriser cette profusion inédite de documents audiovisuels? Il s’agit là d’un sujet central pour les humanités numériques (HN) dont l’un des objectifs est précisément d’outiller les scientifiques et de faciliter leurs pratiques numériques.

Sans prétendre résoudre ici l'ensemble des problématiques soulevées par la vidéo numérique, on peut suggérer que la première action possible est sans doute celle de l'annotation. La prise de notes étant le support naturel de l'activité du chercheur (qu'elle porte ou non sur la vidéo), il est capital pour lui d'avoir la possibilité d'associer ses notes, ses analyses ou ses commentaires aux images qui constituent son corpus. L'objectif de ce chapitre est donc de faire un état des lieux des pratiques actuelles, à partir d'une enquête conduite auprès de chercheurs qui travaillent sur des corpus vidéo. Nous dresserons également ici un panorama des technologies existantes et de ce qu'elles permettent en matière de production scientifique, nous inscrivant en cela dans la perspective des *Software Studies* (Gras, 2015). Pour cela, nous prendrons notamment appui sur le projet de R&D «Celluloid» dans lequel nous sommes plus directement impliqués. Dans cet article, on évaluera donc les besoins des chercheurs, on examinera les pratiques ainsi que la pertinence des réponses logicielles existantes.

Le développement des discours vidéographiques

La circulation des images animées a connu plusieurs étapes depuis une centaine d'années jusqu'à l'entrée dans ce que Régis Debray (1992) nomme la «vidéosphère», une ère dominée par les discours vidéographiques. Parmi ces étapes, on peut citer rapidement la diffusion des images analogiques au cinéma et à la télévision, puis la période de démocratisation des caméras personnelles et du magnétoscope. La poussée récente et massive des outils numériques constitue une nouvelle étape décisive dans la production d'images et leur consultation, notamment grâce à des plates-formes de partage qui permettent de consulter instantanément d'innombrables vidéos libres d'accès (productions amateurs, vidéoclips, tutoriels) ou encore des films tombés dans le domaine public (Bourgatte, 2014).

Ce changement paradigmatique est un enjeu crucial pour les HN. Les chercheurs qui prennent part à ce mouvement se saisissent actuellement des corpus d'images, voire de films ou de vidéos. Précisons d'emblée que les outils d'annotation d'images fixes ne seront pas présentés dans ce chapitre, même si on peut par exemple signaler l'existence d'outils comme A.nnotate ou ThingLink. D'autres applications comme Hypothes.is permettent plus largement d'annoter des contenus Web, qu'ils soient visuels ou textuels.

Les chercheurs s'appuient alors sur les méthodes et les outils d'abord développés pour l'exploration et l'analyse de corpus textuels. Certes, les HN ne se limitent pas à ces pratiques, et celles-ci ne leur sont pas réservées (les systèmes d'information géographique ou l'analyse des réseaux numériques en sociologie reposent, l'un et l'autre, sur des logiques semblables), néanmoins de nombreuses disciplines sont ainsi concernées. En littérature, en linguistique ou en archéologie, les travaux menés ont pris la forme de la lexicométrie qui implique l'exploration et l'analyse quantitative et qualitative des données et des écosystèmes textuels: repérage des occurrences, des correspondances, développement de typologies (Moretti, 2008), *Text Mining* pour la fouille instrumentée de textes, ou encore *Topic Modeling* en vue de la modélisation des thèmes contenus dans un texte.

Or, comme l'ont souligné les auteurs du récent volume des *Cahiers du numérique* consacré au thème des «humanités délivrées», l'un des principaux enjeux des HN est précisément de se saisir de nouvelles formes de littératies «hors du livre» (Clivaz, Vinck, 2014, p. 9): qu'il s'agisse de nouvelles formes de cultures écrites, orales, visuelles ou multimédiatiques. C'est en ce sens que nous avons engagé, en 2013, le programme de recherche Celluloid sur l'annotation vidéo et les cultures audiovisuelles sur lequel nous reviendrons à la fin de ce chapitre, et dont le

blog celluloid.hypotheses.org diffuse les résultats tandis que l'espace celluloid.camp présente les expérimentations méthodologiques. Au-delà de cette initiative particulière, ces dernières années ont vu émerger dans le champ francophone un véritable courant d'études visuelles. Il s'est notamment incarné autour de la plate-forme Culture Visuelle (http://culturevisuelle.org/) pour la diffusion des recherches tournées vers l'image et la vidéo, ainsi que la fondation d'un groupe de travail en sociologie visuelle et filmique au sein de l'Association française de sociologie (http://www.test-afs-socio.fr/drupal/GT47). En parallèle, on retiendra aussi qu'un important consortium d'acteurs du champ de l'audiovisuel (laboratoires de recherche, fonds documentaires et acteurs de l'Internet) s'est constitué dans le cadre du projet de recherche Cinecast (http://cinecast.fr/). Ce consortium a permis de déployer une réflexion d'envergure sur les enjeux liés aux discours vidéographiques et aux besoins d'outillage technologique des chercheurs.

L'annotation vidéo: un enjeu pour la recherche

L'une des pratiques fondatrices du travail de recherche en SHS réside, on l'a dit, dans la prise de notes et l'annotation de matériaux. Annoter un texte est une pratique, pour ainsi dire, aussi vieille que l'écrit lui-même qui s'est vue démultipliée par les possibilités du numérique. Originellement, il s'agit d'insérer un commentaire, des éléments d'analyse ou une note marginale sur un document textuel, graphique ou iconographique. Il peut exister différents types d'annotations: écrites, visuelles ou multimédiatiques. Leur point commun est d'anticiper un usage asynchrone. Autrement dit, une annotation est un commentaire produit pour être consulté plus tard, par soi-même ou par un tiers.

Face à cette pratique immémoriale, le principal apport du numérique réside dans l'insertion d'une annotation qui est prise entre deux balises informatiques et donc techniquement indépendante et dissociable du contenu premier (Cousins, Baldonado et Paepcke, 2000), ce qui n'était pas possible avec l'annotation traditionnelle, imprimée dans le papier et donc difficilement dissociable du contenu annoté.

Comment transposer ces principes à l'annotation vidéo? Cette question n'a pas encore trouvé de réponse définitive. Si une plate-forme comme YouTube propose un service d'annotation de ressources vidéo, celui-ci correspond à des usages ludiques et ne vient pas satisfaire les besoins des chercheurs, dans le respect des droits associés aux corpus analysés. L'annotation de vidéos, en tant que pratique de recherche, reste une opération complexe, même pour les chercheurs technophiles. Cela est d'autant plus étonnant que cette pratique concerne des pans entiers des SHS. On pensera prioritairement aux études cinématographiques, pour lesquelles l'image est au centre de l'attention. Dans ce champ, les pratiques pédagogiques les plus courantes impliquent la visualisation, l'analyse, la comparaison et le découpage de séquences (Bordwell et Thompson, 2014).

En sciences de l'information et de la communication, en anthropologie ou en sociologie, il n'est pas rare de réaliser ou d'analyser des vidéos, des entretiens et des situations d'enquête filmés qu'il convient de sémantiser. L'observation sociale reposant sur la captation d'images à l'aide de caméras est notamment au cœur des *Workplace Studies* (Heath, Hindmarsh et Luff, 2010), un courant de recherche dont la méthode consiste à capter des activités sur une longue période pour ensuite analyser finement les vidéos et consigner les attitudes, les comportements, les interactions. Comme cela a été évoqué en introduction, cette pratique est également répandue en histoire. La géographie, les sciences de l'éducation et les sciences et techniques des

activités physiques et sportives (STAPS) ne sont pas en reste: on y étudie des films d'urbanisme ou météorologiques (comme le fait CNRS Images en diffusant des vidéos sur les problèmes de l'urbanisme, de l'habitat et de la société: http://www.cnrs.fr/fr/multimedia/podcast-urbanisme-habitat-societe/description.html), des documents à caractère pédagogique (par le réseau Canopé: https://www.reseau-canope.fr/bsd/) ou des séquences permettant d'observer les mouvements du corps (à l'UFR STAPS de l'Université de Lorraine grâce à sa plate-forme technologique).

Dans certains de ces travaux, l'étude de l'image elle-même constitue une finalité. Pour d'autres, elle n'est qu'un matériau, soit principal, soit complémentaire d'autres sources (textuelles, orales). Ce matériau peut être ou non produit par le chercheur. Mais dans tous les cas, ces recherches nécessitent un outillage technologique. L'historien Rémy Besson (2014) en a fait l'expérience en travaillant sur un corpus d'entretiens filmés qu'il souhaitait annoter pour expliciter les propos tenus par les protagonistes. Pour cela, il a dû trouver des solutions pour répondre à ses besoins en encapsulant des vidéos *streamées* (diffusées en direct depuis un site distant) depuis la plate-forme Vimeo vers un site dédié, puis en insérant des contenus périphériques: textes de présentation, mots-clés, etc. (http://archivesnumeriques.org/).

Au problème de l'analyse elle-même s'ajoute celui de la médiatisation des recherches en ligne (contenu vidéo et annotations périphériques). Sur ce point, on évoquera le cas de la chercheuse Hélène Fleckinger (2011) qui a dû faire face à des besoins de présentation de son corpus et des annotations associées. Elle a d'abord fait un travail de repérage des acteurs des mouvements féministes dans un corpus de films documentaires. Pour cela, elle a utilisé un logiciel pour insérer des marqueurs sur le film chaque fois qu'elle repérait, seule ou avec l'aide d'une protagoniste, d'autres actrices-clés de ces mouvements. Mais rendre

compte de ce travail sur Internet n'a pas été possible autrement que par des transcriptions textuelles au sein d'un projet baptisé «Bobines féministes» (http://cinevideo.labex-arts-h2h.fr/content/bobines-f%C3%A9ministes).

Des pratiques généralisées de bricolage

Depuis 2010, dans le cadre du programme de recherche Cinecast, nous avons conduit sous forme d'entretiens semi-directifs une ethnographie des chercheurs en SHS travaillant sur des matériaux vidéo (films, séries, documentaires, journaux télévisés), dans un contexte d'avènement des technologies centrées sur l'utilisateur. Comme cela apparaît dans les exemples précédemment évoqués, cette enquête nous a permis de découvrir que les pratiques relèvent le plus souvent d'un «bricolage» nécessitant de faire avec «les moyens du bord» (Lévi-Strauss, 1990 [1962]). Les chercheurs que nous avons observés utilisent des crayons et du papier ou des logiciels ordinaires de traitement de texte. Aucun logiciel dédié à l'analyse de films ne s'est imposé au sein de la communauté scientifique, malgré quelques tentatives de développement technologique comme Ant (http://ant.umn.edu/), Anvil (http://www.anvil-software.org/) ou Pad.ma (http://pad.ma/).

Ce travail d'analyse manuel et linéaire se fait généralement en plusieurs étapes. Les chercheurs s'installent devant leur écran (télévision ou ordinateur) et ils prennent des notes factuelles consistant à relever des éléments-clés de la structure narrative: personnages, décors ou effets de montage. Dans un second temps, ils reviennent sur le contenu pour affiner leur analyse, confirmer des intuitions, consigner de nouvelles données. La prise de notes est donc nécessairement linéaire, ce qui ne facilite pas la circulation dans le matériau filmique. Autant cette manière de procéder est aisée lorsqu'on manipule un ouvrage papier, autant elle devient fastidieuse et chronophage lorsqu'on explore un contenu audiovisuel. Par

ailleurs, il est extrêmement difficile de mener des analyses comparées, à moins d'employer des stratagèmes, à l'exemple de cette chercheuse qui crée des lignes d'analyse temporelle sur des feuilles de papier qu'elle assemble ensuite avec de l'adhésif, à la manière d'un rouleau de papyrus.

Certains chercheurs, qui travaillent sur des sources DVD/Blu-ray ou des fichiers numériques (type MP4), trouvent des solutions avec les technologies qui sont à leur disposition. Ils utilisent notamment les signets des lecteurs vidéo QuickTime ou VLC pour faciliter leur circulation dans le contenu vidéo (l'objectif étant de marquer les moments-clés du film qu'ils analyseront ensuite en profondeur). D'autres réalisent des captures d'écran pour illustrer ou argumenter un propos, notamment en vue d'une publication. Mais ces actions restent sommaires et beaucoup de chercheurs peinent même à trouver des astuces pour télécharger des ressources accessibles en ligne ou convertir des fichiers vidéo qui pourront ensuite être manipulés plus aisément.

Le détournement des outils professionnels de montage

Pour les chercheurs les plus technophiles, il est cependant inconcevable de rester dans cet état de dénuement technologique. C'est pourquoi certains d'entre eux se sont emparés des logiciels traditionnels de montage: ceux qui sont mis à disposition des usagers à l'achat d'un ordinateur, comme Movie Maker ou iMovie; mais surtout des logiciels payants, initialement destinés à un public de professionnels comme Adobe Premiere ou Final Cut Pro. Ils font alors un usage stratégique et tactique de ces outils, selon Michel de Certeau (2010 [1990], p. 57-63), au sens où ils n'utilisent qu'une partie des fonctionnalités dédiée à l'annotation.

En effet, ces logiciels n'ont pas été conçus pour la recherche et ils ne répondent que partiellement aux besoins

des analystes de l'image. Les fonctionnalités d'annotation occupent une place secondaire dans leur ergonomie générale, car elles n'ont qu'une place de pense-bête en vue du montage d'un film. S'ajoute à cela le lourd investissement financier que leur usage suppose (achat d'un ordinateur suffisamment puissant équipé du logiciel), ainsi que la complexité d'usage générée par le grand nombre de fonctionnalités inutiles et donc parasites pour le chercheur. Par ailleurs, la pratique du montage (qui peut s'avérer utile dans certains cas de figure) demande des compétences poussées que seuls les chercheurs en études cinématographiques peuvent se permettre de développer.

On le voit, pour la plupart des scientifiques appartenant aux disciplines issues des HN, l'usage d'un logiciel adapté à la recherche s'impose. Dès 2010, l'idée d'un logiciel d'annotation vidéo qui leur est dédié mobilise des équipes de recherche (LIRIS, INA, IRI). Un tel outil doit notamment ouvrir de nouvelles perspectives en favorisant la systématisation et le partage des résultats. Il doit aussi donner la possibilité de présenter une recherche, un corpus et des commentaires associés de manière formalisée lors d'une conférence ou d'un séminaire.

Les premiers outils d'annotation vidéo et leur positionnement

Dans un contexte de poussée de la vidéo et des technologies numériques centrées sur l'usager, on a vu apparaître, en France, trois outils: Advene, Mediascope et Lignes de temps. Advene, développé par le laboratoire LIRIS, est le résultat d'un projet de recherche technologique très novateur intégrant de nombreuses fonctionnalités d'annotation et de montage. Cependant, le développement de cette technologie s'inscrit dans une démarche expérimentale qui n'a pas donné lieu à une large diffusion dans le monde de la recherche. On peut malgré tout mentionner une application ayant permis de montrer le

potentiel de cet outil: un projet d'analyse de parcours de visites filmés dans des musées.

Mediascope (http://www.inatheque.fr/consultation/mediascope.html) est un outil d'aide à l'analyse de programmes audiovisuels implémenté à l'INAthèque, la bibliothèque de recherche de l'INA. Librement téléchargeable jusqu'à une période récente, Mediascope reste utilisé par des chercheurs qui se rendent directement dans ce fonds documentaire pour travailler sur des corpus télévisuels. L'outil est surtout utilisé pour ses fonctionnalités de lecture (pause, stop, avancer, reculer), le découpage de segments (isoler différents sujets dans une émission, chapitrer une séquence longue), la ponction d'images (sous la forme de captures d'écran) et la constitution de corpus personnalisés. En revanche, il est peu mobilisé pour la prise de notes. La plupart du temps, les analyses de séquences continuent à être menées sur papier ou à l'aide d'un deuxième écran et d'un logiciel de traitement de texte.

Lignes de temps est le premier outil d'annotation vidéo dont on peut dire qu'une communauté de chercheurs l'a utilisé de manière systématique. On l'a imaginé en collaboration avec des critiques de cinéma. Autant dans son fonctionnement que dans son ergonomie, il s'inspire donc des interfaces de montage filmique. Il est centré sur le caractère temporel du film et fonctionne selon un principe de fabrication de lignes d'analyse (Puig et Sirven, 2007; Bourgatte, 2012). À l'ouverture d'un projet, Lignes de temps génère un plan par plan: une ligne découpée en autant de plans que le film en contient. Ce découpage permet d'en visualiser la quantité et leur durée. Il facilite surtout la navigation dans le contenu filmique et la conduite d'analyses.

En ce sens, on peut dire que Lignes de temps favorise l'exploration libre et personnalisée des contenus vidéo. Il permet de marquer des séquences en les numérotant, en

les colorisant ou en les taguant (insertion de mots-clés). Il est possible de prendre des notes; de démultiplier les lignes d'analyse; de faire une étude comparée entre plusieurs ressources vidéo; de segmenter ses ressources et même de mettre des scènes en regard grâce à une fonctionnalité dite de bout à bout. Ces fonctionnalités d'ajout de descripteurs, de manipulation et de comparaison des contenus permettent de repérer des procédés techniques récurrents, des personnages, etc. Elles agissent incontestablement comme un facilitateur pour le chercheur. Le logiciel est utilisé à des fins pédagogiques avec des publics scolaires, y compris dans des petites classes (CP). Dans le cadre d'ateliers d'éducation à l'image, il permet de revenir sur les films vus en salle, de les explorer et d'approfondir ainsi une culture audiovisuelle.

Annoter et collaborer autour des images: le projet Celluloid

Que ce soit pour des problèmes de financement, de moyens humains, ou d'autres motifs, aucun des outils cités précédemment n'a connu de développement lui permettant d'être largement adopté au sein de la communauté scientifique. Ces expériences ont cependant permis de déclencher une dynamique de recherche et ont mis au jour le besoin d'un outil. À partir de ce constat, la technologie Celluloid, dont le développement a commencé en 2015, s'inscrit dans cette dynamique. Elle s'appuie sur l'assemblage de briques technologiques *Open Source*, des bibliothèques *Javascript*, et sur une philosophie importée de la sphère de l'annotation textuelle adaptée à la vidéo. Le module d'annotation Annotator (http://Annotatorjs.org), développé par l'Open Knowledge Foundation Network, utilise le format «Open Annotation Data Model» (http://www.openannotation.org/spec/core/). Nous avons associé ce module au lecteur vidéo Video.js (http://videojs.com), développé en HTML5.

Le défi est de réaliser un outil accessible et simple d'utilisation. Le premier objectif est de pouvoir naviguer dans une vidéo, sélectionner et marquer des points ou des segments en y associant du texte, des images ou d'autres vidéos. Le deuxième objectif est de permettre la collaboration de plusieurs acteurs, pouvant avoir des statuts différents (on pense notamment au cas du professeur qui collabore avec ses étudiants dans le cadre d'un séminaire et qui aura alors une position d'autorité). Le troisième objectif est de s'assurer que les droits relatifs à l'exploitation des contenus vidéo sont respectés, ce qui nécessite de sécuriser l'accès à la plate-forme par un dispositif d'authentification.

Il importe également de prendre en compte la pérennité des technologies utilisées, afin que ces outils constituent un véritable gain pour les chercheurs, les enseignants et les étudiants. À cette fin, il est indispensable de composer avec l'offre du marché numérique, en faisant attention aux spécificités des systèmes d'exploitation (Windows, OS X, iOS, Linux, Android, etc.) et des navigateurs (Internet Explorer, Firefox, Chrome, Safari, etc.) qui ont chacun leurs propres méthodes de décodage des vidéos. Il faut, en outre, anticiper les évolutions technologiques. Il s'agit de concevoir une interface dite adaptative – ou *responsive* – qui fonctionne avec tous les types d'écrans (ordinateurs, tablettes, smartphones). L'ensemble de ces contraintes techniques a un effet direct sur les pratiques: on sait qu'un outil qui ne lève pas, d'emblée, ces verrous ne sera pas utilisé par la communauté des chercheurs.

Une fois l'inventaire de ces contraintes techniques réalisé, il faut souligner que le projet Celluloid repose sur un parti pris de fond qui consiste à désacraliser le film, en prenant au sérieux la mutation des discours vidéographiques. L'accès aux contenus audiovisuels s'est massifié; la production et la diffusion d'images ne sont plus l'apanage des seuls professionnels. Tout le monde peut désormais regarder, capter et diffuser des images sur la plate-forme

YouTube, ainsi qu'avec des applications comme Vine ou Periscope. Sur le modèle des pratiques d'annotation textuelle, qui se sont démocratisées avec la naissance du livre de poche, Celluloid promeut une pratique d'annotation réalisée directement sur la matière filmique. Techniquement, on insère donc des marques d'annotation sur une trame transparente qui est elle-même posée sur l'image filmique. Au final, on obtient une constellation de marques sur le film que l'on peut faire apparaître ou masquer à sa guise.

Pour travailler avec Celluloid, le chercheur ou le groupe de recherche peut constituer son corpus en mobilisant plusieurs types de sources: des vidéos libres de droits qu'il télécharge sur la plate-forme ou qu'il *streame*; des vidéos produites directement par le ou les chercheur(s); des vidéos sous droits pour peu qu'il s'en acquitte. Ce problème fondamental de la gestion sécurisée de films soumis au droit d'auteur a d'ailleurs fait l'objet d'un travail exploratoire mené avec le consortium UniversCiné (http://www.universcine.com). Sur ces vidéos, l'outil d'annotation permet non seulement d'intégrer du texte avec mise en forme (fonctions basiques de traitement de texte), mais aussi des images ou d'autres vidéos, de manière à créer des liens entre différents contenus (montrer qu'un plan est inspiré d'un tableau; intégrer un commentaire filmé) selon un principe multimédiatique ou d'hypervidéo.

Celluloid est un outil qu'on a conçu pour la production d'annotations collaboratives. Il ne s'agit donc pas seulement de donner les moyens à un chercheur d'annoter une ressource, mais bien de penser les modalités de partage, d'échange et de recherche collective. Celluloid se présente ainsi comme un module intégré à un espace numérique de travail (ENT), au sein de l'environnement Moodle, ce qui offre la possibilité de créer des groupes de travail (un groupe de chercheurs; une classe et un professeur) autour d'une ressource vidéo. Les participants peuvent également

échanger à distance ou en présentiel et tirer parti de services périphériques intégrés à l'ENT, comme le wiki, le forum ou la carte mentale qui vont leur permettre d'organiser leurs recherches ou de planifier des tâches. Celluloid pourrait, par exemple, faire l'objet d'une utilisation pour la mise en route de colloques ou de séminaires «inversés» en suivant le modèle pédagogique de la classe inversée (principe consistant pour le professeur à fournir un contenu théorique en amont du cours afin que l'apprenant le consulte chez lui, pour qu'en classe, professeur et élèves consacrent plus de temps à des applications pratiques).

Par exemple, nous avons utilisé Celluloid pour animer un atelier dans le cadre du THATCamp Paris 2015 (http://tcp.hypotheses.org/839). À partir d'un *mashup* préparé en amont, les participants étaient invités à cerner les sources des différentes vidéos le composant et à les partager via l'outil d'annotation. Ils pouvaient en outre commenter, enrichir, voire contredire les trouvailles des autres participants dans une logique collaborative.

En dépit de quelques initiatives ouvertes émanant du monde universitaire (comme https://vialogues.com), beaucoup d'entre elles sont développées par des acteurs privés dans une logique propriétaire. C'est un aspect de l'explosion des technologies pour la recherche et des EdTechs (dénomination qui sert à désigner l'ensemble des technologies pour l'enseignement) qu'on ne doit pas sous-estimer si l'on ne veut pas freiner l'émergence de solutions ouvertes, développées par et pour les chercheurs. Dans cet esprit, on peut par exemple citer les outils développés par la Software Studies Initiative, dirigée par Lev Matovich comme ImagePlot (http://lab.softwarestudies.com/p/imageplot.html), qui permet d'explorer de grands corpus d'images afin de repérer des motifs récurrents, ou encore ceux du Center for History and New Media de la George Mason University comme Omeka.

Les mutations liées à l’explosion de la vidéo numérique doivent aussi être appréhendées dans l’environnement plus large de l’enseignement supérieur. De récentes études montrent que 80% des étudiants utilisent aujourd’hui des vidéos en ligne pour préparer leurs cours, approfondir leurs enseignements et préparer leurs examens. Il s’agit de vidéos réalisées par les enseignants ou trouvées sur Internet (le plus généralement sur YouTube ou sur des plates-formes d’enseignement en ligne comme la Khan Academy). Selon les mêmes sources, 70% de ces étudiants disent consulter de plus en plus régulièrement ces vidéos directement pendant leurs cours. Pour les chercheurs, le développement de ces usages pédagogiques remet en question les pratiques de recherche d’au moins deux manières. D’une part, ces vidéos pédagogiques constituent un terrain d’enquête, notamment en sciences de l’éducation, où les usagers peuvent avoir recours aux outils d’annotation pour traiter ces corpus particuliers. D’autre part, et au-delà même des questions d’annotation, ces nouveaux usages posent la question de la *production* d’images de recherche, un champ qui reste encore largement à explorer.

CHAPITRE 2

L’usage des réseaux sociaux pour chercheurs

EMMANUEL MOURLON-DRUOL

Les réseaux sociaux s’imposent progressivement comme une composante à part entière du profil d’un universitaire. L’image du chercheur ermite, rétif aux nouvelles technologies et volontairement détaché de tout lien électronique est en train progressivement de s’effacer. Deux influences y concourent: d’une part, le développement de réseaux sociaux spécifiquement dédiés aux chercheurs universitaires, et d’autre part, une demande grandissante de la présence d’universitaires en dehors de leur milieu professionnel.

L’antichambre des réseaux sociaux pour chercheurs: les listes de diffusion

Sans être un réseau social à proprement parler, les listes de diffusion par courriel demeurent, à l’heure où sont écrites ces lignes, l’un des moyens les plus aisés de partager et collecter des informations scientifiques. Les procédures d’abonnement et de désabonnement à ces listes de diffusion électronique sont généralement simples, immédiates, et flexibles. Les listes de diffusion sont multiples, et reflètent l’existence de diverses associations de chercheurs qui s’organisent traditionnellement par disciplines. Dans mon champ de recherche, l’histoire (économique) des relations internationales, plusieurs listes me sont utiles, parmi lesquelles (dans l’ordre alphabétique) l’Association française d’histoire économique (AFHÉ: https://afhe.hypotheses.org), EH.Net, H-Net (dont fait notamment partie H-Diplo: http://h-diplo.org), H-Soz-u-Kult

(Humanities – Sozial und Kulturgeschichte), et RICHIE (Réseau international des chercheurs en histoire de l'intégration européenne: https://www.europe-richie.org).

L'avantage d'un abonnement à de telles listes de diffusion est évident: il permet de recevoir toutes les informations relatives à un champ de recherche donné, allant d'un appel à contributions à un programme de colloque en passant par une offre d'emploi. Il permet également de diffuser à un public très large ses propres informations que l'on souhaiterait partager. L'inconvénient majeur est la contrepartie de cet avantage: la circulation de l'information étant large, l'abonné(e) peut vite se retrouver submergé(e) par une vague d'informations gigantesque. L'immense majorité des informations partagées n'intéressera pas directement l'utilisateur, le tri sera long (classement et conservation des courriels), et, au moyen du «répondre à tous», on peut se retrouver spectateur d'un débat interne à une discipline auquel on ne trouve pas nécessairement d'intérêt. Mais les quelques informations qui intéresseront bel et bien l'utilisateur peuvent justifier à elles seules l'abonnement. L'abonné(e) devra donc penser à affiner les réglages de sa souscription: fréquence des courriels reçus (un par annonce? quotidien? hebdomadaire?), consultation dès la réception ou de façon regroupée pour éviter des dérangements répétés, etc.

Les réseaux sociaux spécialisés dans la recherche: Academia.edu et ResearchGate

Des réseaux sociaux exclusivement dédiés à la recherche se sont récemment développés. Academia.edu, lancé en 2008, est un réseau social pour universitaires qui leur permet de partager leurs travaux, d'en mesurer les répercussions, et de suivre les recherches des autres utilisateurs selon leurs intérêts personnels. J'y ai ouvert un compte en 2008 et je l'ai maintenu depuis. Academia.edu a plusieurs mérites: le site est plutôt facile d'utilisation, son indexation sur Google

est excellente, et il permet effectivement de développer des liens avec d'autres chercheurs et de faire connaître sa recherche au-delà de ses cercles professionnels habituels. Comme tout réseau social, il offre indéniablement un certain nombre d'opportunités qui n'auraient pas été possibles il y a une dizaine d'années. Mais Academia.edu est également irritant: son indexation sur Google conduit à ce qu'une entreprise privée à but lucratif obtienne une visibilité supérieure à celle des universités (un profil Academia.edu arrive par exemple souvent en tête d'une recherche Google avant même celui de l'université de rattachement du chercheur); et si ce réseau social permet d'observer qui a pu trouver l'un de vos travaux, où, et quand, la compilation de ces statistiques reste obscure. En outre, depuis quelques mois, Academia.edu propose des solutions de monétisation de son service auprès des chercheurs, qui ont abouti à de nombreux articles encourageant les utilisateurs à effacer leurs comptes (Bond, 2017).

ResearchGate fut également fondé en 2008. J'y ai personnellement ouvert un compte beaucoup plus tard que sur Academia.edu, en 2013 seulement. Ceci tient au fait que jusqu'à récemment, ResearchGate avait une image et un réseau essentiellement liés aux sciences «dures», puisque le site avait été fondé par le virologue Ijad Madisch. Cette image a progressivement évolué, j'ai appris à connaître la plate-forme et j'ai vu un certain nombre de mes collègues s'y inscrire, ce qui m'a conduit à leur emboîter le pas. ResearchGate fonctionne sur le même principe que Academia.edu: l'utilisateur se crée gratuitement un profil, détaille ses intérêts de recherche, liste ses publications, et peut suivre d'autres chercheurs.

ResearchGate se distingue d'Academia.edu essentiellement par deux aspects. D'abord, le site offre la possibilité de poser des questions aux autres utilisateurs à propos d'une recherche – libre à chacun, bien sûr, de répondre ou non à ces sollicitations. Même si je ne l'ai

personnellement jamais expérimentée, cette fonctionnalité semble bel et bien utilisée par nombre d'autres abonnés au site. Ensuite, ResearchGate a établi un indicateur – le RG score – décrit par le site comme «une nouvelle manière de mesurer votre réputation scientifique». Là où Academia.edu offre des statistiques brutes de consultation de pages et de liste de mots-clés utilisés, ResearchGate apporte un chiffre qui permet de comparer les chercheurs entre eux.

ResearchGate, comme Academia.edu, crée donc autant d'opportunités que d'irritation. Le site permet indéniablement de tisser des liens et de faire des découvertes qui n'auraient pas été possibles sans ce réseau social; et il contribue à une meilleure visibilité de sa propre recherche. Mais, comme Academia.edu, il tombe dans un certain nombre de travers critiquables. Le calcul du fameux RG score est assez mystérieux. Si le site explique bien les facteurs qui président à son calcul – publications, questions, réponses, *followers* – la production tout comme la pondération de ceux-ci restent opaques. Or, si ces modes de calcul sont inconnus, il est très clair qu'un article dans certains journaux vaudra toujours plus qu'un livre, ce qui est un biais de calcul considérable, surtout au vu de la diversité des disciplines universitaires utilisant le site. Comme Academia.edu, ResearchGate pousse à la mise en ligne du texte intégral de votre recherche, ce qui produit le même effet: offrir à une entreprise privée à but lucratif une recherche souvent produite sur fonds publics. À l'heure des compteurs sur les sites des maisons d'édition – notamment Taylor & Francis – une telle politique revient également à altérer le calcul: certaines publications seront vues sur le site de ResearchGate ou d'Academia.edu, et non pas sur le site de la maison d'édition d'origine, dont le compteur de vues ne sera donc plus un reflet statistique fidèle. L'universitaire doit donc réfléchir à deux fois avant d'offrir son texte intégral à Academia.edu ou ResearchGate, suivant l'usage (et la cohérence) qu'il souhaite obtenir des

statistiques produites par ces différents sites. Personnellement, je ne mets pas le texte intégral de mes travaux sur ces réseaux sociaux s'il existe un compteur sur le site de la maison d'édition, afin de pouvoir maintenir ce compteur aussi fidèle que possible.

On peut également noter la création plus récente, en 2011, du réseau social Piirus, géré par l'Université de Warwick au Royaume-Uni (www.piirus.com). Piirus vise à permettre aux chercheurs de trouver plus facilement d'autres chercheurs aux intérêts similaires – le réseau Piirus a d'ailleurs débuté sous le nom de «Research Match». L'utilisateur se crée un profil gratuitement, détaillant ses intérêts de recherche et ses méthodologies. Une fois qu'il a entré ses informations, le site suggère des profils d'autres chercheurs aux intérêts connexes avec lesquels il pourrait vouloir entrer en contact ou collaborer. Je n'ai personnellement découvert ce réseau que très récemment et n'y ai pas (encore) ouvert de compte. Il est à noter que ce réseau, à l'inverse de ResearchGate et Academia.edu, est soutenu directement par une université.

Alors que l'*Open Access* est au cœur des débats entre universitaires, États et maisons d'édition, la question de la disponibilité du texte intégral des recherches devient en effet cruciale. À partir du 1er avril 2016, au Royaume-Uni, toute recherche devra avoir été déposée sur un site institutionnel une fois sa publication acceptée par une revue, à défaut de quoi le texte en question ne sera pas éligible au prochain cycle d'évaluation de la recherche des universités britanniques – le Research Excellence Framework (REF) – prévu pour 2020. Chaque université britannique possède son propre dépôt institutionnel – celui de l'Université de Glasgow s'appelle Enlighten – et chaque chercheur est tenu d'y référencer ses travaux (http://eprints.gla.ac.uk et http://eprints.gla.ac.uk/view/author/16738.html pour ma page personnelle). En France, la loi pour une République

numérique d'octobre 2016 permet aux auteurs de mettre à disposition en libre accès leurs articles, s'ils sont issus d'une recherche financée en tout ou partie sur fonds publics. La plate-forme HAL (https://hal.archives-ouvertes.fr/) permet déjà à tout chercheur d'y déposer ses productions scientifiques et de disposer de certaines fonctionnalités que l'on retrouve sur les réseaux sociaux pour chercheurs, comme des statistiques ou un CV en ligne. En outre, un *widget* pour le CMS WordPress, disponible pour la plate-forme Hypothèses, permet d'intégrer certaines fonctionnalités de HAL dans son carnet de recherche.

La multiplication de ces réseaux, dépôts institutionnels et autres sites Internet dédiés aux universitaires a encouragé la création d'un identifiant unique des chercheurs – l'Open Researcher and Contributor ID, ou ORCID – depuis 2012. ORCID (http://orcid.org) est un code alphanumérique non propriétaire permettant d'identifier les chercheurs de façon unique et permanente. C'est une entreprise à but non lucratif soutenue par des membres institutionnels comme des organismes de recherche et des maisons d'édition. D'une certaine façon, l'ORCID s'apparente au DOI (*Digital Object Identifier*) en ce qu'il cherche à permettre un repérage unique et pérenne de contenus (un article, un profil de chercheur) alors même que les adresses URL peuvent changer, ou que des homonymies entre chercheurs peuvent créer la confusion. Un nombre grandissant d'organismes de financement, comme le Wellcome Trust et l'European Research Council (ERC), encouragent les candidats à créer un compte ORCID. La création de ce compte est très simple, rapide, gratuite, peut s'effectuer en plusieurs langues, et résout effectivement nombre de problèmes de cohérence car il est international.

Les réseaux sociaux généralistes et les chercheurs: Facebook, Twitter, LinkedIn

Les réseaux sociaux (Facebook, Twitter, LinkedIn…) ne sont toutefois pas l'apanage des chercheurs. C'est d'ailleurs plutôt l'inverse, les deux réseaux sociaux les plus emblématiques – Facebook et Twitter – étant antérieurs à Academia.edu et ResearchGate. Les réseaux sociaux généralistes se sont progressivement introduits dans la recherche universitaire car ils servent de plus en plus comme plates-formes de partage et de collecte d'informations pour les chercheurs. Facebook, créé en 2004 puis accessible à tous en 2006, est sans doute devenu le plus célèbre au monde. Pour le chercheur, toutefois, il crée la suspicion, étant perçu comme frivole et prompt à susciter la controverse. Mais avec plus d'un milliard d'utilisateurs depuis 2015, Facebook joue un rôle dans la diffusion des travaux des chercheurs, et ceux-ci peuvent difficilement faire l'économie d'une réflexion sur le rôle de Facebook à cet égard. Certes, ce n'est pas un réseau pour universitaires comme ceux que l'on vient d'évoquer. Mais c'est peut-être là que réside son intérêt: il permet de faire connaître ses travaux scientifiques à des utilisateurs qui ne seront pas inscrits (et ne s'inscriront jamais) sur Academia.edu ou ResearchGate.

Facebook offre en effet la possibilité de scinder son profil d'utilisateur en deux: l'un privé, regroupant ses «amis» (l'accès à ce profil pouvant être lui-même encore limité), où les informations postées ne seront accessibles – selon les réglages de l'utilisateur – qu'à ce même cercle restreint; et l'autre public, ou toute publication sera accessible librement, y compris à des personnes avec lesquelles l'utilisateur n'est pas «ami», ou encore via une simple recherche Google. Un chercheur peut donc poster en mode public un article qu'il vient de publier, et ainsi atteindre un lectorat potentiel beaucoup plus vaste que son simple réseau d'amis. Un utilisateur lambda peut lui aussi décider de suivre la page publique d'un chercheur, voire d'interagir avec lui. De plus, Facebook permet la création de groupes

selon un centre d'intérêt donné. Le groupe La fabrique de l'histoire, consacré à l'émission d'Emmanuel Laurentin sur France Culture, compte plus de 15 000 abonnés; le groupe Les coulisses de Bruxelles, lié au blog du correspondant à Bruxelles de *Libération* Jean Quatremer, compte près de 4 000 abonnés. Suivant les règles de partage de chacun de ces groupes, on pourra donc imaginer qu'un chercheur pourrait y partager l'annonce d'une publication et ainsi atteindre potentiellement plusieurs milliers d'autres utilisateurs. Dès que possible, je cherche à partager mes travaux sur Facebook, que ce soit sur ma page publique personnelle ou via les groupes Facebook pertinents. Ceci a indéniablement offert une plus grande visibilité et un plus large lectorat à certains de mes billets, en particulier ceux liés à l'actualité: la disparition des archives de Claude Guéant, la publication des premières minutes de la Banque centrale européenne, l'union bancaire, l'affaire de la publication des courriels d'Hillary Clinton (www.e-mourlon-druol.com/). Cette meilleure visibilité m'a permis d'être contacté par des journalistes et des chercheurs, et plus généralement de montrer, en libre accès, mes travaux et mon profil de recherche.

Twitter ne souffre pas du même a priori quant à la réputation que Facebook, mais plutôt d'un auditoire plus limité, et d'un modèle économique moins lucratif. Né en 2006, Twitter a environ moitié moins d'utilisateurs que Facebook, et surtout sans doute encore moins d'utilisateurs actifs. En effet, de nombreux comptes sont créés soit de façon robotique pour artificiellement augmenter le nombre de *followers*, soit en tant que comptes institutionnels et ont été abandonnés par leur créateur sans pour autant avoir été supprimés. Twitter est un réseau dit de «microblogging»: l'utilisateur *tweete* un message de 140 caractères maximum. L'utilisateur suit d'autres comptes, et d'autres comptes peuvent suivre l'utilisateur: aucune réciprocité dans le suivi n'est obligatoire. Les utilisateurs de Twitter

peuvent répondre à un *tweet*, le citer, ou le *retweeter*. Twitter offre donc des opportunités similaires à celles de Facebook tout en ayant l'image (et la réalité) d'un réseau social plus sérieux et professionnel, en grande partie parce que la longueur des messages postés est fortement limitée. Un message succinct n'empêche certes pas la frivolité du commentaire, mais si celui-ci se veut réellement pertinent, la brièveté du message dévoilera les talents de concision et d'esprit de son auteur. J'utilise Twitter essentiellement pour partager et récolter des informations ayant trait à ma recherche, annoncer et promouvoir la publication d'un nouvel article ou billet de blog, et échanger avec d'autres utilisateurs. Comme tout réseau social ouvert, des règles de prudence élémentaires doivent être respectées, surtout sur Twitter, qui par défaut est public. Attention donc à ce que l'on publie… Tout message posté sur Twitter (sauf si on a un compte volontairement verrouillé, ce qui n'a pas vraiment d'intérêt) devient visible. Il peut être supprimé par la suite, mais suppression n'équivaut pas à disparition, puisque rien n'empêche un autre utilisateur de l'immortaliser par une capture d'écran, ou d'utiliser un dispositif technique plus complexe pour collecter des *tweets*.

Certains utilisateurs ne prêtent toutefois guère attention à ce qu'ils écrivent: Twitter et Facebook ont vu apparaître un type d'utilisateur bien spécifique, prompt à nourrir la controverse, baptisé familièrement le «troll». Le troll est un utilisateur dont le but – souvent le seul et unique – est de perturber un débat, générer une polémique, engendrer le conflit, en cherchant constamment à empêcher un réel échange. Le troll provoque intentionnellement et ne cherche qu'à nuire, plutôt qu'à débattre sereinement. Un chercheur utilisant Facebook ou Twitter ne fera pas nécessairement face à ce type de situation – le cas échéant, la meilleure réponse est le silence – mais au moment de réfléchir à son implication dans les réseaux sociaux, le chercheur doit clairement réfléchir au *trolling* dont son sujet de recherche

pourrait faire l'objet. Un sociologue travaillant sur les camps de migrants, ou sur les phénomènes de radicalisation religieuse, s'exposera très probablement à un *trolling*. Cela doit-il pour autant le décourager d'utiliser Facebook ou Twitter? Il lui appartiendra de peser le pour et le contre en fonction de son champ de recherche, de ses attentes par rapport à l'utilisation des réseaux sociaux, et de sa propre capacité à gérer l'activité de ces trolls. Il faut surtout savoir que deux ripostes simples existent: en plus de garder le silence, on peut bloquer les trolls ou les signaler à Facebook/Twitter (si leurs propos incitent à la haine, par exemple).

Les réseaux sociaux de photos, comme Flickr et Pinterest ne sont pas utiles dans mes recherches, en tout cas pour le moment, mais ils ne doivent pas être délaissés pour autant. D'abord, ces sites constituent des bases de données et peuvent servir de source d'analyse pour de futures recherches (qui s'en sert? Comment et à quelles fins?). Et puis, certains comptes ou groupes peuvent être constitués dans une perspective de collection de sources, comme The Great War Archive Flickr Group (https://www.flickr.com/groups/greatwararchive/). Ce groupe est la suite du projet The Great War Archive, géré par l'Université d'Oxford pendant quelques mois en 2008, qui a dû s'interrompre faute de financement (www.oucs.ox.ac.uk/ww1lit/gwa/). L'Université a souhaité continuer d'offrir la possibilité de recueillir des documents sur une autre plate-forme, Flickr.

LinkedIn, créé en 2003 et ayant aujourd'hui environ moitié moins d'utilisateurs que Facebook, est un réseau social purement professionnel. L'utilisateur s'y crée gratuitement un profil qui ressemble à un CV, où l'on détaille sa formation, ses expériences, et son parcours. LinkedIn devient ainsi un carnet d'adresses en ligne, qui remplace la pile de cartes de visite collectées au gré des divers colloques et séminaires. LinkedIn reste certes dépendant de

la mise à jour régulière de leurs informations par ses utilisateurs, mais il sera de toute façon plus à jour que la vieille carte de visite reçue il y a dix ans. L'utilisateur peut poster des informations strictement professionnelles qui seront diffusées dans un fil d'information – comme sur Facebook – à tous ses contacts. LinkedIn permet de maintenir un contact avec diverses personnes, et de se maintenir mutuellement informé des évolutions de carrière.

Les blogs de chercheurs

Le blog, ou carnet, permet à un auteur – le blogueur – de publier un texte, généralement succinct, sur un site Internet. Du point de vue du chercheur, le blog fut d'abord regardé avec scepticisme. Ce scepticisme venait, un peu comme pour Facebook, de la frivolité de l'exercice. Les premiers blogs parus sur Internet se distinguaient par leur caractère très égocentrique: ils étaient des lieux de partage d'expériences personnelles culinaires, touristiques, cosmétiques, jardinières ou sportives. Le *blogging* était souvent présenté comme une activité de partage d'une information dont tout le monde aurait pu (dû?) se passer. Les blogs intelligemment menés par des journalistes, des juristes ou des économistes ont toutefois permis de progressivement dissiper ces a priori. L'émergence du *blogging* scientifique a parachevé cette tendance. Il est désormais courant pour un chercheur d'écrire des billets de blog, ou bien de tenir un blog.

On peut distinguer deux grands types de blogs scientifiques: les blogs solo, et les blogs multiauteurs (je reprends ici une distinction opérée par le blog Writing For Research – https://medium.com/@write4research). Le blog solo est le carnet établi par un chercheur en son nom propre, soit sur son site Internet (ce qui est mon cas: www.e-mourlon-druol.com), soit via une plate-forme dédiée (http://hypotheses.org/ ou www.medium.com). Le blog pluriauteurs est un carnet institutionnel qui accueille des

billets invités, écrits de façon ad hoc, comme La Vie des idées (www.laviedesidees.fr – Books and Ideas dans sa version anglaise), ou les différents blogs hébergés par la London School of Economics and Political Science, comme le LSE Impact Blog ou le LSE Europp (blogs.lse.ac.uk/europpblog/). Le blog pluriauteurs permet d'atteindre un lectorat potentiel bien plus large que le blog solo – un peu comme les groupes Facebook. L'un n'empêche pas l'autre: je maintiens mon blog personnel tout en ayant déjà publié sur des blogs pluriauteurs.

Les avantages du *blogging* scientifique sont nombreux, mais peuvent se résumer en un seul: il permet d'ouvrir la boîte noire que constitue la recherche. Il offre la possibilité d'expliquer les aléas d'un travail, de partager ses doutes sur une nouvelle méthode, éventuellement de tester une idée de recherche ou de présenter des résultats intermédiaires. Le carnet permet de diffuser des informations au-delà du cercle universitaire restreint, dans un style moins corseté que celui de la revue scientifique, tout en conservant rigueur et précision. Au Royaume-Uni, le site Write4Research (https://medium.com/@write4research) explique comment transformer un article de revue scientifique en un post de blog digeste pour le commun des mortels («*How to write a blogpost from your journal article*»). J'utilise mon blog à ces fins. J'y ai partagé plusieurs réflexions liées au champ des humanités numériques: en plus des exemples (archives de Claude Guéant, publication des premières minutes de la Banque centrale européenne, et affaire des courriels d'Hillary Clinton), j'ai évoqué l'usage du numérique dans la recherche en histoire économique des relations internationales, et me suis essayé à un exercice de lecture distante en analysant le texte numérisé de l'intervention du chancelier allemand Helmut Schmidt à la Bundesbank en 1978 (www.e-mourlon-druol.com/an-exercise-in-text-mining-and-distant-reading-helmut-schmidts-visit-to-the-

bundesbank-in-1978/). Ceci m'a permis de partager certaines réflexions, d'échanger avec de nombreux chercheurs (et ainsi de faire vivre et évoluer ces réflexions), sans avoir pour autant poussé ma recherche jusqu'à la publication formelle d'un article dans une revue scientifique. Je publierai peut-être ultérieurement ces travaux, mais cela ne constitue pas une priorité pour moi actuellement, et je me suis pour l'instant contenté de regrouper tous mes billets relatifs aux humanités numériques dans une seule et unique section sur mon site. Certains billets m'ont également permis de vulgariser ou synthétiser quelques aspects précis de mes recherches, comme l'union bancaire, mais aussi les débats sur les excédents commerciaux dans la zone euro et la Capital Markets Union. À nouveau, ces billets m'ont permis d'échanger avec des journalistes ou des membres de *think tanks*, et d'attirer leur attention sur mes publications scientifiques. Finalement, mes deux billets sur l'union bancaire précités présentent certaines réflexions que j'ai publiées dans un article dans le *JCMS: Journal of Common Market Studies*. Mon blog me sert donc à tester des idées, partager des réflexions, et diffuser les résultats de ma recherche au-delà du cercle universitaire.

Mais le *blogging* scientifique n'en présente pas moins des inconvénients ou des risques. Le maintien d'un carnet ou la rédaction d'un billet est simple, s'apprend vite, et demande relativement peu de temps. Il n'empêche que même les choses les plus simples doivent s'apprendre, et donc demandent un minimum d'investissement personnel. À défaut, l'auteur risque de faire un blog décalé qui n'aura pas l'effet qu'il devrait avoir. De plus, si le blog demande relativement peu de temps, il doit tout de même être vivant, et donc se doit d'être alimenté régulièrement. Sans quoi l'auteur prend le risque que son blog/site paraisse abandonné. Le futur blogueur doit donc réfléchir à ses besoins et ses objectifs. En début de thèse, décider d'ouvrir un blog personnel et d'y publier un billet nouveau par

semaine sera sans doute trop ambitieux: la priorité du moment n'est pas au *blogging* scientifique. Publier un billet dans un carnet multiauteurs pour faire connaître son sujet de recherche sera sans doute plus adapté. Pour autant, un chercheur qui s'est adapté facilement au style du *blogging* pourrait envisager, même tôt dans sa recherche, de développer un blog. Le sociologue multipliant les enquêtes de terrain dans un camp de migrants pourrait très bien publier une fois par mois un billet transcrivant une rencontre, ou corrigeant une erreur entendue dans le débat public. Mon expérience d'écriture de billets suit un peu cette tendance: je n'ai pas tenu de blog durant ma thèse, mais j'ai commencé à écrire des billets depuis 2010, la soutenance de ma thèse coïncidant avec l'intensification de la crise de la zone euro. Par la suite, les opportunités n'ont pas manqué pour mettre en valeur l'utilité d'une perspective historique sur les débats actuels, et le blog s'est avéré un moyen privilégié pour partager mes réflexions, en complément de mes publications scientifiques.

Il n'y a pas de formule miracle pour établir quelle pratique conviendra le mieux à un chercheur. Le plus sage est d'en connaître la diversité, les atouts et les avantages, et de bien réfléchir à ses propres besoins, attentes et priorités. Les besoins sont multiples et chaque plate-forme y concourt différemment: l'usage sera-t-il passif ou actif? S'agit-il de collecter ou de diffuser de l'information? Les réseaux sociaux pour chercheurs procèdent d'une dualité qu'il convient de garder à l'esprit, entre l'émetteur et le receveur. Le chercheur en début de doctorat sera sans doute principalement receveur, et devra tâcher d'être le plus disponible possible pour recueillir des informations susceptibles de l'intéresser. Mais une fois plus avancé dans sa thèse, et surtout une fois celle-ci achevée, le chercheur deviendra autant (si ce n'est plus) émetteur d'information que receveur. À ce titre, il devra faire attention aux

différents publics susceptibles de suivre certains réseaux sociaux et pas d'autres; et de ce fait à diffuser de façon plus large ses informations afin qu'elles atteignent leur public potentiel. Il convient également de composer avec la complémentarité des réseaux sociaux. En 2016, il serait dommage de tenir un blog et de l'alimenter régulièrement tout en restant rétif à l'utilisation de Facebook, Twitter et LinkedIn. Cela reviendrait un peu à construire un magnifique navire tout en refusant obstinément de le mettre à l'eau. Le chercheur doit avant tout penser à la stratégie globale qui lui convient, à sa cohérence d'ensemble. En étant bien conscient de ses propres compétences et limites, il pourra jongler efficacement entre les différentes plates-formes, plutôt que de se perdre dans la jungle des réseaux sociaux.

CHAPITRE 3

Zotero: la gestion de références bibliographiques et de corpus documentaires

CHLOÉE FABRE

Utiliser au quotidien un logiciel de gestion de références bibliographiques permet d'optimiser son travail de recherche. Le choix de présenter plus particulièrement Zotero repose sur le fait qu'il s'agit d'un logiciel libre et gratuit. Conçu dans un centre américain de recherche en histoire, il répond clairement aux attentes de tout chercheur. Plus concrètement, nous verrons dans quelle mesure ce logiciel peut vous accompagner dans les projets de recherche.

Lire la suite sur
www.parcoursnumeriques-pum.ca/zotero-la-gestion-de-references-bibliographiques-et-de

CHAPITRE 4

De la carte mentale au carnet de recherche en ligne

PASCALE ARGOD

Une méthodologie du projet de recherche et de publication associée à des outils de l'édition ouverte revêt un grand intérêt à chaque étape du travail de chercheur. Mon expérience d'enseignante, de professionnelle de la documentation et de chercheure l'illustre bien ici, en ce qui a trait à l'investigation, l'exploration, la production et le développement de projets. La démarche heuristique aboutit à la réalisation d'un tableau de concepts et de mots clés. L'exploration utilise la carte mentale ou conceptuelle (logiciel de *mind mapping*) et des carnets de recherche. La publication d'un blog de recherche encourage le chercheur à structurer son projet sur le long terme et à répondre aux appels à projets. En effet, le rayonnement du carnet favorise la participation à des projets collectifs, mobilise des compétences et permet de créer une communauté d'intérêts. La créativité et la diffusion des recherches sont les atouts principaux de ces outils d'édition ouverte, pour cheminer de l'investigation à l'écriture.

Lire la suite sur
www.parcoursnumeriques-pum.ca/de-la-carte-mentale-au-carnet-de-recherche-en-ligne

CHAPITRE 5

La création d’une base de données théâtrales

LUISA GIARI SICH

Le travail de recherche de notre thèse sur le rôle du théâtre traduit à Venise au XVIIIe siècle a nécessité la création d’une base de données complexe, traitant les informations sur le théâtre à la fois édité et joué sur scène. Ce texte met en évidence la progression du projet, analyse la résolution des divers problèmes rencontrés et décrit l’approche scientifique utilisée pour concevoir cette base de données.

Lire la suite sur
www.parcoursnumeriques-pum.ca/la-creation-d-une-base-de-donnees-theatrales

DEUXIÈME PARTIE

L’OUTILLAGE COLLECTIF

CHAPITRE 6

Wiki, boîte à outils ou de Pandore?

MARIE CHAGNOUX ET PIERRE HUMBERT

La technologie wiki peut être perçue à la fois comme instrument, comme objet et comme outil de la recherche en sciences humaines et sociales (SHS), nous inspirant ainsi de la distinction opérée par Gilbert Simondon (2012 [1958]) entre deux de ces notions: instrument dans son appréhension des mécanismes de coopération et d'intelligence collective (Mabillot, 2012; Jacquemin, 2012), objet dans l'exploration des espaces de collaboration ainsi construits (Martine *et al.*, 2013) et outil dans les fonctions d'aide au travail de recherche (Wang, Vezenov, et Simboli, 2014; Young et Perez, 2012). Si les deux premières dimensions ont déjà fait – et font toujours – l'objet de nombreuses publications, il nous semble que la dimension du wiki comme outil reste insuffisamment explorée. On doit donc dépasser la vision classique du wiki comme formidable réceptacle d'écriture collaborative et espace d'expression de nouveaux rapports au savoir pour explorer ses usages structurant la construction de l'activité de recherche. Outil de production et outil de gestion, il permet selon nous d'appréhender deux actions essentielles au travail du chercheur: apprivoiser le temps et apprivoiser la mémoire dans une même dynamique.

Ce chapitre fera donc l'économie de thématiques pourtant fécondes pour se focaliser sur cet objectif: montrer qu'au-delà des fonctions bien définies et largement popularisées, de l'édition de contenu de type encyclopédique et du travail collaboratif, utiliser les wikis permet de disposer d'un outil

polyvalent qui répond parfaitement aux différentes facettes du métier de chercheur en SHS.

Nous ne convoquerons donc pas le mythe fondateur, l'informaticien américain Ward Cunningham qui en 1995 inventa la technologie «wiki» afin d'améliorer la collaboration au sein d'une communauté de développeurs informatiques. Nous ne ferons pas la narration de l'épopée Wikipédia dont le succès planétaire ancra l'utopie collaborative dans la réalité du Web 2.0. Nous ne ferons pas le panorama historique des grands wikis de la recherche en SHS, de glossaires partagés de sociétés savantes en wikis dédiés à un auteur ou une théorie. Nous n'offrirons pas un «wiki pour les nuls» qui permettrait une prise en main aussi rapide qu'efficiente. Nous ne développerons pas les théories de la collaboration pour vanter les mérites d'une intelligence collective que pourtant nous défendons.

En revanche, à partir de notre utilisation intensive et enthousiaste de l'outil et d'expérimentations dans des champs disciplinaires et des projets de différentes natures, nous essayerons, plus modestement mais non sans conviction, de montrer ce que cet outil peut apporter au chercheur pour véritablement faciliter son travail au quotidien. Enfin, pour compléter notre regard, forcément partial et subjectif, nous avons sollicité l'avis d'une dizaine de collègues qui, dans le cadre d'un projet commun utilisent un wiki, pour enrichir cette réflexion au prisme d'autres cultures – disciplinaires, générationnelles, philosophiques. Ce projet de *Publictionnaire* (http://publictionnaire.huma-num.fr/), qui vise la construction d'un dictionnaire critique et encyclopédique, mobilise actuellement plus d'une centaine de rédacteurs dans une perspective interdisciplinaire pleinement assumée (linguistique, sociologie, information et communication, histoire, philosophie…). La gestion et l'édition de ce dictionnaire sont assurées via un wiki, dont sont extraites certaines des illustrations que nous solliciterons au fil de la rédaction pour illustrer notre propos.

Afin de limiter les préjugés, c’est un collègue extérieur au projet, Benjamin Kelm, qui a mené les entretiens. Nous tenons à le remercier d’avoir accepté de le faire.

La technologie wiki

Projets de publication «collective», appels à projet «pluridisciplinaire», incitations au développement de «transversalités» entre pôles ou équipes de recherche au sein d’un même laboratoire… Les injonctions à collaborer ne se sont jamais faites aussi fortes dans le domaine de la recherche et impliquent l’utilisation d’outils de gestion de contenu et de méthodes de rédaction à plusieurs mains présumés opérationnels, disponibles et maîtrisés. L’un des outils qui devraient être naturellement convoqués pour ces usages est le wiki dont la grande plasticité et la relative technicité ouvrent par ailleurs des perspectives qui dépassent ses prétentions originelles.

Toute technique emporte avec elle un ensemble de potentialités et de contraintes d’usage formant des scripts ou des scénarios d’usage. Ici, le wiki apparaît initialement comme un espace vierge (page blanche) et il appartient à l’usager chercheur de construire le dispositif le mieux adapté aux problématiques qu’il traite. Cette construction s’effectue par la mise à disposition d’un catalogue de fonctions, parfois appelées modules, visant à travailler la structuration de l’information (blocs, icônes, styles), le traitement de données (tableaux, fonctions arithmétiques, tris) ou encore à assurer l’interopérabilité avec d’autres systèmes (aperçus de fichiers, messagerie instantanée, messagerie, notifications). L’avantage est que la construction et la personnalisation de ces solutions liées aux différentes tâches du métier ne nécessitent que très peu de connaissances techniques.

La technologie wiki repose sur trois principes de base que nous allons présenter brièvement, le lecteur peut se

reporter à l'ouvrage de Jérôme Delacroix (2005) pour en savoir plus:

- la création et la modification de contenu grâce au balisage du texte;
- la structuration implicite et minimale qui s'accroît dynamiquement par le jeu de liens hypertextes;
- la gestion de contenu et le suivi de l'activité.

Le langage de balisage

Aujourd'hui, comme la plupart des logiciels d'édition de contenu, la majorité des wikis n'utilisent plus explicitement la syntaxe du langage à balises qui pouvait par ailleurs différer d'un wiki à l'autre. Les outils proposent dorénavant des interfaces WYSIWYG conviviales qui permettent des mises en forme élaborées sans connaissance préalable de la syntaxe wiki. Néanmoins, comprendre le principe de la syntaxe wiki permet de mieux appréhender son fonctionnement. Au fur et à mesure de l'édition de contenu, le rédacteur ajoute des commandes sous forme de balises qui vont influencer l'apparence du texte. Ainsi, le plus souvent, l'ajout d'un point d'exclamation indique un titre de rang 1, l'ajout de deux points d'exclamation un titre de rang 2. Il est possible de préciser la fonte d'une police d'écriture, la présence d'une liste, d'une image, d'une URL, etc.

Une structuration dynamique

À partir d'une première page – la page d'accueil – plusieurs pages peuvent ainsi être créées en ajoutant un simple lien entre crochets ou en syntaxe *CamelCase*, une notation qui met en majuscule l'initiale des mots associés afin d'éviter d'ajouter des espaces. Cette combinaison de signes est interprétée par l'application comme l'ajout automatique d'une page vierge portant le titre des termes placés entre crochets ou dans le mot composé. La structure d'un wiki ne

préexiste donc pas forcément à l'écriture, elle émerge au fur et à mesure des besoins et des actions des usagers. La plasticité est ici pleinement exprimée puisque des nouveaux lieux de réflexion ou de dialogue s'ouvrent au fil des besoins.

Ainsi la page d'accueil du wiki d'un chercheur pourra contenir une liste de tâches, ordonnée ou non, par thématique (recherche, enseignement, vie personnelle) ou par priorité (urgent, à faire, à lire, à regarder). La possibilité d'ajouter des liens, sous forme d'URL locale ou distante ou d'ajouter des documents (images, notes…) simplifie la navigation et permet un accès facilité aux ressources. Il est par exemple possible d'avoir une liste des appels à publications classés par date, avec un lien vers l'appel proposé par le comité d'organisation.

Ces liens peuvent renvoyer vers des notes personnelles, des documents (la page de suivi d'un mémoire, un texte en rédaction pour une agence nationale de recherche) ou des pièces d'ordre logistique (un ordre de mission), ce qui permet de structurer toutes les activités hétérogènes du chercheur au sein d'un même espace.

À ce premier niveau de structuration qui s'appuie sur le caractère hypertextuel du wiki, s'ajoute un second niveau de structuration. Celui-ci repose sur le filtrage à l'aide de mots-clés (ou *tags*) permettant de repérer les contenus parfois transversaux qui se rapportent à une ou plusieurs thématiques, à un ou plusieurs projets. De nombreux logiciels, comme Zim, offrent une fonctionnalité de placement d'étiquettes dans les pages. Par la suite, la fonction «sélection-filtrage» permet d'accéder très rapidement aux pages contenant l'étiquette ou de les regrouper.

Une gestion de contenu et un suivi de l'activité

Enfin, fondamentalement conçu pour l'activité collaborative, le wiki offre généralement des outils avancés de gestion de

contenu comme:

- la gestion des versions via un historique;
- la possibilité de comparer des versions deux à deux (en général, les versions sont présentées côte à côte, les différences sont colorisées);
- la réversibilité pour permettre le retour à des versions antérieures (restauration).

Outre ces aspects se focalisant sur les contenus, d'autres fonctionnalités tournées vers l'activité sont également présentes, comme des listes de contenus récemment mis à jour par d'autres utilisateurs, les notifications d'actualisation de pages ou d'espaces que nous avons choisi de surveiller ou encore, sur certaines solutions logicielles, la possibilité d'être alerté lorsque son nom d'utilisateur est cité. Ainsi, des bandeaux de menu vont proposer des fonctions en accès rapide: la possibilité de restreindre les modifications, d'ajouter des documents ou des *tags* à une page, et de choisir d'envoyer ou non des notifications aux observateurs de la page, éventuellement en documentant les modifications.

Le wiki en pratique

Une fois ce cadre technique posé, nous souhaitons exposer comment le wiki peut devenir un outil de gestion de la recherche, un véritable compagnon du chercheur au quotidien. Pour cela, nous allons nous focaliser sur quelques utilisations possibles, illustrées par des éléments tirés de nos expériences et de nos pratiques.

Un système d'information documentaire

Qui n'a jamais fait face au problème de classement de documents aux statuts hétérogènes, aux multiples versions dont il est tantôt le destinataire, tantôt l'auteur? Agissant parfois dans l'urgence, le chercheur ne prend pas

nécessairement le temps de contextualiser les supports avec lesquels il travaille. Le wiki constitue un outil aidant à reprendre un tant soit peu le contrôle sur le chaos dont sont en partie responsables les froides et mécaniques arborescences de fichiers. Dans un wiki, tout document est automatiquement placé dans un contexte, au sein d'un discours, même minimal, servant à comprendre les circonstances de son utilisation ou de sa création. La plate-forme permet ainsi de renouer avec une forme de documentarisation certes potentiellement non contrôlée, mais répondant à des logiques «métiers» et possédant une pertinence propre aux acteurs concernés. Ainsi, le dépôt de pièces jointes, l'ajout d'étiquettes, la constitution de collections dynamiques et le suivi des versions de documents qu'offrent certains wikis constituent, selon nous, un apport indéniable de cet outil.

Un système de gestion de projet

Le travail du chercheur s'inscrit de plus en plus dans des projets collectifs. Loin de nous l'idée de considérer les dispositifs techniques de collaboration comme déterminants dans une telle démarche (les aspects humains et sociaux sont, selon nous, bien plus prégnants), cependant, nous tendons à penser que les diverses médiations, y compris techniques, contribuent à soutenir la production du groupe, sa cohésion et son information.

La structuration dynamique des différentes pages du wiki sert généralement la structuration du projet. Le chercheur peut alors s'appuyer sur une structuration arborescente de pages afin de définir l'ensemble des éléments qui décrivent le périmètre de son projet, que ce soit en matière de phases ou de livrables. Ainsi l'arborescence du wiki du *Publictionnaire* articule des éléments organisationnels (comptes-rendus des réunions, listes de tâches, coordonnées des auteurs) et des contenus scientifiques (notices de dictionnaire, biographie des auteurs).

La plupart des wikis proposent également aux utilisateurs de constituer des listes de tâches avec la possibilité, pour chacune d'elles, de définir son statut (terminé, non terminé), la personne responsable de sa réalisation et la date à laquelle celle-ci doit être réalisée. Ces listes peuvent constituer des tableaux de bord pour assurer le suivi d'avancement du projet. Elles peuvent aussi rendre compte des résultats obtenus par les personnes responsables de chacune des tâches ou des étapes. De même, il est également possible de prévoir un flux de travail (*workflow*), impliquant par exemple des auteurs et des relecteurs interagissant dans plusieurs itérations. Le projet *Publictionnaire* s'articule autour d'un tableau de bord qui synthétise toutes ces possibilités. On y trouve pour chaque entrée l'auteur envisagé, le chercheur en charge de la communication avec l'auteur, l'état de la notice (en attente, reçue, validée, en discussion), ainsi que des commentaires qui permettent au comité éditorial d'échanger.

Un espace de construction et de discussion

Le wiki se fonde sur une certaine idéologie du partage et participe à construire informatiquement un lieu d'échange entre acteurs d'une même communauté, un espace collaboratif spécifiquement orienté vers la pratique de l'écriture. Le modèle cognitif de Flower et Hayes (1981, p. 370-374) aborde l'écriture comme processus en trois temps et offre un éclairage intéressant quant au rôle que peut jouer le wiki dans un processus d'écriture collaborative. Il distingue:

- une phase de *planification*: les différents contributeurs peuvent émettre de nombreuses propositions soumises au jugement ou à la créativité des autres participants. Le wiki fonctionne ici comme aide à la créativité en

favorisant les associations d’idées et en assurant la traçabilité des échanges entre participants;

- une phase de *mise en texte* durant laquelle la mise en forme du texte rapproche l’utilisateur du wiki de l’expérience de l’écriture par l’usage de logiciels de traitement de texte. L’attention du contributeur est centrée sur la production et la structuration du contenu, au moyen d’outils de mise en forme et d’organisation hiérarchique de l’information (titres, listes et blocs). La publication progressive de versions intermédiaires permet ainsi à chaque auteur de suivre l’évolution du texte auquel il contribue;
- une phase de *révision*, où le wiki permet l’ajout ou la modification du texte final, l’annotation de parties ou de l’ensemble sous forme de commentaires. Des discussions peuvent s’engager, comme sur un forum, à propos d’éléments de connaissance explicités sur le support. L’intérêt majeur du wiki, à ce stade, réside à notre sens dans le fait que ces échanges ne sont pas dissociés de leur objet (élément textuel ou paratextuel) ce qui facilite la compréhension du cheminement intellectuel de la discussion. Dans le cadre du projet *Publictionnaire*, de nombreuses discussions sont menées dans la partie commentaires, les échanges, moins formels, y sont conviviaux et il est possible d’exprimer son approbation (*like*) à propos des interventions de chacun.

Typologie de wikis

La plupart des institutions mettent à disposition des chercheurs des wikis adaptés et déjà installés, ce qui les dispense des questions techniques de l’hébergement, de l’installation, de la sécurisation et du paramétrage de l’outil. Il peut néanmoins être intéressant, selon les caractéristiques et les besoins du projet et des utilisateurs, de mettre en place un wiki dédié et indépendant. La

typologie suivante synthétise les différentes options offertes pour la mise en place d'un wiki. Nous avons volontairement limité le nombre d'exemples en choisissant parmi les dizaines de solutions mises à disposition celles qui semblaient le plus en adéquation avec les besoins et les compétences d'un chercheur en SHS. Pour aller plus loin, le site wikimatrix.org offre une bonne recension en anglais et permet la comparaison entre plusieurs solutions.

Les **wikis de bureau**sont destinés à un usager qui souhaite saisir et organiser ses notes personnelles, ils permettent d'intégrer des ressources (images, vidéos), de hiérarchiser des documents et de gérer via une interface unique tous les aspects de la vie du chercheur (planning, liste de tâches, liste d'appels à soumission ou de projets, liste de contacts, etc.). Le logiciel TiddlyWiki, compatible avec les navigateurs Firefox, Safari, Chrome ou Zim (logiciel libre), disponible sur Windows, BSD ou GNU/Linux correspond bien à ce premier type.

Les **fermes de wiki**comme Wikispaces ou Wikia (libre) prennent en charge l'hébergement, les mises à jour techniques et les questions de sécurité. Elles offrent des espaces dédiés qui peuvent être facilement mis en place.

Les **applications hébergées**(DokuWiki, PmWiki, MediaWiki, tous trois *Open Source*) sont destinées à être installées sur un serveur (local, sur un ordinateur, ou distant, sur Internet), l'utilisateur doit donc prendre en charge les questions techniques mais cette solution évite de confier à d'autres l'hébergement de données, parfois sensibles dans le cadre de projets de recherche, et permet de contrôler l'intégralité des fonctions.

Les **logiciels institutionnels**(mais non moins hébergés) comme Confluence ou le libre XWiki sont, par la qualité des interfaces et la richesse des fonctionnalités offertes, des solutions idéales pour la gestion de projets ou d'équipes. Répondant explicitement à des besoins professionnels, ils proposent des outils particulièrement avancés (import des

formats Office, connexion à un annuaire, etc.) et on les désigne sous le terme de «wikis seconde génération».

Le wiki en contexte

Le discours qu'on tient dans cet article est celui de deux adeptes convaincus et, même si les expériences nous ont instruits des limites de l'outil – nécessaire travail d'animation, hétérogénéité des investissements, subjectivité de la structuration, rigidité des formats… –, il nous a paru intéressant de solliciter les regards d'utilisateurs moins familiers des wikis. À partir du projet *Publictionnaire*, nous avons recueilli la parole de collègues, membres du comité d'organisation, dont la diversité garantissait une certaine représentativité. Ces chercheurs, de générations et de disciplines multiples, ouvrent notre propos à d'autres problématiques que nous synthétisons ici.

Les complémentarités des compétences

Dans le cadre de l'écriture ou de la gestion collaborative d'espaces wikis, le partage des rôles est souvent spontané: un des participants œuvrera plus naturellement aux corrections orthographiques et syntaxiques, un autre s'attardera sur le formatage, tandis qu'un troisième portera un soin particulier à la gestion de l'espace de travail en supprimant les pages obsolètes ou en modifiant la structure de l'arborescence. Le wiki permet alors d'appréhender de manière concrète les rouages d'une intelligence collective en construction:

> Chacun ayant non pas un rôle particulier défini mais une manière qui est propre à chaque personne d'investir. Il n'y a pas de responsable en titre donc c'est un vrai collectif quasiment autogéré qui fait, qui avance [Entretien 2].

La dimension structurante

Spontanément, les participants évoquent l'intérêt de rassembler en un seul lieu, hiérarchisé, les documents et les échanges. Les fonctions d'historique qui permettent de garder toute trace des interactions comme des versions de documents sont jugées particulièrement pertinentes dans le cadre d'un projet sur le long terme où les objectifs sont amenés à évoluer au cours du temps.

> Voyez ce tableau de bord, il est quand même relativement cadré, il y a des zones, ça a l'air de rien, mais c'est très aidant, enfin c'est très structurant [Entretien 3].
>
> Ça permet aussi de se retrouver comme sur toute plate-forme documentaire: on a tous les documents liés au projet qui sont disponibles, auxquels on peut se référer […] si on a un souci, il y a des documents archivés et c'est véritablement une ressource [Entretien 1].
>
> Les avantages donc c'est de garder une mémoire à laquelle éventuellement on peut se reporter en réunion c'est-à-dire qu'on peut projeter [le wiki sur un écran] et dire: «Tiens à telle époque on avait dit ça» [Entretien 3].
>
> C'est tout à fait pratique, cela ne nous fait pas perdre de temps, au contraire ça en fait gagner, ça rationalise tout à fait bien les choses […] Je vois sa plus-value, ça évite des *mails* à n'en plus finir et c'est parfait. Ça c'est tout à fait pratique, donc c'est l'apport principal [Entretien 5].

La gestion du temps et de l'espace

Dans un contexte professionnel d'accélération temporelle, où tous ne peuvent toujours être présents, en même temps, au même moment, quand des acteurs peuvent rejoindre tardivement des projets, le wiki peut apparaître comme une solution asynchrone intéressante.

> Comme je n'étais pas au démarrage du projet, au début je me suis baladée dans le wiki pour essayer de comprendre, pas tellement pour comprendre l'outil,

> mais de comprendre à travers l'usage qui avait déjà été fait de l'outil, ce qui se dessinait comme projet [Entretien 1].
>
> La mise en place d'un wiki qui est tout à fait commode et qui a tout à fait du sens, d'une part au vu du nombre de participants, il est vraiment difficile de réunir beaucoup de personnes en un même lieu donc c'est bien pratique […] [Entretien 5].

Les craintes liées à l'outil

Questionnés sur leurs postures au démarrage du projet, les participants mentionnent souvent des craintes que l'usage n'a pas toujours défaites.

> J'ai toujours un peu peur d'un parasite de l'activité de recherche par des considérations techniques alors même que la technique peut faciliter la recherche [Entretien 2].
>
> La première fois, j'ai eu peur de perdre du temps dans l'outil au détriment d'objectifs plus urgents […], peur de perdre des données si on ne maîtrise pas l'outil [Entretien 1].
>
> Les commentaires dans un wiki, y a toujours des ambiguïtés, y a toujours la potentialité du moins qu'il y ait des ambiguïtés, des mécompréhensions, des choses qui ne vont pas et qui finalement nous donnent plus de travail [Entretien 4].

Quelles qu'aient été ces craintes initiales, leurs auteurs font néanmoins le constat que tous ont finalement adopté l'outil.

> Il y a des gens qui étaient un peu réticents, alors je ne sais pas si c'est par peur de l'outil ou parce qu'ils ne se sentaient pas à l'aise et qu'ils pensaient qu'ils n'allaient pas forcément y arriver mais qui s'y sont mis finalement et qui l'utilisent pas mal [Entretien 4].

> Ça a évolué parce que moi je me rappelle des réunions et des collègues me disant en aparté: «Oh là là, moi je suis pas du tout à l'aise avec ça» et je pense que maintenant ça va. À chaque fois qu'il y a des mises à jour […] je vois les collègues qui mettent à jour et c'est pas seulement le petit noyau de quelques-uns [Entretien 3].

> Je me suis aperçue que certaines personnes [ça avait] été pour elles l'occasion d'utiliser pour la première fois un outil wiki, ça c'est sûr, ça a été une découverte pour certaines personnes je pense [Entretien 3].

Les personnes interrogées expliquent spontanément que cette prise en main sans résistance ni difficulté technique facilite l'adoption:

> Par définition, il faut forcément du temps mais j'étais tout de suite très à l'aise, d'une part parce que l'interface est familière, ce n'est pas une interface complètement nouvelle ou dérangeante, il y a tout de suite une impression de familiarité et au-delà de l'impression, les fonctionnalités ressemblent aux outils de bureautique que j'ai l'habitude d'utiliser donc je n'ai pas eu le sentiment d'être en terre inconnue et en plus je n'ai pas rencontré de problèmes réels. J'ai peut-être hésité… oui j'ai hésité deux ou trois fois mais soit j'ai trouvé rapidement la solution, soit je ne sais pas… c'est un peu confus mais il n'y avait pas de réel problème [Entretien 4].

> Le logiciel en tant que tel, moi je n'ai pas rencontré de limites particulières. […] Ce n'est pas extrêmement complexe ce que l'on fait avec le logiciel. […] Ce sont des solutions qui sont faciles à prendre en main. […] Donc même des gens qui sont un peu réticents ou qui n'ont pas de temps à y investir comprennent tout de suite comment on fait pour aller modifier un document collaboratif [Entretien 5].

Ceux qui étaient réfractaires à l'outil évoquent deux arguments au départ, la complexité technique et la dénaturation du travail scientifique, qui se sont finalement dissous au fur et à mesure des utilisations.

> Ça reste un outil et ça ne va pas influencer ma manière de travailler, la manière dont j'organise le travail, dont je donne des priorités… enfin ça ne va pas infléchir mon travail [Entretien 4].

Une boîte à outils ou un métaoutil?

Tant qu'on ne les prend pas en main, les outils ne sont que de simples objets de notre environnement. Nous avons souhaité proposer au chercheur de se saisir de cet outil car il nous semble aujourd'hui que, face à la prolifération et à l'hétérogénéité des outils de gestion et d'édition, dans un contexte scientifique et politique qui encourage la dynamique de l'action collective, cet objet pouvait se révéler pertinent. La fluidité de l'utilisation et la plasticité du dispositif l'investissent d'une dynamique autonome qui permet de concevoir des espaces intellectuels et organisationnels dans une même dynamique, en dehors de toute compétence technique. Ces qualités en font un vecteur d'intelligence collective autant qu'individuelle:

> L'être le plus intelligent est celui qui est capable de bien utiliser le plus grand nombre d'outils: or, la main semble bien être non pas un outil, mais plusieurs. Car elle est pour ainsi dire un outil qui tient lieu des autres. […] Car la main devient griffe, serre, corne, ou lance ou épée ou toute autre arme ou outil. Elle peut être tout cela parce qu'elle est capable de tout saisir et de tout tenir (Aristote, *Les Parties des animaux,* Livre IV, ch. X, 687a).

Traversées par la tension numérique, les humanités font face au besoin impérieux de développer des outils pour appréhender des territoires encore inconnus, sans contraindre le chercheur dans des formats stérilisants ou

chronophages en matière d’apprentissage. Dans ce contexte, le wiki ne pourrait-il pas être cette main, ce méta-outil, qui lui permet de tout saisir et de tout tenir, de devenir griffe ou serre, selon ce qu’il convient au chercheur?

CHAPITRE 7

Du digital naive au bricoleur numérique: les images et le logiciel Omeka

CÉCILE BOULAIRE ET ROMEO CARABELLI

L'objet qui nous a réunis, alors que notre ancrage disciplinaire et nos objets de recherche sont très éloignés les uns des autres, est le logiciel Omeka (https://omeka.org/), développé par le Roy Rosenzweig Center for History and New Media (RRCHNM), qui promeut aussi le logiciel de récupération et de gestion de références bibliographiques Zotero. Omeka est un *Content Management System* (CMS) dédié à l'organisation, à l'exposition, à la mise en ligne de données iconographiques, avec leurs métadonnées, qui permet d'en faire très facilement une publication sur le Web. Ce chapitre sera pourtant consacré moins à Omeka qu'au rapport que nous entretenons, comme chercheurs, aux logiciels que nous utilisons. Dans l'exemple d'Omeka, nous essaierons d'analyser l'usage que nous avons fait des diverses applications nécessaires à nos recherches, de comprendre les impasses que nous avons connues, et surtout, d'extraire de cette expérience quelques préceptes de bon sens destinés à la communauté des chercheurs en lettres, sciences humaines et sociales (LSHS).

C'est en effet comme chercheurs en LSHS que nous nous définissons, et c'est à ce titre que nous souhaitons réfléchir à notre posture dans le champ des humanités numériques (HN), en racontant comment nous sommes passés du stade de *digital naive* à celui, tout aussi modeste, de bricoleur

numérique – le bricolage s'entend ici au sens où Claude Lévi-Strauss le défend, dans *La Pensée sauvage:* «une science que nous préférons appeler "première" plutôt que primitive», une science liée à l'accident, au «mouvement incident: celui de la balle qui rebondit, du chien qui divague, du cheval qui s'écarte de la ligne droite pour éviter un obstacle» (1960, p. 26). Le bricoleur, selon l'anthropologue, est «celui qui œuvre de ses mains, en utilisant des moyens détournés par comparaison avec ceux de l'homme de l'art». Nous souhaitons ici présenter les démarches de deux chercheurs en LSHS qui, parce qu'ils sont pris dans les réalités concrètes de la vie universitaire (en matière de moyens, de formation, de personnel d'accompagnement de la recherche), mais aussi parce qu'ils sont absorbés dans ce qui fait l'essence de leur métier, la recherche elle-même, en sont amenés à *adapter* les outils numériques à un faisceau de contraintes et de besoins qui leur sont propres… et qui changent avec chaque nouveau projet. Nous revendiquons l'appellation «bricoleurs» dans l'univers des HN. Ni développeurs, ni ingénieurs, ni documentalistes de notre propre recherche, ni photographes, ni chargés de communication ou de vulgarisation scientifique, nous avons pourtant besoin d'un petit peu des compétences de chacun de ces professionnels, pour faire avancer nos travaux. C'est donc armés de savoir-faire hétéroclites et animés par des besoins extrêmement précis et définis par notre recherche que nous abordons les outils des HN. Un peu comme nous utilisons notre voiture: nous aimons qu'elle fonctionne, et nous savons où nous voulons aller parce que nous avons quelque chose à y faire; mais nous ne souhaitons pas apprendre la mécanique.

Nos projets

Pourquoi Omeka? Au début des années 2010, nous sommes l'un comme l'autre engagés dans des projets de recherche (nationaux, européens) qui nous amènent à brasser des

quantités importantes de données iconographiques et textuelles. Nous avons besoin de stocker ces données, de les classer, de les rendre accessibles aux autres chercheurs du projet, dispersés géographiquement. Nous pensons aussi à l'*après*, et imaginons déjà comment mettre ces données à la disposition d'une communauté plus vaste (autres chercheurs, simples curieux) en les rendant notamment accessibles sur le Web. Idéalement, elles devraient non seulement être accessibles à d'autres utilisateurs, mais mises en valeur, expliquées, éditorialisées, enrichies perpétuellement. L'un des projets concerne l'entreprise Mame, une maison d'édition de livres religieux et de livres pour enfants installée à Tours depuis la fin du XVIIIe siècle. Une équipe financée par l'Agence nationale de la recherche (programme Jeunes chercheurs) se consacre à son histoire, en tentant pour ce faire de reconstituer virtuellement une partie de son catalogue, car toute l'entreprise a brûlé, avec ses archives, en 1940. L'équipe a commencé à recenser les livres, en utilisant une base de données propriétaire, FileMaker, qui donne toute satisfaction… sauf pour la mise en commun des données, extrêmement fastidieuse. De plus, ce logiciel commercial n'est pas du tout pensé pour un débouché sur le Web alors qu'on souhaite pouvoir «montrer» la production Mame, la rendre publique, la partager. L'équipe recherche donc une solution de remplacement, d'autant qu'elle souhaite associer aux fiches bibliographiques rédigées par les chercheurs les photographies des livres, réalisées dans les bibliothèques et les fonds privés qui conservent une partie des publications. Ces photographies encombrent les disques durs des chercheurs, et sont à la fois difficiles à partager et mal répertoriées, car aucune réflexion sur leurs métadonnées n'a été envisagée d'emblée: nous avons pris des photos d'instinct.

L'autre projet, Mutual Heritage, concerne le patrimoine architectural et urbain récent dans le monde méditerranéen;

il a pour cadre plus large le programme européen Euromed Heritage 4. Ce projet vise à inventorier, documenter et promouvoir le patrimoine récent des XIXe et XXe siècles, afin d'encourager l'intégration du patrimoine culturel dans la vie économique et sociale actuelle. Les acteurs de ce programme, de provenances multiples, ont donc collecté des images photographiques d'éléments architecturaux patrimoniaux, pour les villes d'Alger, Casablanca et Tunis. Les chercheurs réalisent et documentent ces clichés géolocalisés sur le terrain. Ils sont la matière première de la recherche et de la réflexion; ils peuvent aussi, à terme, constituer une base d'exemples à la disposition des professionnels concernés par les questions d'urbanisme et de patrimoine; ils sont enfin destinés à être mis en valeur sous la forme d'expositions virtuelles qui récapituleront les principaux enseignements de cette recherche. Deux des points communs de nos projets résident dans le fait qu'ils sont l'un comme l'autre pris dans une temporalité contrainte (projets financés), et dans le fait qu'ils nous conduisent à brasser des données sérielles (ce qui est loin d'être le cas général en LSHS).

Après quelques errances (pour le projet Mame, elles sont exposées au jour le jour dans le carnet de bord Mame & fils, en ligne: http://mameetfils.hypotheses.org/), chacun de nous trouve au sein de la Maison des Sciences de l'Homme (MSH) de Tours (désormais MSH Val de Loire) une aide technique essentielle. Au sein de l'atelier numérique, nous sommes aiguillés vers la solution Omeka; c'est à cette occasion que nous faisons connaissance, et mettons en commun nos interrogations de chercheurs.

Nos attentes de *digital naive*

Omeka est un *Content Management System* (CMS), c'est-à-dire un logiciel permettant de créer un site Web assez complexe avec une assistance technique limitée. En effet, une interface d'administration, avec des champs à remplir

en langage naturel, sert d'intermédiaire à l'usager. À cet égard, si l'on est familier d'une interface de blog comme WordPress par exemple, on est en terrain (relativement) connu. La particularité d'Omeka, ce qui fait tout son intérêt pour des projets qui concernent le patrimoine, pour lesquels on cherche souvent à collectionner et mettre en ordre des images, c'est qu'il est précisément dédié à la mise en valeur de collections importantes. C'est d'ailleurs le terme utilisé dans la nomenclature même du logiciel pour organiser les données: celles-ci sont rangées en «collections». À l'intérieur d'une collection, chaque élément (dans l'un des cas évoqués, c'est un livre des éditions Mame; dans l'autre, c'est un bâtiment présentant un intérêt patrimonial) fait l'objet d'une description précise et normée. Le schéma de métadonnées choisi pour Omeka est le *Dublin Core*, qui a l'avantage d'être une norme internationale, permettant éventuellement le moissonnage automatisé des données (récupération de données comparables depuis des entrepôts *Open Archives Initiative – Protocol for Metadata Harvesting* (OAI-PMH), par exemple la Bibliothèque nationale de France (BNF), mais aussi mise à disposition, dans un entrepôt OAI-PMH, des données constituées par les chercheurs, à l'intention d'une communauté large). C'est par ailleurs un *Dublin Core* «enrichi»: en plus des 15 champs réglementaires du *Dublin Core* de base, Omeka offre la possibilité d'utiliser 15 autres champs (on parle alors de *Dublin Core* «qualifié») qui viennent s'ajouter aux premiers, ou plutôt les raffiner (grâce au *plugin Dublin Core Extended*). Enfin, il ouvre encore la possibilité de définir soi-même certaines entrées (*Item Type*), selon une nomenclature personnelle, en plus des champs de base.

On peut associer des images à chaque élément de la collection: pour les livres, nous pouvons ainsi inclure une photographie du premier plat de couverture, de la page de titre, des illustrations, voire de toutes les pages qui présentent un intérêt particulier par rapport au projet

(préface ou *ex-præmio* par exemple); pour les bâtiments, il est possible de proposer plusieurs vues, ainsi que des détails pertinents. Et puisque chacune de ces photographies est en elle-même une source, nous pouvons éditer une notice descriptive pour chaque photo; elle donne accès, par exemple, aux données EXIF et XMP de la photo, mais aussi à tout commentaire singulier que souhaiterait inclure le chercheur. Il est possible aussi d'adjoindre à chaque document un lien hypertexte vers une ressource Web, ou une source extérieure, par exemple un PDF. Omeka remplit ainsi une fonction de base de données. Lors de l'affichage Web, l'internaute a accès à des champs de requête, qui peuvent lui permettre de demander par exemple tous les livres publiés dans la Bibliothèque de la jeunesse chrétienne en 1850, ou tous les bâtiments d'Alger dessinés par l'architecte Gabriel Darbeda.

Mais Omeka peut aussi servir à mettre en forme ces données, de manière plus linéaire et plus construite que sous la forme d'un vaste réservoir de données brutes. Le *plugin Exhibit Builder* permet en effet de créer des expositions virtuelles. Autant dans la première phase du travail le chercheur se sert d'Omeka comme d'un outil d'appui à sa recherche, lui permettant de rationaliser l'indexation et le partage des données collectées sur le terrain, autant dans cette seconde phase il se retrouve dans sa posture intellectuelle favorite, celle qui consiste à donner sens à un amas de faits, d'images, de phénomènes. Il va sélectionner les images qui lui paraissent les plus à même de faire percevoir les réalités essentielles; il va mettre en valeur ces images en les commentant; il va organiser le parcours intellectuel de son lecteur, construire un discours élaboré, scénariser la mise en forme des conclusions de sa recherche, comme on le fait en établissant la table des matières d'un ouvrage de synthèse ou en assurant un commissariat d'exposition. Ici, l'intérêt d'Omeka réside dans la possibilité d'organiser très facilement ces expositions sur

le Web. L'interface qui contient de manière préprogrammée quelques types de mise en page facilite la mise en forme; on puise les images dans les collections, et le chercheur rédige librement les textes explicatifs. Le programme Mutual Heritage a ainsi monté une exposition virtuelle intitulée «Des styles architecturaux variés», qui permet une entrée synthétique dans le corpus et les réflexions de l'équipe de recherche (cette exposition n'est actuellement plus visible, car elle n'a pas supporté la montée de version).

Le programme Mame a créé une exposition consacrée aux albums publiés durant l'entre-deux-guerres (http://mameetfils.univ-tours.fr/exhibits/show/albums_expo). Ces expositions sont complémentaires à la mise à disposition des données intégrales.

Nos déceptions

Tout cela semble parfait, et facile. Mais dans la réalité, les choses ne se déroulent pas de manière linéaire, loin de là. Peut-être parce qu'on nous propose la solution Omeka en cours de projets, et non à l'ouverture de ceux-ci – mais à l'ouverture précisément, nous ne savions ni l'un ni l'autre que nous allions avoir besoin de ce type de produit! C'est bien l'expérience de nos terrains qui fait émerger le besoin d'outils permettant de négocier les données produites par l'enquête. De ce fait, l'un comme l'autre, dans nos équipes, nous avons mis en place (ou laissé s'installer de manière anarchique) des méthodologies et des démarches disparates, qui se révèlent à ce stade peu compatibles avec la rigueur (très souvent perçue comme de la rigidité) d'un CMS comme Omeka.

Un exemple: les chercheurs du projet Mame sont à parts égales des historiens et des littéraires, des usagers de catalogues de bibliothèques, mais n'ont pas reçu de formation au catalogage normé. Lorsqu'ils ont entre les mains un ouvrage de la maison Mame qui leur paraît présenter de l'intérêt pour la recherche collective, ils le

décrivent donc en privilégiant des informations pertinentes pour eux, par rapport à leur orientation épistémologique – certes, en respectant quelques fondamentaux concernant les références bibliographiques, mais en ajoutant des champs de description très personnels, et très liés à la nature des informations que l'équipe a besoin de réunir, parce que justement aucun catalogue de bibliothèque *existant* ne les a fournies jusque-là. Le cas est loin d'être anecdotique. Un débat récent entre Stéphane Vial et Pierre Lévy sur le blog *Design & Digital Humanities* rappelle que l'une des grandes activités des chercheurs en LSHS consiste justement en ce que Lévy appelle la «curation», c'est-à-dire l'établissement de ces corpus et bases de données (Vial, 2015), à partir d'observations de terrain et de dépouillements.

L'équipe de Mame a donc constitué sa propre base de données, aidée en cela par la souplesse de prise en main de FileMaker… or la convertir en *Dublin Core* pour passer sur Omeka va se révéler très compliqué et fastidieux. Techniquement, ce n'est pas impossible: il est toujours imaginable d'exporter les données de FileMaker vers un tableur; et le module CSV Import d'Omeka permet de récupérer ces données. Encore faut-il pouvoir «glisser» les informations que l'équipe a considérées comme importantes, et qu'elle a organisées selon sa logique propre, dans le schéma contraignant du *Dublin Core*! L'opération est possible, mais elle conduit souvent à fusionner des informations initialement distinctes pour les intégrer à un même champ, ce qui prive de la possibilité de faire des requêtes fines sur ces champs «écrasés». Or il est hors de question, pour les chercheurs, que l'effort de normalisation technique fasse disparaître des informations scrupuleusement amassées pour leur qualité inédite. L'exercice de bascule représente donc une acrobatie intellectuelle qui freine la recherche plus qu'elle ne l'aide – à ce stade, plusieurs chercheurs de l'équipe s'interrogent sur

la pertinence d'un outil dont la prise en main demande tant de temps et d'efforts, pour un résultat moins satisfaisant.

Certes, on aurait pu éviter facilement une telle dépense d'énergie par une réflexion *anticipée* sur la structuration des champs de saisie. Mais nous avons tous fait l'expérience de données non prévues dans un protocole de recherche initial – parce que nous n'avions pas anticipé l'existence de ces informations, ou l'intérêt qu'elles pourraient présenter, une fois réunies – et qui se révèlent pourtant centrales, en cours de route, pour produire un regard entièrement neuf sur nos objets. Un exemple: pour indexer les livres Mame, nous n'avions pas prévu de détailler les informations liées aux *ex præmio*, les étiquettes collées dans les livres et indiquant à quelle occasion scolaire le livre a été offert et à quel enfant. Or ces *ex præmio* étaient présents dans la majorité des livres que nous trouvions (alors qu'ils sont évidemment absents des ouvrages de la BNF, qui par définition n'ont jamais connu d'usagers), et ils fournissaient des informations qualitatives d'une importance essentielle: à quel âge étaient destinées telles collections et séries, à quel sexe tel livre était prioritairement offert, combien de temps s'écoulait entre la date d'impression d'un livre (présente sur le colophon) et la date de distribution lors des prix de fin d'année (manuscrite sur l'étiquette d'*ex præmio*), un indice précieux pour comprendre la circulation concrète du livre sur le marché des ouvrages scolaires. FileMaker permet de créer des champs de saisie de natures diverses, de les nommer librement, de les insérer n'importe où dans un masque de saisie, à n'importe quel stade du travail. Découvrant l'intérêt de ces *ex præmio*, nous improvisons d'emblée, et pour notre plus grande satisfaction de chercheurs, une manière d'en enregistrer les informations. De même que pour le projet Mutual Heritage, c'est seulement une fois sur le terrain que les chercheurs chargés de la ville de Tunis ont imaginé qu'ils pouvaient analyser les dynamiques de transformation d'un quartier patrimonial en

s’appuyant sur une donnée facile à obtenir: le prix du mètre carré locatif moyen du quartier considéré (fourni par les journaux d’annonces immobilières). Mais où ranger cette information, donnée essentielle à la recherche, dans un schéma *Dublin Core*? Tout chercheur sait donc bien à quel point c’est le terrain qui construit empiriquement sa réflexion, et combien sa collecte évolue dans ses modalités à mesure qu’il confronte ses hypothèses au réel. Mais lors du passage au rigide *Dublin Core*, les choses se compliquent.

Une autre difficulté se présente aux deux équipes: le temps nécessaire à l’indexation des données dans Omeka. Le chercheur en effet est toujours pris entre deux exigences, coincé entre deux temps: le temps de la collecte des données, qui est en réalité constitutif de ces données – car, au mépris de l’étymologie, les choses sur lesquelles s’exerce la réflexion ne sont pas «données» mais fabriquées par le regard du scientifique; et le temps de la réflexion, qui donne un sens à un amas de phénomènes, et engage l’action. La question du temps intervient ici. Le chercheur fait face à la difficulté d’engranger le plus possible d’informations, pendant une campagne à la durée forcément limitée (un séjour dans telle bibliothèque éloignée, une mission à l’étranger contingentée, ou encore un chantier de fouilles à l’existence restreinte par des délais légaux), en satisfaisant à l’exigence de précision, de méthode et d’exhaustivité induite par la perspective, à moyen terme, de l’exploitation puis de la mise en valeur des éléments collectés. Il est conscient que le temps passé à indexer et documenter des clichés photographiques empiète là sur la campagne de photos elle-même, ici sur la rédaction du compte-rendu de terrain, voire sur les nuits de récupération – sans compter que, sans être cynique, on évalue le chercheur sur sa production bibliographique… pas au nombre de données indexées! La nature fastidieuse d’une indexation normée, indispensable à l’utilisation du

corpus a posteriori, doit impérativement être atténuée par l'outil qui accompagne le chercheur dans sa collecte. Or Omeka ralentit le temps de saisie, obligeant à utiliser des nomenclatures peu ergonomiques et à travailler sur des interfaces peu intuitives.

Comme responsables d'équipe, nous savons d'expérience qu'il existe une loi simple régissant la collecte de données dans un cadre de recherche: n'étant pas payé en fonction du nombre de fiches réalisées, le chercheur ne se livre à une constitution fine de corpus que dans la mesure où il est convaincu que cet investissement temporel sera rentable au moment de la phase d'élaboration du savoir. Un exemple très concret: le logiciel FileMaker offre une énorme puissance de calcul et de croisement des champs de requête. Il est donc très facile à un chercheur, travaillant sur une masse de données recueillies collectivement, de décider soudainement, pour les besoins d'une publication, de chercher combien de titres différents Just-Jean-Étienne Roy a publiés aux éditions Mame entre 1855 et 1867: la requête est simple, la réponse immédiate, l'information utilisable, on peut la croiser avec autant de requêtes similaires qu'en exige la logique de l'article en cours. Les efforts que le chercheur a produits auparavant (et que ses collègues ont produits) en saisissant manuellement toutes sortes de données (notamment des pseudonymes) se voient donc immédiatement récompensés. Omeka offre moins de souplesse dans le croisement des requêtes, il classe moins bien les réponses et produit plus d'erreurs; le chercheur peut être déçu, au moment où il élabore du savoir, en constatant que le temps qu'il avait consacré à la production des données lui est si mal rendu.

À ce stade de nos deux programmes de recherche, nous constatons donc, l'un comme l'autre, qu'Omeka n'est pas le miracle de fluidité que nous avions (naïvement?) imaginé. Efficace dans l'idéal, il suppose, pour être vraiment opérationnel, que l'équipe ait préalablement mis sur pied

des procédures de travail finement calibrées. Or ces procédures, justement, il n'est pas toujours possible de les anticiper a priori, car le terrain apporte son lot de découvertes, qui orientent de manière rétroactive les démarches du chercheur, se dégageant d'une impasse ou à l'inverse investissant brusquement un territoire qui n'était pas cartographié au démarrage du projet. De *digital naive*, nous voilà devenus *digital disappointed*. Comment surmonter ce sentiment de relatif échec?

Le bricolage comme solution

Avec du recul, la solution qui nous apparaît la plus pertinente à développer devant une communauté de chercheurs est donc celle du bricolage. On ne pense pas le bricolage ici négativement, comme un rejet de la technicité ou du professionnalisme, mais positivement, comme une capacité d'adaptation à un contexte qui n'est, techniquement, humainement et économiquement, jamais idéal, et par définition, jamais prévisible – c'est ce qui distingue la recherche de l'expertise. Partant du principe que la montre la moins chère, c'est la montre que l'on possède déjà; que le mieux est l'ennemi du bien; que les journées ne comptent que vingt-quatre heures même chez les chercheurs publiants, notre réflexion nous a amenés à tempérer notre enthousiasme devant l'objet Omeka, et à revenir à une réflexion épistémologique moins tributaire du mirage techniciste qui semble parfois envahir les discours sur les HN.

Nos équipes ont trouvé Omeka d'un maniement difficile, chronophage, pour un rendu qualitatif inférieur aux outils empiriques que nous étions pourtant tentés de rejeter parce qu'ils n'étaient pas optimaux. Or l'honnêteté nous conduit pourtant à reconnaître les atouts qu'il présente. Comment obtenir d'Omeka (et du *Dublin Core* qui le structure) qu'il nous offre à la fois la rigueur qui nous permettra d'utiliser et de partager les résultats de la recherche, et à la fois la

souplesse dont nos équipes ont impérativement besoin durant le temps de l'enquête, qui est une exploration intellectuelle plus qu'une simple récolte? Nous avions choisi FileMaker ou Excel dans le tâtonnement préliminaire des deux projets, parce qu'ils sont faciles à comprendre, et très souples. Plutôt que de nous contraindre à rentrer dans la logique un peu raide d'Omeka, une solution d'avenir pourrait-elle être de lui inventer un «assouplisseur»? C'est là qu'intervient la notion de bricolage.

Imaginons un charpentier de marine qui, pour parfaire la courbure de la coque d'un bateau, se plaint que son herminette ait un fer trop droit. Si on lui répond «Tu dois faire avec cette herminette, car l'outil est ainsi conçu; pour lui conserver sa rigueur, il doit avoir cet angle de courbure, et pas un autre, adapte-toi», il y a fort à parier qu'il renoncera à terminer sa coque… ou que le bateau prendra l'eau. Mais l'expérience réelle est tout autre: il suffit en fait que le charpentier aille voir le forgeron, explique à quoi doit exactement ressembler l'outil dont il a besoin pour son art, afin que le forgeron produise précisément la forme de tranchant adaptée à l'usage ici et maintenant. Ou, si aucun forgeron ne se trouve à proximité, que le charpentier aménage lui-même son outil, certes sans la rigueur qui serait celle du forgeron, mais dans le but que le produit, et non l'outil, soit conforme à la rigueur la plus absolue: que la courbure soit bonne, et que le bateau flotte. Le charpentier et son forgeron sont-ils plus près du bricolage que de l'informatique? Sans doute. C'est cet esprit de bricolage adaptatif que nous revendiquons. Que le charpentier ne soit pas un forgeron; que le forgeron adapte l'outil à l'art du charpentier selon les exigences de ce dernier; que personne ne vienne dire à l'artisan qu'il doit se plier à la rigidité de son outil, mais qu'au bout du compte le travail soit fait, et bien fait; et qu'au besoin, le forgeron s'inspire des «aménagements bricolés» qu'en son absence le charpentier aura apportés à son outil, pour améliorer lui-même la forme

des prochaines herminettes qui sortiront de sa forge. Pour la plus grande satisfaction des charpentiers de marine, certes, mais surtout des navigateurs, pour qui les bateaux, *in fine*, sont faits.

Nous ne voulons pas bricoler Omeka – nous le voudrions, que nous ne le pourrions pas, car nous ne sommes pas compétents pour entrer dans le code, de même que le charpentier n'est pas forgeron. Mais nous voulons bricoler quelque chose qui nous offre les avantages d'Omeka sans nous imposer ses inconvénients – pour gagner du temps de recherche, pas pour en gaspiller. Il nous apparaît progressivement que la solution n'est pas technique, mais procédurale. Omeka est un excellent outil de gestion de données et de valorisation des résultats de la recherche. Mais il nécessite l'intervention d'un ingénieur, d'une part pour l'installer et le maintenir – les montées de version peuvent se révéler fastidieuses, comme nous en avons fait l'expérience –, d'autre part pour réfléchir avec l'équipe au meilleur protocole de distribution des tâches, celui qui garantit l'exploitation la plus rationnelle possible des capacités de chacun, sans les alourdir, et si possible en les allégeant. Plusieurs années après la fin de ces deux projets, nous sommes convaincus que la fonction de l'ingénieur, au sein de l'équipe de soutien, n'est pas d'imposer son outil, serait-il excellent (Omeka, d'un certain point de vue, l'est), mais d'aménager, avec les chercheurs, les *passerelles* qui permettront d'intégrer dans Omeka (qui offrira sa rigueur) les données qui auront été collectées sur le terrain avec les outils les plus souples et les plus élémentaires, ceux que les chercheurs connaissent bien, ont bien en main, ceux qui leur font gagner du temps et ne nécessitent ni connexion à internet ni disque dur surpuissant.

Encore faut-il que les procédures soient élaborées *ensemble*: pour que l'ingénieur puisse sans difficulté aucune rapatrier les données des chercheurs, ceux-ci doivent accepter de s'imposer un certain nombre de normes, dont le

nombre et la précision doivent être négociés; afin que les chercheurs aient la garantie que la nature et la qualité de leurs données soient reflétées de manière exacte dans l'objet final, l'ingénieur doit s'astreindre à bien comprendre l'intérêt suscité chez le chercheur par chacune d'entre elles; enfin, ingénieurs et chercheurs doivent se mettre d'accord sur les procédures, non pas accidentelles mais consubstantielles au travail de terrain, de modification et d'amélioration du protocole de saisie, quand soudain apparaît «quelque chose qu'on n'avait pas prévu» mais dont l'analyse est au cœur de la démarche de recherche. Le chercheur doit alors accepter de ne pas modifier anarchiquement et unilatéralement un champ de saisie ou la nature d'une donnée; l'ingénieur ne doit pas être autorisé à répondre «Ce n'est pas possible». C'est en cela que réside le bricolage: dans ce mouvement d'allers et retours entre chercheurs pris dans les réalités du terrain, et ingénieurs chargés du soutien technique – mais en gardant à l'esprit que ce qui prime est évidemment la démarche de recherche, et non la logique d'ingénierie. Bricoler une «moulinette» qui permettrait aux chercheurs d'utiliser les outils qu'ils connaissent et apprécient pour automatiser le versement des données vers Omeka, voilà ce qui, à notre sens, constituerait une véritable avancée en matière d'instrumentation numérique.

Pour en revenir à un cas concret: dans le cadre du projet Mame, nous avons basculé de FileMaker à Omeka. Il a fallu extirper les données de FileMaker, en les convertissant sous la forme d'un fichier CSV; puis concaténer certains champs, pour produire un équivalent *Dublin Core*; enfin réintroduire ces données CSV dans Omeka, via un *plugin*. À partir de cette date, on invitait les chercheurs à saisir directement dans Omeka les nouvelles observations menées sur le terrain. Avec du recul, nous pensons qu'il aurait fallu considérer cette étape intermédiaire (la moulinette FileMaker-CSV-Omeka) non pas comme une procédure

ponctuelle, mais comme la procédure ordinaire: que les chercheurs n'aient jamais à saisir sous Omeka, mais que la bascule se fasse à intervalles réguliers via la moulinette; et en économisant la déperdition liée à la concaténation des données, parce qu'une réflexion collective aurait permis de normer et réadapter en permanence les champs de saisie sous FileMaker pour les rendre d'emblée compatibles avec le *Dublin Core*.

Si le bricolage nous semble essentiel, c'est aussi, ne l'oublions pas, parce qu'en LSHS, il n'est pas rare qu'une équipe (équipe d'accueil, petit laboratoire sans moyens) ne dispose pas d'un ingénieur à demeure. Il est illusoire de considérer qu'un chercheur répondant à un appel à projets puisse se faire aider d'emblée, au moment de la rédaction de son programme, par un ingénieur maison, puisque la précarisation de l'emploi à l'université a raréfié les postes d'ingénieurs statutaires, en invitant à ne recruter que ponctuellement ces accompagnateurs de la recherche – donc *après* la réponse à l'appel à projets. Généraliser le bricolage, c'est une manière pragmatique de répondre au besoin d'instrumentation de la recherche, tout en s'adaptant à son contexte concret d'exercice. Toutes les équipes n'ont pas d'ingénieur; mais tous les chercheurs utilisent un tableur. Aidons les chercheurs à bricoler leurs moulinettes, à mieux utiliser les outils qu'ils connaissent déjà, ceux qu'ils possèdent déjà, mais en formalisant quelques règles simples qui leur éviteront les considérables déperditions d'énergie dont nos projets ont souffert lors des bascules vers des outils à l'ingénierie complexe.

Le bricolage n'est pas l'absence de rigueur

C'est donc en ayant à l'esprit les réalités concrètes de la vie du chercheur que nous avons cherché à aller plus loin dans notre réflexion sur l'instrumentation numérique de la recherche en LSHS. Il nous semblait en effet que, en amont d'Omeka, en amont même des «moulinettes» qui nous

avaient sauvé la mise, nous avions la possibilité d'améliorer des démarches mises en place empiriquement par les chercheurs dans leur usage de l'outillage numérique devenu ordinaire. Nous avions en effet le sentiment qu'une partie des difficultés que nous avions rencontrées aurait pu être évitées dès l'étape de la création des images qu'Omeka allait nous aider à mettre en ordre et en valeur. Une enquête que nous avons décidé de mener au sein d'une large communauté de chercheurs a confirmé cette intuition, autour de l'usage de la photo numérique dans les recherches en LSHS. Elle a en effet révélé qu'aujourd'hui, très nombreux sont les chercheurs à utiliser la photographie numérique comme un outil de la recherche (comme nous l'avions fait); que très peu d'entre eux ont reçu une formation spécifique; et que fort peu de chercheurs savent réellement optimiser leurs prises de vues, trier et archiver convenablement leurs photos, c'est-à-dire de manière rationnelle et surtout rentable pour l'utilisation future. Notre intention n'était nullement de réfléchir à l'épistémologie de cette pratique photographique – les chercheurs font ça très bien eux-mêmes – mais de penser cette pratique sur le plan de l'*outillage*, afin d'aider les chercheurs à l'améliorer, ce qui, espérions-nous, simplifierait les utilisations ultérieures de ces données iconographiques, notamment dans des procédures impliquant de l'ingénierie un peu rigide.

Faire de meilleures photos pour en faire moins; connaître le fonctionnement de son appareil, et l'existence des logiciels de traitement les plus simples; s'initier à quelques réflexes et bonnes pratiques de nommage, d'indexation et d'archivage; apprendre à automatiser des tâches qui se feront en quelques secondes sans saisie manuelle, voilà quelques-uns des services que nous avons souhaité rendre à la communauté des chercheurs, en nous appuyant sur notre propre expérience, mais aussi sur les témoignages issus de l'enquête. C'est l'objet d'un petit manuel que nous avons rédigé, et que nous publions par morceaux sur notre

blog (http://iconorezo.hypotheses.org/), non pas *La photo pour les* [*chercheurs*] *nuls*, pour paraphraser une collection populaire, mais précisément *La photo pour les chercheurs compétents qui ont besoin de faire de bonnes photos*, sachant qu'une «bonne» photo est une photo qui est exploitable dans le contexte disciplinaire de la recherche. Nous imaginons que ce partage de pratiques élémentaires aidera les chercheurs dans le quotidien de leurs différents terrains, et favorisera les usages ultérieurs de solutions numériques comme Omeka, parce que nous les aurons aidés à croiser rigueur de la nomenclature et efficacité dans le traitement de l'information, sans jamais les contraindre, mais à l'inverse en les aidant à gagner du temps.

Si la première étape de la rationalisation du travail se situe dans l'acte même de prendre une photo et de la transférer sur son ordinateur, la seconde intervient au moment de documenter cette photo: qu'est-ce qui a été photographié? Où, quand? À quel ensemble plus large se rattache cet objet, ce phénomène? Quelles informations de nature qualitative le chercheur souhaite-t-il ajouter à cette donnée brute qu'est la photographie, et qui ne figurent pas sur l'image (le nom de l'architecte qui a construit le bâtiment, par exemple)? Et surtout, où et comment stocker ces métadonnées: dans le nom même de chaque photo, dans un fichier de métadonnées (mais lequel?), dans un tableur, dans une base de données? Quelle solution fournira le meilleur rapport simplicité/réutilisabilité, pour le chercheur d'abord, pour une éventuelle équipe d'ingénierie ensuite? Ce que nous souhaitons apporter, c'est notre expérience dans ce domaine, en explicitant les choix que nous avons faits, et mettant en garde contre les erreurs que nous avons commises. Mais nous savons que là s'arrête le bricolage – et donc notre fonction – et qu'au-delà commence la compétence d'un professionnel qui n'appartient que rarement à nos équipes, le documentaliste. Il est sans doute indispensable de l'intégrer à un programme de recherche

qui nécessite de traiter des données sérielles; peut-être même de proposer ses compétences à tout chercheur en LSHS, au sein de son laboratoire ou d'une MSH: mais ces questions relèvent de la politique de la recherche, et non de l'épistémologie.

C'est à cette étape en effet qu'idéalement un chercheur devrait pouvoir s'appuyer sur les compétences de ce professionnel dont le métier est justement le tri et l'archivage rationnel des informations. Spécialiste de la gestion de l'information documentaire, c'est lui qui saura par exemple conseiller le chercheur dans la manière de comprendre le fonctionnement du *Dublin Core*, et surtout, de traduire en langage ordinaire «l'esprit» des champs. Dans quelle rubrique faut-il entrer le nom de l'illustrateur d'un livre: comme «*creator*» ou comme «*contributor*»? À quelle notion renvoie le champ «*is part of*» lorsqu'on est en train de réaliser la fiche d'identité d'un bâtiment patrimonial? Au-delà même de ces opérations simples de répertoire, le documentaliste, familier des démarches de recherche, saura aussi conseiller le chercheur quant à l'organisation même du corpus: faut-il créer une seule collection Omeka pour l'ensemble des données, ou scinder le corpus en plusieurs collections, et dans ce cas, selon quelle logique? Ces choix auront des répercussions importantes sur l'affichage et la recherche, à moyen terme. C'est encore au documentaliste qu'on devrait pouvoir demander, à terme, d'utiliser les ontologies pour proposer aux chercheurs des démarches d'organisation de l'information qui rendent compte de la richesse des objets étudiés, au lieu de les forcer à se plier à des formats standards qui tendent peu ou prou à en réduire la complexité, parfois jusqu'à la caricature, comme nous l'avons subi avec le *Dublin Core*. Mais là, encore, ce n'est plus tout à fait la compétence du chercheur…

Les chercheurs en activité aujourd'hui n'ont pour la plupart pas reçu de formation aux HN. Pour eux, les démarches intellectuelles directement liées à l'épistémologie de leurs champs disciplinaires priment sur toute autre considération. Nous appartenons à cette frange des chercheurs en SHS qui considèrent que les naissantes HN ne constituent pas forcément un nouveau paradigme de la recherche, mais pour l'instant surtout un faisceau d'outils – qui, comme tous les outils, notamment technologiques, auront leur temps, mais dont il serait absurde de ne pas se saisir s'ils peuvent aider le chercheur pour sa recherche, et pour la mise en valeur de ses résultats. Omeka aujourd'hui, un autre demain, s'avère être un bon outil. Mais nous sommes convaincus que c'est la main de l'artisan qui doit guider l'outil, et non l'inverse. Nous ne souhaitons ni nous laisser imposer par un outil logiciel un modèle d'organisation de la pensée qui nous détournerait de notre mission, nous empêchant de *penser* nos objets; ni nous livrer à de fastidieux apprentissages du maniement d'outils subtils, mais très compliqués. Nous voulons des outils de structuration et de partage des données de nos recherches qui soient aussi faciles à conduire que nos voitures personnelles, qui se plient à notre style de conduite comme aux terrains sur lesquels nous avons décidé de les mener, et qui nécessitent peut-être un permis, mais pas de leçons de mécanique. Que les outils des HN permettent aux chercheurs *digital naive* de rester des bricoleurs numériques, du moment qu'ils sont d'excellents anthropologues, géographes, historiens. Que les progrès des usages communautaires de ces nouveaux outils mènent à la création de ces «moulinettes» qui rendront les opérations indolores, de sorte que l'outil numérique leur permette de décupler leur force de travail, sans jamais alourdir leur charge.

Il ne s'agit pas d'opposer la rigueur du numérique à l'amateurisme foutraque du terrain, mais bel et bien de faire

dialoguer deux rigueurs: celle de l’outillage informatique, qui de fait impose sa structuration, gage d’efficacité mais poids au quotidien; et celle de l’épistémologie propre à chaque démarche de chercheur, et adaptée à chaque terrain, chaque problème. Ce n’est qu’au prix de cette médiation, de ce dialogue, que pourront naître, aux confins des humanités et de l’outillage numérique, de véritables HN. Il nous semble, quant à nous, que les «bricoleurs» sont l’avant-garde tâtonnante de ces nouvelles conquêtes de la recherche.

CHAPITRE 8

L’édition électronique de manuscrits modernes

FATIHA IDMHAND, CLAIRE RIFFARD ET RICHARD WALTER

Dans les années 1970, la critique génétique a bouleversé le champ des études littéraires en introduisant les manuscrits et les brouillons des écrivains dans l’interprétation critique de l’œuvre. Après que la «nouvelle critique» a révolutionné l’analyse des textes littéraires en leur appliquant les procédés de la linguistique structurale et que Roland Barthes a théorisé ces méthodologies dans des écrits de référence, l’approche dite «génétique» a proposé de dépasser cette forme d’analyse du texte relativement fermée en prospectant du côté de sa genèse. Si la génétique a hérité des avancées du structuralisme, elle s’est, toutefois, immédiatement située à contre-courant du modèle en vogue en remettant la figure de l’auteur au cœur de l’analyse et en décloisonnant l’étude critique du texte qu’elle situe dans un processus d’écriture. L’introduction de concepts novateurs comme ceux d’«avant-texte» (Bellemin-Noël, 1971) et de «dossier génétique», mais aussi la publication d’écrits et d’ouvrages de référence (Hay, 1989; Grésillon, 1994; De Biasi, 2000), ont permis aux généticiens de développer et d’asseoir une méthodologie aujourd’hui reconnue et diffusée dans le monde entier. À l’instar de la *variantistica* italienne, la génétique a une approche multidimensionnelle de l’écrit: elle tient compte de la chronologie des événements qui ont donné naissance à l’œuvre, du contexte dans lequel celle-ci a émergé et des moyens mis en œuvre par l’auteur pour exécuter son projet. L’étude génétique vise surtout à mettre en évidence ce processus de fabrication de l’œuvre. La singularité de cette

approche constitue l'un de ses obstacles, depuis des années, lorsque cette science se propose de publier les sources de sa recherche. Une édition rigoureuse de ce matériau, accompagné des interprétations scientifiques du chercheur, suppose la publication d'une masse documentaire très importante, bien supérieure au texte final et signifie la mise en place d'un protocole éditorial modulable selon l'écrivain, selon le corpus et l'hétérogénéité du dossier de genèse. Aussi, à moins de centrer la publication sur une unité réduite (un poème, un chapitre, un croquis), les contraintes de l'édition sur papier n'ont jamais permis la réalisation de telles publications: il semblerait donc que seules les technologies numériques soient en mesure de résoudre ce problème éditorial. Un tel changement de paradigme ne soulève-t-il pas de nouvelles questions?

Les enjeux de l'édition critique et génétique

Il faut avant tout insister sur un préalable non négligeable, celui de l'existence des sources. La conservation de traces du processus de création (archives, premiers jets, ébauches, brouillons manuscrits ou dactylographiés, etc.) est une condition nécessaire à l'analyse génétique; or s'il est vrai qu'en France, depuis la fin du XIXe siècle et le legs historique de Victor Hugo (testament du 31 août 1881) à la Bibliothèque nationale de France (BNF), les écrivains ont régulièrement légué leurs fonds aux bibliothèques, dans les autres pays d'Europe ou du monde, le legs ou le dépôt à l'institution publique restent inhabituels; il est plutôt courant que les manuscrits soient vendus à des privés ou à des bibliothèques, engendrant ainsi la dispersion des fonds. Parfois, ce sont des ayants droit privés qui se proposent d'assumer la conservation des fonds et qui agissent sur ceux-ci en fonction de leurs capacités financières, comme cela a été le cas des archives africaines et latino-américaines qui sont au cœur de notre expérience.

Le généticien qui travaille sur de telles archives dépend de la disponibilité de ces sources primaires et des aléas juridiques inhérents à l'histoire de ces patrimoines et des pays dans lesquels ils sont conservés. Lorsqu'on a levé l'ensemble de ces verrous privés ou institutionnels, le travail du généticien est d'abord celui d'un archiviste: il récolte, classe et traite les documents en vue de leur étude scientifique et de leur communication au public. À cette étape, et en raison de ces contextes, la numérisation revêt deux fonctions essentielles: elle permet une première préservation des archives, préalable au dépôt (physique) des originaux dans un lieu de conservation pérenne, et elle facilite l'accès des chercheurs aux données génétiques.

Lorsqu'il a pu accéder aux fonds, le généticien suit alors un protocole très précis: il parcourt la masse documentaire dans le but de reconstituer la chronologie de l'écriture puis organise cet ensemble en «phases génétiques» et en «états rédactionnels» ou «génétiques». Ce premier classement est capital car il permet de constituer le «dossier génétique», c'est-à-dire le dossier qui rassemble l'ensemble des documents utilisés par l'écrivain pour réaliser son œuvre, depuis les brouillons et les ébauches scripturaires (feuillets libres, dessins, carnets, etc.) jusqu'aux documents qu'il a consultés (on parlera de sources exogénétiques comme les journaux, les revues, les livres, etc.) et aux échanges qu'il a eus avec d'autres sur son projet (lettres, courriers, etc.). Ce dossier, qui est une production critique du chercheur, est une «part essentielle de la recherche, non seulement par les difficultés propres au déchiffrage et à la transcription des documents mais surtout par l'exigence d'un classement exhaustif qui suppose, en réalité, que le dossier publié ait été au préalable presque entièrement élucidé» (De Biasi, 2007): il est l'élément-clé de l'édition scientifique des brouillons.

L'édition génétique ne vise donc pas uniquement à établir un texte, elle ambitionne de présenter les étapes situées en

amont et rendues compréhensibles grâce au récit de genèse. Bien sûr, elle pourrait «légitimement se contenter [...] de reproduire en fac-similé le dossier génétique complet, en l'accompagnant d'un classement génétique des documents et de leurs sous-parties [...] et d'une représentation ou explication des processus génétiques sous-jacents» (D'Iorio, 2010) mais la richesse d'une telle édition génétique réside dans la plus-value apportée par l'éditeur scientifique qui révèle les chemins de traverse empruntés par l'écrivain pour construire son œuvre. Deux possibilités s'offrent à lui pour mettre en évidence cette genèse: une présentation qui permette de découvrir un moment précis de la genèse en ciblant, dans le dossier génétique, une étape ou une phase particulière (dans son épaisseur spatio-temporelle) ou la mise en lumière d'une succession de strates différentes parcourues dans leur chronologie. Pour distinguer ces deux approches, consubstantielles en réalité, Pierre-Marc de Biasi parle d'horizontalité et de verticalité; pour lui, l'édition doit tenir compte de ces «mécanismes génétiques inséparables de leur inscription temporelle dans une durée complexe et de leur déploiement dans le volume qu'occupent concrètement les documents» (2007). Les notions d'éditions horizontale et verticale ne sont peut-être pas appropriées pour refléter les possibilités de l'édition électronique, elles traduisent néanmoins les deux dimensions dans lesquelles s'inscrit l'édition génétique des manuscrits. Quels sont donc, à ce stade, les enjeux de l'édition numérique pour notre domaine?

Soulignons d'abord que le basculement technologique ne modifie aucunement certains préalables de la recherche en génétique, celle-ci requiert toujours la même approche méthodologique: d'abord le classement du corpus et l'organisation des avant-textes dans l'ordre chronologique de leur apparition et selon les différentes phases de la genèse du projet, ensuite, la constitution du dossier

génétique qui regroupe tous les documents qui ont eu, selon les phases, une incidence sur le processus de création. Ce dossier, qui rassemble les différents états génétiques de l'œuvre et toutes les sources de l'«exogenèse», suppose un traitement singulier en raison de son caractère multidimensionnel (et exponentiel). D'abord parce qu'en fonction de l'écrivain étudié, de son époque et de l'évolution des pratiques scripturaires, la nature des objets qui le constituent se diversifie avec le temps: manuscrits, copies dactylographiées et photos imprimées s'enrichissent, au fil des années, de nouveaux objets et de nouvelles ressources. Témoins des différentes révolutions technologiques que nous avons connues, les supports évoluent (cassettes, CDs, DVDs, disquettes, clés USB, disques durs, etc.), bouleversent le profil matériel des documents du dossier de genèse et induisent des traitements (numérisation, archivage, etc.) qui engendrent à leur tour de nouvelles réflexions épistémologiques sur les conditions de sauvegarde de ces sources, notamment lorsqu'elles sont nativement numériques. On pourrait d'ailleurs penser que l'édition électronique est en mesure d'intégrer, plus facilement que le papier, les éléments du dossier de genèse nativement numériques qui ne faisaient pas partie de l'édition papier, et pour cause! Pourtant, le processus n'est pas si simple.

En effet, si la transition massive vers le numérique semble inéluctable, la perte de données l'est également si on ne réalise aucune structuration de l'information à archiver et si on n'anticipe pas le respect des normes et des standards, des méthodes et des coutumes, actuelles et futures. Le passage au numérique exige par conséquent une préparation en amont très fine (mesure des répercussions de l'édition, choix des outils et des méthodes, etc.) car chaque option peut avoir des conséquences irréversibles. L'expertise du généticien doit aujourd'hui intégrer les nouveaux horizons de cet environnement numérique avec,

forcément, de nouveaux enjeux. Il semble ainsi évident, par exemple, que les possibilités de visualisation des processus génétiques fins (diverses campagnes d'écriture sur un état, rapprochement de différents états du même texte, relations avec des éléments d'exogenèse) sont démultipliées. Toutefois, il faut souligner que le basculement technologique ne modifie ni les étapes de la recherche génétique ni les principes éditoriaux d'une publication scientifique; on ne sera donc pas étonné d'apprendre que la publication du dossier génétique reste l'enjeu principal de l'édition électronique et que, comme toute édition de ce type, elle répond préalablement aux règles scientifiques du domaine de recherche et à des règles éditoriales propres au numérique, à des normes et des standards destinés à faciliter l'échange d'informations d'une part, mais également la conservation à long terme de celles-ci.

Numériser et éditer un manuscrit, dans notre domaine, suppose par conséquent la prise en compte des différentes dimensions de l'objet (documentaire, archivistique, génétique), dont le contenu comporte des informations de multiples natures qui devront être structurées (usage de descripteurs, de normes et de règles différentes). À chaque étape du projet, le chercheur fait face à des choix de différentes natures selon les niveaux de l'édition qu'il envisage: d'un côté l'édition scientifique, de l'autre les contraintes techniques; d'un côté l'interopérabilité nécessaire en vue de la mutualisation, de l'autre des choix technologiques spécifiques, en fonction de besoins propres. L'édition génétique numérique doit obéir à ces exigences, avec la même qualité que l'édition imprimée et en intégrant des impératifs techniques supplémentaires comme le respect des normes du domaine numérique.

e-Man: pour l'édition de textes et de manuscrits modernes?

Face à la masse croissante de contenus numériques produits, à la variabilité et aux contraintes du virage vers le *tout numérique*, les problèmes posés par la gestion de telles sources sont, indéniablement, plus importants qu'autrefois et il n'existe, pour l'instant, aucune méthode ni solution clés en main. Nous avons fait le choix d'une nouvelle approche: adapter un outil générique à nos besoins pour tenter de répondre aux règles de la discipline scientifique et à celles du numérique. D'autres projets avaient déjà tenté, avant nous, de trouver des solutions techniques pour permettre à tous d'accéder aux sources des œuvres dans le but de rendre compte de leur processus de création. Nous pouvons citer parmi ceux-ci celui de Julie André et Elena Pierazzo, fondé sur une visualisation des manuscrits de Marcel Proust à partir d'un codage en Text Encoding Initiative (TEI) mais qui n'a donné lieu qu'à un prototype (http://research.cch.kcl.ac.uk/proust_prototype). Le projet d'édition des avant-textes de *Paisajes después de la batalla* de l'écrivain Juan Goytisolo, mené par Bénédicte Vauthier (http://goytisolo.unibe.ch), a abouti à une solution convaincante mais difficile à transposer pour qui souhaite éditer une grande quantité de manuscrits en raison, comme pour l'édition des Cahiers de Proust, du temps considérable que requiert l'encodage en TEI. Le projet d'édition des dossiers de *Bouvard et Pécuchet* de Flaubert, mené par Stéphanie Dord-Crouslé (http://www.dossiers-flaubert.fr) ou des journaux de Stendhal, mené par Cécile Meynard et Thomas Lebarbé avaient aussi abouti à des solutions relativement transposables; pour le premier, l'encodage en TEI a constitué le pivot, pour le second, on a mené l'édition à l'aide d'une plate-forme, CLELIA. Ensuite, un système a été mis en place pour permettre la transposition de l'édition vers la TEI. Tous ces projets, basés sur des solutions *Open Source* (développements et langages, formats, etc.), contrastent avec d'autres, comme le Samuel Beckett Digital Project, dont la solution éditoriale est difficile à consulter, et

à transposer, en raison de la fermeture du site, avec accès payant (http://www.beckettarchive.org/).

Le projet e-Man (plate-forme pour l'édition de manuscrits modernes) avait pour but le développement d'une solution ouverte, simple, fondée sur un outil de publication interopérable favorisant l'exploitation numérique de documents. Il s'agissait pour nous de conjuguer, ainsi que le suppose l'édition génétique, la publication en ligne d'images, de textes et de diverses sources numériques dans le respect d'un état de l'art éditorial et avec les exigences d'une édition critique enrichie des résultats de recherches scientifiques sur les processus de création des œuvres. Pour élaborer cet outil, nous avons fait le choix d'une volumétrie haute en envisageant le traitement de volumes d'archives importants (des dizaines de milliers de feuillets). Nous nous sommes appuyés sur des manuscrits et des archives dont les droits et les autorisations de reproduction et de diffusion ont été négociés avec des propriétaires qui ont accepté la numérisation et la diffusion des sources sous licences *Creative Commons*, à des fins de recherche scientifique. Il s'agit (pour l'instant) de fonds francophones (Jean-Joseph Rabearivelo et Mouloud Feraoun) et hispaniques (José Mora Guarnido, Carlos Denis Molina, Carlos Liscano) dont la genèse s'étend entre la fin du XIXe siècle et le début du XXIe siècle; ils sont actuellement conservés soit chez leurs propriétaires ou ayants droit (en Amérique latine, en Algérie), soit dans des bibliothèques (Service commun de la documentation Lille 3, Biblioteca Nacional de l'Uruguay), soit à l'Institut français d'Antananarivo (Madagascar). Ces archives n'ont pas été traitées par des institutions nationales comme la BNF ou l'Institut mémoires de l'édition contemporaine (IMEC). Notre projet illustre en réalité la plupart des cas scientifiques: ceux de chercheurs qui travaillent sur des corpus encore situés hors des bibliothèques ou des académies. Ils doivent, en quelque sorte, prouver le potentiel de ces documents de façon qu'ils

puissent être préservés un jour par ces institutions. C'est donc pour cette raison que des questions d'ordre archivistique et documentaliste se sont posées à nos équipes.

Après avoir comparé et évalué les différentes solutions de publication disponibles, nous avons décidé de développer, à partir d'un outil existant, une plate-forme générique et modulaire avec trois ambitions principales: 1) éditer les manuscrits numérisés avec des métadonnées qui respectent les normes et les standards internationaux du World Wide Web Consortium (W3C) (http://www.w3.org); 2) associer le manuscrit numérisé à sa transcription enrichie; 3) associer, au sein de la plate-forme, les documents entre eux grâce à des liens logiques, temporels et génétiques pour constituer les dossiers génétiques.

L'édition électronique génétique conjugue ainsi la publication d'images (édition de fac-similés) et de textes (édition de textes transcrits et d'analyses scientifiques) avec d'autres types de ressources, numériques ou numérisées, présentées dans un ordre significatif: celui du processus de création. La difficulté d'une telle publication réside dans le caractère multidimensionnel des dossiers génétiques qui demandent des traitements complexes. Ce sont ces réflexions qui nous ont amenés à envisager la mise en place d'un modèle éditorial modélisable, adaptable à différents types de corpus et d'une plate-forme destinée à favoriser l'édition génétique numérique: **e-Man**, une plate-forme d'**é**dition de **man**uscrits modernes (http://eman-archives.org).

Le logiciel Omeka

Nous avons donc choisi d'adapter un outil *Open Source* à nos besoins: Omeka. Ce système de publication Web spécialisé dans l'édition de collections muséales, de bibliothèques numériques et d'éditions savantes en ligne se situe à la croisée du système de gestion de contenus (CMS),

de la gestion de collections et de l'édition d'archives numériques; il a été développé par une équipe de chercheurs du Roy Rosenzweig Center for History and New Media (RRCHNM) de la George Mason University (Virginie, États-Unis) dans le but de faciliter l'édition en ligne d'images. Les créateurs de cet outil ont conçu un logiciel facile à prendre en main de façon à permettre aux utilisateurs, selon le site d'Omeka, de «se concentrer sur le contenu et sur l'interprétation plutôt que sur la programmation» (http://omeka.org/about/). Omeka permet de publier des contenus de façon simple et flexible, son architecture souple est adaptable aux besoins du projet grâce à l'ajout de modules (*plugins*) et à l'utilisation de modèles visuels (*templates*). En revanche, le logiciel impose un schéma pour la structuration des métadonnées, celui du *Dublin Core*.

Le *Dublin Core* a été créé en 1995 pour décrire des ressources de façon simple et compréhensible, afin de les diffuser et de les échanger facilement. Ses métadonnées sont compatibles avec les autres recommandations internationales de description des archives comme l'International Standard Archival Description-General (ISAD(G)) dont la DTD EAD (*Document Type Definition – Encoded Archival Description*) est la traduction en XML. Le fait que le *Dublin Core* soit imposé par Omeka peut paraître réducteur pour décrire des documents aussi complexes que des manuscrits d'écrivains, mais en réalité Omeka n'impose que quelques champs du *Dublin Core* (quinze), ensuite, l'utilisateur peut adjoindre aux contenus qu'il édite des métadonnées personnalisées. Celles-ci sont déclinables à l'infini et permettent d'enrichir la description en fonction des besoins du projet, en associant à chacune des unités documentaires publiées des informations descriptives, administratives et structurelles.

Une typologie applicable pour les manuscrits

Ainsi, pour publier nos manuscrits numérisés, nous avons choisi de les décrire à l'aide des quinze champs du *Dublin Core* et de compléter cette description par des éléments personnalisés qui correspondent à la fois aux caractéristiques de nos fonds (poésie, théâtre, référents spatiaux…) et aux intérêts scientifiques repérés et décrits par les chercheurs spécialistes. Notre description des manuscrits s'est appuyée sur le plan proposé dans la recommandation DeMArch (2010). Ce plan, qui intègre la notion de dossier génétique (cette approche est de plus en plus présente dans les milieux archivistiques et dans les bibliothèques qui gèrent des fonds d'écrivains contemporains), nous a permis d'établir une typologie respectueuse des recommandations en vigueur et qui peut potentiellement s'appliquer à toute étude de manuscrits modernes et contemporains. Dans cette utilisation d'Omeka, nous avions donc fait le choix d'associer l'encodage des informations en *Dublin Core* à des informations encodées, dans des champs personnalisés, en XML.

Une approche par dossier génétique

Omeka permet par ailleurs de mettre en relation des documents entre eux grâce à des liens spécifiés: en élaborant la plate-forme e-Man, nous avons fouillé cette fonction de façon à permettre au lecteur de visualiser chaque manuscrit, au sein de son dossier génétique, en spécifiant la complexité de ses relations. Pour ce faire, on a associé chaque manuscrit – document – à l'ensemble des éléments liés à sa genèse (avant-textes, ébauches, documents préparatoires et sources exogénétiques). Cette «relation» a été établie à partir d'une extension d'Omeka, *Item Relations* qui permet d'associer des contenus entre eux selon le type de lien que le chercheur spécifie: «fait partie de», «est avant-texte de», etc. Cette extension ne construit toutefois que des relations univoques, qui fonctionnent dans

les deux sens. Or pour situer le document dans la relation chronologique qu'il entretient avec les autres documents à la fois au sein du même dossier génétique ou d'un autre, la solution que nous avons mise en place a consisté à modifier le fonctionnement de l'extension de façon que la relation exprimée ne soit effective que dans un seul sens et, surtout, que la relation chronologique soit indiquée précisément. Pour cela, il a fallu modifier légèrement le code de façon à visualiser la relation qui unit chaque document à d'autres lorsque celui-ci fait partie d'un dossier génétique. Pour aller plus loin, nous nous sommes proposé d'explorer d'autres pistes de travail, notamment celles qui concernent la façon de distinguer, au sein du dossier de genèse, les différents avant-textes et surtout, d'afficher les liens génétiques dans leur complexité à la fois diachronique et synchronique. Est-il possible de poursuivre un tel développement *ad hoc* dans le cadre de l'utilisation d'un outil comme Omeka? Dans notre utilisation novatrice du logiciel, nous tentons de modifier, le moins possible, son code informatique. La poursuite de l'expérience déterminera la validité du procédé mais à ce stade, nous pouvons déjà dire que le logiciel a comblé la plupart des besoins éditoriaux, et partiellement permis de répondre aux enjeux des questions de genèse.

Les autres enjeux d'Omeka et au-delà

Lors de notre expérience, nous avons testé d'autres *plugins* d'Omeka: la géolocalisation, la visualisation de corpus par mots-clés et ensembles de *tags*... mais sans résultats significatifs. Omeka présente par ailleurs l'avantage de constituer un entrepôt OAI-PMH (*Open Archives Initiative – Protocol for Metadata Harvesting*, une fonctionnalité que l'utilisation de métadonnées structurées en *Dublin Core* simplifie) destiné à faciliter la présentation de nos données auprès des grands «moissonneurs», les tests menés sur ces entrepôts ont été, en revanche, assez prometteurs. L'outil a quelques limites que nous avons tenté de repousser afin de

continuer à l'adapter selon nos besoins; une nouvelle fois, c'est par un subterfuge que nous sommes parvenus à associer à l'image du manuscrit numérisé la transcription: pour cela, nous avons utilisé l'un des champs de métadonnées de l'image. En effet, dans notre approche, la collection concerne l'auteur de l'œuvre, à chaque auteur d'œuvre, nous avons associé des contenus qui sont les œuvres et chacune de ces œuvres comporte à son tour des images, qui sont en réalité les pages numérisées. Chaque contenu est décrit avec des métadonnées associant du *Dublin Core* et du XML et nous pourrions encore enrichir la description du contenu en remplissant, pour chaque image, les champs des métadonnées. C'est donc dans les métadonnées des images que nous avons trouvé un champ permettant d'afficher, à côté de l'image, la transcription: le champ «description». En effet, le projet inclut la possibilité de consulter la transcription de certains manuscrits, notamment ceux qui sont les plus difficiles à déchiffrer. Une fenêtre HTML autorise la transcription directement dans l'outil. L'une des possibilités évoquées à ce stade du projet a concerné la capacité du logiciel à intégrer un langage XML structurant comme celui de la *Text Encoding Initiative* (TEI), par exemple.

Le Consortium TEI a développé le langage XML-TEI dans le but, d'une part, de structurer les métadonnées liées au texte et, de l'autre, d'enrichir le texte publié avec des informations scientifiques. Depuis les travaux du Special Interest Group dédié à la génétique et coordonné par Elena Pierazzo, Malte Rehbein et Amanda Galley, la TEI s'est enrichie de balises permettant d'encoder des opérations génétiques importantes comme les ratures, les suppressions, les substitutions ou les déplacements (voir l'expérience menée sur le cahier n°46 de Marcel Proust: http://research.cch.kcl.ac.uk/proust_prototype/). L'encodage des transcriptions des textes en TEI paraît évident lorsqu'il s'agit d'enrichir l'édition électronique des textes dans la

mesure où l'immense bibliothèque de balises proposées par la TEI complète l'édition scientifique électronique.

Dans un premier temps, nous avons testé l'extension *TEIDisplay*, pour décider, rapidement, de l'abandonner car elle n'était plus maintenue dans les versions les plus récentes d'Omeka. Ensuite, nous avons recherché d'autres solutions pour l'encodage XML-TEI de nos métadonnées et de la transcription, toujours au sein de la plate-forme e-Man. Nos contraintes de départ étaient multiples: d'abord, nous souhaitions éviter de recourir à un logiciel d'encodage payant en raison du coût d'un tel outil, de sa difficile prise en main et du fait que les personnes chargées des transcriptions ne sont pas des encodeurs professionnels et qu'ils ont souvent du mal à se familiariser avec de tels dispositifs. Ensuite, nous souhaitions que les deux actions, transcriptions et encodage, puissent être réalisées sur une unique plate-forme afin d'éviter une double manipulation informatique. Cela signifiait donc d'utiliser un outil intégrable au sein d'Omeka; or le logiciel n'a pas prévu la saisie du texte avec des balises spécifiques. Omeka propose certes une extension, *Scripto*, qui donne à première vue l'impression de centraliser les interfaces de transcription et de saisie des notices. Nos premières utilisations de *Scripto* nous ont posé quelques problèmes pour travailler avec nos équipes, notamment le passage par la gestion de contenus via MediaWiki qui compliquait le travail de validation de la transcription (par exemple, il ne permet pas de supprimer la dernière modification effectuée) et obligeait à travailler hors de l'environnement d'Omeka. Nous avons pensé qu'il serait possible d'intégrer à la plate-forme une barre d'outils permettant d'encoder en XML-TEI, d'ajouter à la barre HTML des fenêtres de saisie (notamment dans le champ «description» des images), des boutons additionnels destinés à coder, plus précisément, des phénomènes génétiques (sur ce point nous nous sommes inspirés du projet Transcribe Bentham). Cette démarche vise à simplifier

la tâche du transcripteur qui, comme avec un traitement de texte, n'aurait plus qu'à sélectionner le texte ou l'élément qu'il veut encoder et à cliquer sur l'icône correspondant à l'action pour entourer l'élément des balises adéquates. Il ne serait ralenti ni par l'apprentissage du langage informatique ni par le risque d'erreur dans la saisie manuelle des balises. Nous n'avons toujours pas testé cette option à ce jour. Sa faisabilité dépend, une fois de plus, du degré d'intervention sur le code informatique et des développements que nous pourrions être en mesure, non pas de réaliser, mais de maintenir, dans la durée notamment à l'occasion de mises à jour. En 2016, l'équipe d'Omeka avait annoncé la diffusion d'une nouvelle version du logiciel.

En guise de conclusion provisoire, revenons d'abord sur les objectifs de notre expérience. Nous voulions initialement éviter la création et le financement d'un outil *ad hoc* (qui aurait demandé des développements et une maintenance importants) par l'exploration des potentialités d'un outil générique, robuste, plébiscité par une large communauté et maintenu par une équipe de recherche. Pour cela, nous avons pensé que la création de fonctions spécifiques en lien avec les besoins précis de l'édition numérique génétique permettrait à la fois d'améliorer l'outil, mais également de répondre à des besoins que les créateurs d'Omeka n'avaient pas envisagés. Or de telles interventions, que nous avions initialement jugées mineures, finissent par rendre l'outil générique… non générique! Certes, il est vrai que nos propositions de développement pourront être soumises dans l'avenir à la communauté internationale Omeka, sous forme de propositions d'extensions, afin qu'elles bénéficient au plus grand nombre et que le partage de cette expérience, et des compétences acquises, sera l'une des voies de la pérennisation des solutions par d'autres projets comparables. Mais la question qui demeure en suspens reste celle de la relation du projet avec l'outil et d'une forme de réflexe des sciences des humanités à n'opter que pour

une seule et unique solution technique/numérique. Face aux nouvelles questions épistémologiques soulevées par l'expérience, nous avons donc repris le temps de la réflexion.

Ainsi, s'il est vrai que le numérique a contribué à lever certains des verrous financiers, matériels, techniques et quantitatifs des généticiens éditeurs en leur ouvrant les possibilités de la publication des textes, des images et de tous les médias de natures différentes (sources sonores, vidéos, iconographiques, etc.) qui peuvent composer le dossier de genèse d'une œuvre, il a également bouleversé l'organisation du projet scientifique avec le passage de l'édition des sources de la recherche depuis la fin du projet vers son amont. L'édition accompagnerait dorénavant le projet dans toute sa durée. Or la question est-elle vraiment éditoriale? Le renversement de la démarche scientifique n'est-il pas moins lié à la question de l'édition, de ses moyens, de son objectif et de sa qualité, qu'à celle de l'objet d'étude lui-même? En effet, nombreux sont les (nos) travaux qui ont signalé que la transformation/conversion numérique de l'objet d'étude génère de nouvelles questions, requiert de nouvelles approches, d'autres méthodes; or avons-nous toujours affaire au même objet une fois celui-ci devenu donnée informatique?

Le numérique a contribué à dégager l'horizon de travail d'un certain nombre de verrous quantitatifs qui ont longtemps freiné les ambitions, pour le domaine qui nous concerne, de l'édition génétique des œuvres, comme la limitation du nombre de pages, d'images et le rapprochement d'éléments issus de médias différents (sources sonores, vidéos, iconographiques, etc.). De plus, grâce à l'édition numérique, on peut enfin envisager un rapprochement heureux avec l'édition savante traditionnelle (philologique). Grâce à ces technologies, il est possible d'éditer l'un des éléments-clés de l'analyse génétique: le dossier génétique de l'œuvre, un élément crucial du travail

critique du chercheur, en représentant les différents niveaux de relation qui existent entre les documents, sans limite, c’est-à-dire, en donnant à voir l’interprétation des phénomènes, résultat de l’herméneutique scientifique. Mais tout cela est-il possible avec un seul outil? Notre expérience de développement de la plate-forme e-Man a tenté de répondre à cette question.

TROISIÈME PARTIE

LA GESTION DE PROJET

CHAPITRE 9

Le projet DFIH: humanités numériques et histoire financière

JÉRÉMY DUCROS, ELISA GRANDI,
PIERRE-CYRILLE HAUTCŒUR, RAPHAËL HEKIMIAN,
ANGELO RIVA ET STEFANO UNGARO

La disponibilité de bases de données vastes, harmonisées, de qualité et pérennes ouvre de nouvelles perspectives de recherche en sciences humaines et sociales (SHS). La collaboration entre les SHS et les sciences et technologies de l'information et de la communication (STIC) crée les conditions pour la réalisation d'une *big data revolution*.

L'équipement d'excellence «Données financières historiques» (EQUIPEX DFIH) se propose de construire une base de données contenant les informations relatives aux actifs cotés à la Bourse de Paris (titres, taux de change, matières précieuses) et à leurs émetteurs de 1795 à 1976. L'équipement contient les prix bimensuels (comptant, terme, reports et options), les dividendes et les autres informations (comme les opérations sur titres) nécessaires à l'établissement de séries de prix homogènes dans le temps. Il recense en outre l'historique juridique de la société (date de fondation et évolution du statut juridique), les principales règles de gouvernance, le siège, les administrateurs, des bilans simplifiés, etc. ainsi que des informations sur les émetteurs publics (par exemple la situation budgétaire et les garanties éventuelles). Les données seront mises gratuitement à la disposition des chercheurs en 2020 via la très grande infrastructure de recherche (TGIR) PROGEDO.

Pourtant, l’ambition de ce projet est plus large. L’équipement DFIH est une base de données «ouverte». Il vise initialement à inclure les autres bourses françaises, mais pendant la phase de conception du modèle de données, tous les efforts ont été faits pour prévoir des évolutions possibles de la base. La flexibilité des bases de données relationnelles permet d’espérer que des évolutions non anticipées (par exemple l’intégration de données prosopographiques sur les salariés d’une entreprise) puissent trouver leur place au sein de l’équipement.

La construction de la base se fonde sur deux sources sérielles principales:

- les cotes officielles de la Bourse de Paris publiant les informations sur les actifs;
- les annuaires boursiers officiels et privés publiant les renseignements sur les émetteurs.

La stratégie de capture des données dépend de plusieurs variables principalement liées aux sources et à l’état de l’art de la technologie. L’accessibilité des sources, la qualité des imprimés, l’organisation graphique des informations et la fréquence des changements de format (donc le type d’information contenue) sont les éléments à prendre en compte. D’autre part, c’est principalement l’état de l’art de la technologie qui détermine les coûts relatifs des différentes solutions de capture possibles. Il faut arbitrer entre deux types de solutions: soit le recours à des techniques éprouvées, mais caractérisées par un «coût unitaire» élevé; soit des investissements en faveur d’innovations qui soient susceptibles de repousser la frontière technologique afin de pouvoir réduire le prix de revient des données.

Dans le cadre de l’EQUIPEX, les contraintes et les opportunités rencontrées ont conduit à mettre en place

deux technologies de capture de données requérant la numérisation des sources. Les sources numérisées constituent un bien public *in se* et un exceptionnel instrument de documentation de la base. Ces sources ont par conséquent vocation à être mises en ligne via une infrastructure comme la TGIR Huma-Num. Les deux technologies adoptées sont les suivantes: d'une part, la saisie manuelle dans un environnement numérique *ad hoc* des informations publiées sur les cotes boursières jusqu'en 1950 et d'autre part, le traitement semi-automatique des cotes d'après 1950 et des annuaires par un logiciel spécifique fondé sur la reconnaissance optique des caractères (ROC) et sur l'intelligence artificielle. La nécessité de développer un logiciel spécifique a été dictée par trois ordres de besoins: d'abord, l'échec des meilleurs logiciels commerciaux de reconnaissance optique; ensuite, la nécessité de disposer d'un logiciel capable d'apprendre de ses erreurs; enfin, le recours au traitement semi-automatique des données pour les ventiler dans la base de données.

L'EQUIPEX a réalisé ce logiciel en partenariat avec un prestataire privé qui proposait par ailleurs une infrastructure robuste et efficace d'analyse et de traitement des données issues de la ROC, deux étapes nécessaires pour la production d'une quantité importante de données. En prenant en compte les coûts directs et indirects, ce logiciel offre des gains par rapport à la saisie manuelle dans un environnement numérique. Pourtant, son fonctionnement demande encore un important travail humain avant et pendant son déploiement. Dans un premier temps, il a été nécessaire de créer, au sein de la base, une structure informationnelle permettant au logiciel de ventiler correctement les données lors de leur versement. Par la suite, le flux des données issues de la ROC doit être validé par des opérateurs. Or, le travail humain requis pour la constitution de bases de données à partir de sources

anciennes, et notamment tabulaires, demeure un verrou qui rend la production coûteuse. La recherche qui s'appuie sur les humanités numériques (HN) comme terrain de rencontre entre SHS et STIC est ainsi appelée à œuvrer pour réaliser la *big data revolution*, à savoir le développement d'outils conçus pour produire, à partir de sources historiques, des masses importantes de données à un coût raisonnable.

La construction de l'architecture de la base de données

Organisée suivant le modèle relationnel, la base de données, gérée par le système Oracle, consiste en un ensemble de tables, reliées entre elles par des relations et correspondant chacune à une unité de stockage. À ce jour, la base de données regroupe près de 150 tables. L'information y est stockée en ligne (on parle alors de *tuple*) et en colonne, appelée attribut. Une relation entre deux tables est définie à partir d'une clé primaire, c'est-à-dire à partir d'un attribut ou d'une combinaison d'attributs. La longue période envisagée ici nécessite de recourir à une base de données non pas statique mais dynamique où toutes les informations sont liées au temps. Chaque information est donc définie et valide à l'intérieur d'une période de temps donnée, avec une date de début et une date de fin. Les relations entre tables ne sont pas toutes gérées de la même manière. Dans certains cas, la clé primaire de l'une des tables est simplement ajoutée dans l'autre. On appelle alors cette clé, *clé étrangère*. Dans d'autres cas, une table de jointure est créée avec les deux clefs primaires des tables à relier.

Une fois les principales entités définies, plusieurs contraintes fondamentales ont guidé l'élaboration du modèle conceptuel des données.

- La première était de conserver les caractéristiques des sources historiques, afin d'en rendre compte le plus fidèlement possible et de suivre leur évolution au cours des deux siècles;
- La deuxième était de permettre aux utilisateurs d'exploiter la base de façon efficace et d'effectuer facilement des requêtes SQL (*Structured Query Language*) couvrant la période étudiée;
- Il était également important que la base de données finale puisse être compatible avec d'autres bases de données déjà existantes ou futures, notamment en Europe, par exemple les bases SCOB (*Studiecentrum voor Onderneming en Beurs*) en Belgique ou Eurofidai en France. En effet, la fusion de ces bases de données est envisagée à terme;
- Enfin, il a fallu adapter la base de données aux contraintes découlant des techniques de capture des données.

Une fois le modèle conceptuel de la base défini, il a été possible de concevoir le modèle logique des données. Pour cela, nous avons eu recours à l'expertise du centre de recherches SCOB à Anvers qui maintient une base de données contenant les données historiques des Bourses de Bruxelles et d'Anvers. L'une des tâches les plus importantes a alors été d'adapter la base SCOB au cas spécifique français et aux contraintes de production des données. Ainsi, avant de commencer à importer des données dans la base, l'ensemble des possibilités rencontrées dans les sources historiques françaises devait être envisagé afin de ne pas être bloqué ultérieurement. Par exemple, l'une des spécificités de la base de données SCOB est de ne prendre en compte que les prix au comptant et de laisser hors du champ de la base les informations concernant les opérations à terme. Étant donné l'ampleur et l'importance de ces opérations dans le cas français – même pour un

marché régional comme la Bourse des valeurs mobilières de Lyon, les opérations à terme représentent jusqu'à 90% du volume total des transactions au cours de la période traitée –, nous avons décidé de les inclure dans la base de données française. Sans entrer dans le détail de ces opérations financières, signalons simplement qu'elles peuvent être de deux types, fermes ou à option, et être effectuées à plusieurs échéances, ce qui engendre par conséquent un grand nombre de cas possibles. Nous avons donc opté pour la création de nouvelles tables et de nouvelles relations pour pouvoir en tenir compte. La base de données comporte également des tables qui facilitent le processus de production de données. C'est le cas par exemple de celles créées dans le but de permettre une double saisie.

Afin de minimiser le coût, deux technologies de capture de données ont été mises en place. D'une part, la saisie manuelle dans un environnement numérique *ad hoc* construit sur la base de masques de saisie; d'autre part, le traitement des sources par un logiciel spécifique.

La saisie manuelle dans un environnement numérique *ad hoc*

Schématiquement, elle a été impérative pour:

- les données nécessaires à la construction de la structure informationnelle (types d'objets manipulés, listes fermées de secteurs et de valeurs, relations entre objets, notamment entre les titres et leurs émetteurs);
- les données publiées sur les cotes officielles jusqu'en 1950, dont la qualité et l'organisation fluctuantes rendent l'automatisation inefficace ou non rentable.

La saisie des données et l'établissement de la structure informationnelle demandent un niveau élevé d'interprétation, donc d'expertise et de connaissance des

sources. Il s’agit notamment de créer le système permettant à la base de reproduire, pour chaque jour de cotation, la cote officielle telle qu’elle figure sur la source. La reproduction de la cote officielle, aussi banale qu’elle puisse paraître, est en réalité une opération complexe. Elle demande non seulement de suivre les titres dans le temps malgré les changements des noms, des secteurs de cotations (dont le nombre et l’intitulé varient fréquemment) et des caractères techniques, mais aussi de relier le titre au bon émetteur en dépit d’évolutions parfois radicales, comme la disparition d’un État, la fusion entre deux entreprises ou encore le rachat de l’une par l’autre. La mise en place de masques de saisie avancés facilite la saisie de ces informations qu’on a confiée à des opérateurs experts comme des doctorants et des postdoctorants. Ces masques, codés en Java, permettent à l’opérateur expert d’interagir facilement avec la base.

En revanche, une fois la structure informationnelle achevée, la saisie des données publiées sur les cotes requiert un niveau d’interprétation moindre. Des assistants de recherche peuvent l’organiser et la réaliser en interne ou la déléguer à une entreprise spécialisée dans la saisie de données. Les contraintes relatives à l’espace nécessaire pour l’installation et l’encadrement de nombreux opérateurs de saisie nous ont conduit à déléguer le travail à une entreprise spécialisée. La collaboration avec cette entreprise a commencé par une adaptation du modèle logique des données, afin de prendre en compte les spécificités de leurs procédés de production, dont la plus importante est la double saisie. Afin de minimiser les erreurs de saisie, deux opérateurs saisissent en parallèle les mêmes données. En cas de divergence entre ces deux saisies, un troisième opérateur tranche. Pour assurer une productivité satisfaisante, on a conçu des masques de saisie simplifiés. On a spécialisé les équipes d’opérateurs par type de données (prix au comptant, prix à terme, dividendes), et

affecté à chaque type un masque de saisie ne reportant que les champs utiles. Par exemple, les taxes payées par les investisseurs sur les intérêts et les dividendes apparaissent sur la cote officielle au cours de l'entre-deux-guerres. En conséquence, on a créé un masque ne reportant que le champ «dividende brut» pour la période antérieure à 1920 tandis qu'un masque comportant la distinction entre «dividende brut», «dividende net» et «taxe» a été implémenté pour la période suivante.

L'opérateur pouvait ainsi sélectionner, au sein du masque, le titre auquel associer les données saisies, en comparant l'image de la cote insérée directement dans le masque et l'arborescence graphique traduisant la structure informationnelle de la base. Cette possibilité, jointe à la spécialisation des équipes d'opérateurs, réduit le niveau d'interprétation des sources nécessaire à la saisie, mais elle ne l'élimine pas. C'est pourquoi les opérateurs ont bénéficié d'une formation de trois semaines par des opérateurs experts (le superviseur scientifique du projet et un doctorant en histoire financière) et pouvaient s'appuyer sur un manuel de guide à la saisie et des tutoriels vidéo. Ces derniers présentent les règles générales de saisie. Le manuel détaille, quant à lui, les cas particuliers, très nombreux, que les opérateurs peuvent retrouver sur la cote. Il s'agit par exemple de cas rares, d'exceptions, ou de données susceptibles d'être interprétées différemment selon les époques ou selon la partie du texte où elles sont publiées. Ainsi, un prix de «20» peut être lu comme 20 centimes, 20 francs ou encore 20%, selon la période de publication et le type de titre échangé. Malgré tous les efforts accomplis, l'entreprise spécialisée avait parfois besoin d'une assistance dans l'interprétation des données. Pour cette raison, nous avons mis en place une interface en ligne de questions/réponses permettant de charger des images ou toute autre documentation nécessaire pour accompagner les réponses aux questions.

Une fois les données importées, l'entreprise les a envoyées à une équipe chargée d'en vérifier la qualité. Cette vérification était à la fois informatique et manuelle. La vérification informatique reposait sur des règles de validation. Tout d'abord, nous avons établi un ensemble de paramètres que les données se devaient de respecter selon la connaissance acquise du comportement des actifs cotés. Les paramètres étaient principalement fondés sur les volatilités intrajournalières et interjournalières des prix et étaient adaptés à chaque période analysée.

Ensuite, ces paramètres ont été traduits en scripts SQL qui analysaient automatiquement les données et effectuaient l'extraction des prix violant les règles établies. Ces prix «aberrants» ne provenaient pas nécessairement d'erreurs des opérateurs de saisie, mais pouvaient résulter de périodes de haute volatilité ou de coquilles présentes sur les cotes officielles. Les assistants de recherche les ont systématiquement vérifiés. Des vérifications supplémentaires ont enfin eu lieu sur des échantillons aléatoires représentant entre 10% et 20% des données.

Pour procéder à ces vérifications manuelles, on a développé des applications *ad hoc* en java directement installées sur la base Oracle. Ces applications fournissent deux avantages principaux: premièrement, elles permettent d'explorer les données facilement grâce aux mêmes principes que les masques de saisie Java pour les utilisateurs experts; deuxièmement, l'application prend directement en compte les corrections faites, et facilite ainsi leur importation dans la base avec les autres données à la fin du travail de vérification.

Le processus de saisie de données a demandé des temps de traitement différents selon les périodes traitées et le type de données insérées. Ainsi, la quantité d'informations à saisir pour la première partie du XIX^{e} siècle était relativement réduite, mais nécessitait davantage d'interprétation du fait de la moins grande standardisation des sources. À l'inverse,

à la suite du développement du marché boursier et de l'augmentation du nombre de titres cotés au cours de la Belle Époque, les données demandaient davantage de temps de saisie, alors que l'homogénéisation des sources facilitait l'interprétation. En général, on peut estimer que la saisie d'un «paquet de données» – une dizaine d'années environ en début de période, contre une seule année en fin de période – a exigé environ une semaine de travail pour un groupe de cinq personnes en ce qui concerne le début de la période, pour arriver à deux semaines quant aux données portant sur le début du XXe siècle. La vérification de ces mêmes données a demandé un jour-homme pour le début du XIXe siècle et deux jours-hommes pour le début du XXe siècle.

L'évolution de la qualité et du contenu des sources – avec notamment l'apparition des codes SICOVAM (Société Interprofessionnelle pour la Compensation des Valeurs Mobilières) – a permis d'abandonner la saisie manuelle à partir des cotes publiées en 1950 et d'utiliser une technologie fondée sur la ROC et l'intelligence artificielle.

Conception et déploiement du logiciel spécifique

La saisie semi-automatique concerne donc l'exploitation de deux types de sources: les cotes officielles à partir de 1950 et les annuaires des valeurs négociées en Bourse. Les principes du traitement semi-automatique des données sont les mêmes quel que soit le type de source. Pour pouvoir appliquer la technologie de ROC au traitement de ces deux sources, nous avons procédé en plusieurs étapes. Après la numérisation des sources, nous avons développé un logiciel spécifique qui utilise la ROC pour repérer et reconnaître l'information sur la source numérisée et un système d'intelligence artificielle pour la relier à l'entité correspondante (par exemple l'émetteur ou le titre) déjà stockée dans la base et finalement l'insérer dans la table dédiée. Un tel procédé nécessite un important travail en

amont pour concevoir et développer le logiciel. En effet, l'éditeur du logiciel a besoin de deux types d'information pour écrire son code: d'une part, une description fine de la source, afin de savoir où sont physiquement situées les informations sur celle-ci; d'autre part, les tables de la base où on doit insérer ces informations et les liens qu'on doit créer entre les données.

Il est nécessaire de clarifier ces éléments dans un document unique avant de commencer l'écriture du code logiciel. Les chercheurs chargés de rédiger un tel document doivent donc avoir une bonne connaissance des sources et du modèle conceptuel de la base. Ils doivent ensuite s'assurer de la bonne compréhension du document de la part des informaticiens chargés du développement et continuer à dialoguer avec ces derniers pendant toute la phase de développement et de test du logiciel. L'ensemble de ce processus, de la rédaction des spécifications à l'utilisation du logiciel, est un travail de longue haleine: étant donné la complexité des sources et leur évolution dans le temps, il a fallu prévoir une année de travail en amont de la production pour s'assurer d'une saisie correcte des informations et de leur bonne ventilation dans la base de données.

Des opérateurs doivent néanmoins valider et éventuellement corriger, à chaque étape du processus, le flux de données, une fois capturé par le logiciel, avant sa ventilation dans la base. Cette validation/correction se fait par des interfaces conçues *ad hoc* dans le but d'optimiser le traitement.

Les cotes officielles

Les cotes officielles sont traitées par ce logiciel spécifique à partir de 1950. En effet, jusqu'à cette date, les caractéristiques des sources rendent l'usage de cette technologie peu efficace. Tout d'abord, la qualité initiale de l'impression des sources ne permet pas une reconnaissance

optimale des caractères. Ensuite, à partir de 1950, on associe le code SICOVAM, un identifiant unique, à chaque titre admis à la cote. Inséré manuellement, il facilite l'association des informations capturées à celles déjà présentes dans la base et permet de résoudre dans une certaine mesure le casse-tête des nombreux changements de désignation des titres. Enfin, la structure de la source se stabilise après la Deuxième Guerre mondiale. Chaque nouveau format de la source implique un travail de spécification et de développement informatique *ad hoc;* les changements fréquents rendent donc l'utilisation du logiciel peu économique.

La capture et la ventilation automatisées des données dans la base passent par plusieurs étapes. Tout d'abord le logiciel lit la source pour reconnaître la structure tabulaire contenant les données et pour reconstruire les lignes de cotation: c'est pourquoi tout changement dans cette structure demande une adaptation du logiciel. Une fois qu'un opérateur a validé la structure et l'alignement, le logiciel associe le titre détecté sur la source au même titre déjà existant dans la base en se fondant sur un algorithme d'appariement des noms et sur le code SICOVAM, utilisé comme variable de contrôle. Cette association requiert néanmoins l'intervention d'un opérateur pour lever les ambiguïtés signalées par le logiciel. On calcule la fiabilité de l'appariement sur la base de la similarité entre les chaînes de caractères (nom et code) lues sur la source et celles déjà présentes dans la base. Si le taux de fiabilité ne dépasse pas un seuil paramétrable par l'utilisateur, le logiciel soumet une question à l'opérateur en lui présentant une liste de candidats à l'appariement tirés de la base et classés par ordre décroissant de similarité. L'opérateur peut: a) valider l'association avec le premier titre de la liste, proposée par défaut par le logiciel; b) créer une association avec un autre titre de la liste; c) créer une association avec un titre absent de la liste mais présent dans la base ou avec un titre

nouvellement créé en cas d’erreur ou d’omission. Enfin, le logiciel importe et ventile dans la base, en créant automatiquement les relations opportunes, les données qui se trouvent sur la même ligne du titre apparié.

Les annuaires

La qualité et la structure des annuaires permettent d’en traiter la série complète à l’aide d’un logiciel spécifique. Deux types d’annuaires sont disponibles: les annuaires dits officiels, car publiés par la Compagnie des agents de change de Paris et les annuaires d’éditeurs privés, dont le plus connu est l’Annuaire Desfossés. Ces annuaires, officiels et privés, étaient publiés sous forme de livres reliés et imprimés sur un papier d’assez bonne qualité, contrairement aux cotes. La structure, c’est-à-dire la disposition de l’information sur la source, est très stable dans le temps. Pourtant, le travail de spécification des informations publiées, préalable au développement logiciel, s’avère ardu car ces informations sont de nature différente. Les annuaires sont organisés par émetteur: chaque entité, publique ou privée, ayant émis des titres sur le marché officiel de la Bourse de Paris y fait l’objet d’une description détaillée. Toutes les informations considérées comme utiles pour les investisseurs y figurent. Il faut donc prévoir un traitement spécifique pour chacune d’entre elles. Il convient, tout d’abord, de trouver des repères codables pour délimiter la zone de texte, de déterminer des séparateurs permettant à l’outil d’extraire de la zone les informations utiles et de repérer les modalités récurrentes de ces informations pour les insérer préalablement dans la base et les associer à un identifiant. Il s’agit de fournir au logiciel de ROC un dictionnaire de référence et de lui permettre d’établir les relations utiles, une fois qu’on a créé l’association entre la modalité lue et celle présente dans la base.

Le travail de saisie semi-automatique s'effectue alors en trois étapes. La première consiste simplement à exclure toutes les pages de l'annuaire numérisé qui ne sont pas directement utiles à la constitution de la base de données (par exemple la publicité ou la table des matières). La deuxième étape, commune à tout type de travail avec ROC, est plus longue: il s'agit, d'une part, de s'assurer que le logiciel a bien pris en considération toutes les zones du texte et les tableaux nécessaires et, d'autre part, a «typé» correctement les blocs et les tableaux puisque chacun fait l'objet d'un traitement spécifique. Par exemple, on a saisi toutes les modalités possibles de statut juridique d'une société (société anonyme, société en nom collectif, etc.) dans une table de la base afin de fournir un dictionnaire de référence à la ROC, et on a attribué des identifiants uniques à ces modalités pour permettre au logiciel de créer automatiquement les relations nécessaires. Si la zone de texte concernant l'évolution du statut juridique de la société est typée «administrateurs», le logiciel cherchera à associer les formes juridiques prises par l'entreprise avec les modalités relatives aux postes des administrateurs (président, directeur général, etc.) stockées dans une autre table. Un code couleur est généralement utilisé pour distinguer rapidement les blocs et pouvoir les corriger si nécessaire. L'étape suivante, celle des corrections et des validations, est la plus longue du traitement. Ici, l'objectif est double. Il faut d'abord valider l'association entre l'entité lue par le logiciel sur la source et son correspondant dans la base. Comme pour les cotes officielles, l'association est fondée sur un algorithme de distance similaire à celui appliqué aux titres des cotes. Le but est de permettre au logiciel de ventiler de façon adéquate dans la base les informations capturées sur la source. Il s'agit ensuite de corriger éventuellement les informations mal ou non lues par l'outil de ROC et de les rattacher à l'émetteur. L'opérateur vérifie notamment les informations sur

lesquelles le logiciel a émis un doute quant à la reconnaissance. Lorsque l'opérateur a validé les différentes zones, le traitement de l'émetteur en question est terminé et les données sont automatiquement insérées dans la base. Le processus de saisie semi-automatique, qu'une équipe constituée en moyenne par deux doctorants et trois assistants de recherche effectue, nous a demandé environ un mois pour chaque annuaire.

Tant pour le traitement de la cote officielle que pour celui des annuaires, un certain nombre de règles de validation sont implémentées dans le logiciel afin de procéder à une première vérification des données avant insertion. Ainsi, on effectue des contrôles dans le but de corriger les résultats bruts de la ROC, à l'aide d'une application qui regroupe les ambiguïtés trouvées dans le texte pour pouvoir le corriger en une seule fois. On met également en place des contrôles après insertion. Les requêtes SQL servent ainsi à faire ressortir les écarts d'information entre la source et les données insérées dans la base. Pour la cote officielle, il s'agit, comme pour la saisie manuelle, de repérer les données «aberrantes». Concernant les annuaires, nous tirons parti du fait que certains émetteurs sont décrits dans plusieurs volumes (quand ils ont des titres cotés en réalité). Il est alors possible de contrôler que les informations constantes, comme la date de constitution de la société, sont bien identiques dans chacun des annuaires.

La construction de bases de données est l'un des faits massifs du courant des HN. Les avancées technologiques permettent d'aborder la construction de bases de données d'envergure. L'EQUIPEX DFIH a eu les moyens d'investir dans cette recherche et de mettre en place des outils innovants de capture de données. Pourtant, l'important et indispensable travail humain à fournir pour la constitution de bases de données à partir de sources anciennes, et notamment tabulaires, rend la production coûteuse.

Une fois les données insérées et vérifiées, elles pourront être utilisées pour mener des recherches, non seulement en économie, mais également dans les différents domaines des sciences sociales. Par exemple, trois des auteurs – E. Grandi, R. Hekimian et A. Riva – travaillent actuellement sur un article visant à évaluer les répercussions des *interlocking directorates* sur la performance boursière des établissements bancaires cotés à la Bourse de Paris entre 1883 et 1913 (2016). Pour cela, ils tirent des annuaires des données sur les conseils d'administration des banques, en y appliquant l'analyse des réseaux sociaux, ainsi que les données relatives aux bilans de ces mêmes banques afin d'en évaluer la liquidité et la solvabilité. On teste alors les répercussions de ces variables sur les cours des actions, issus du traitement des cotes officielles, à l'aide de l'économétrie des données de panel, permettant d'exploiter simultanément les dimensions individuelles et temporelles des données DFIH.

La recherche en HN, comme terrain de rencontre entre SHS et STIC, est ainsi appelée à œuvrer pour réaliser la *big data revolution*, c'est-à-dire le développement d'outils permettant de produire des masses importantes de données à un coût raisonnable. La multiplication des projets visant à construire des bases de données à partir de sources historiques nous amène à penser que la fédération de ces efforts pourrait générer d'importantes économies d'échelle dans la recherche et le développement de ces outils. La création d'une plate-forme nationale consacrée à la question de la production de données à partir de sources anciennes et ralliant chercheurs en STIC et en SHS, pourrait mettre à la disposition de la communauté scientifique des compétences, des moyens et des technologies dont le secteur privé bénéficierait également.

CHAPITRE 10

Le numérique et les SIG pour présenter et représenter la population

SÉBASTIEN OLIVEAU

Pour réfléchir aux conséquences de la révolution numérique sur les sciences sociales, on utilise maintenant le terme d'humanités numériques (HN) de façon assez consensuelle. Nous proposons de revenir sur un ensemble d'expériences menées dans le champ de la géographie de la population. Ce retour d'expériences a deux objectifs. Le premier est de montrer l'apport du numérique à la discipline géographique, en rappelant notamment comment celle-ci s'en est emparé. Le second est de souligner la diversité des changements qui en ont découlé, autant dans le domaine de la collecte et du traitement que dans celui de la restitution des informations.

La numérisation de la géographie: une histoire ancienne

La géographie a toujours utilisé des supports visuels de représentation de ses données. Du XIIe siècle, avec le *Livre de Roger* du géographe Al Idrissi, aux publications les plus récentes, la géographie emploie une iconographie riche et variée. Parmi ces images, la carte est perçue comme une spécificité, ou au moins une spécialité, du géographe. Elle est en effet un moyen puissant de présenter des matériaux de recherche, qu'ils en soient au stade exploratoire ou final.

La cartographie mobilise des données inscrites à la surface du globe. Ces données sont multidimensionnelles, et leur représentation nécessite un certain nombre de règles. Par ailleurs, la récolte, l'exploitation et la reproduction de ce type de données sont coûteuses, aussi

bien en temps qu'en argent. On comprendra aisément qu'on ait envisagé très tôt le développement du numérique comme une piste intéressante. La géographie a donc participé dès les années 1960 à ce que l'on nomme couramment la numérisation du monde.

C'est dans le Harvard Laboratory for Computer Graphics, autour de Howard Fisher, que commence en 1965 le développement de SyMAP, le premier logiciel de cartographie sur ordinateur. Dix ans plus tard avait lieu la première conférence américaine sur la cartographie assistée par ordinateur. Lorsque les micro-ordinateurs se développent à partir de 1975 (notamment avec l'Apple II en 1977 et le premier PC d'IBM en 1981), c'est un nouveau monde qui s'ouvre aux géographes. Des expérimentations de logiciels de cartographie, dont peu ont survécu et ont pu se répandre, voient le jour un peu partout.

Les éléments nécessaires à l'évolution de la cartographie vers les systèmes d'information géographique (SIG) sont ainsi présents dès le début des années 1980. Les logiciels de cartographie assistée par ordinateur (CAO) sont dédiés à la cartographie thématique, autrement dit essentiellement au traitement statistique et thématique de données. Les SIG, quant à eux, intègrent explicitement la dimension spatiale des données. Ils permettent donc d'exploiter ces caractéristiques en effectuant des requêtes sur les attributs spatiaux des objets géographiques. Dès 1982, la société Esri lance la première version de son logiciel ArcInfo et quatre ans plus tard MapInfo commercialise le premier SIG pour PC. C'est néanmoins au tournant des années 1990 que les SIG se développent réellement.

L'entrée dans le numérique apparaît capitale dans l'épistémologie de la géographie par au moins trois aspects. Tout d'abord, le numérique induit un nouveau questionnement interne à la discipline par rapport au lien existant entre géographie et cartographie. Dans le même temps, ce questionnement s'exporte à l'extérieur de la

discipline, où l’ensemble des sciences sociales a pris le «tournant spatial» et tente d’intégrer la dimension spatiale des phénomènes étudiés. Enfin, cette entrée dans le numérique permet à ses acteurs de manipuler des données de plus en plus importantes en volume, de plus en plus fines en résolution, mais aussi de plus en plus complexes (intégrant le mouvement, la temporalité). Globalement, on peut réenvisager les questions classiques de recherche, revisiter les modèles, au regard de données plus nombreuses, plus détaillées, plus accessibles.

L’information géographique, autrefois très coûteuse, s’est lentement démocratisée, jusqu’à devenir aujourd’hui un bien commun. Il existe à présent des solutions libres de SIG, dont la plus répandue, QGIS, tend à remplacer les logiciels précités autant dans le monde universitaire que dans l’entreprise et les institutions publiques. Depuis 2006, enfin, la Fondation Open Source Geospatiale (OSGeo) soutient et participe à construire une offre de logiciels *Open Source* en géomatique.

À partir des années 2000, un nouveau support apparaît: les «globes virtuels» qui représentent la terre en trois dimensions. En 2004, la NASA donne accès à une partie de ses données via son logiciel World Wind. Un an plus tard, le logiciel Google Earth diffuse ses images satellites au plus grand nombre. Depuis, les globes virtuels se sont multipliés. On notera ainsi l’existence du globe TerraViva! qui s’adresse à la communauté scientifique en diffusant des images portant des informations spécifiques, comme des classifications de la végétation, des évaluations de zones de risques naturels, la répartition de la population et ses répercussions sur le milieu, etc. Le gouvernement indien a développé à partir des images de ses satellites son propre portail, Bhuvan. Tous aujourd’hui peuvent visualiser (voire exploiter) ces images, quasiment inaccessibles il y a vingt ans.

En parallèle, la cartographie en ligne et interactive s'est aussi largement développée. Cela concerne d'abord la cartographie routière, aussi bien via les sites de grandes entreprises que résultant du travail collaboratif de bénévoles, mais aussi la cartographie statistique, par exemple celle des recensements. Le portail Internet de l'Institut national de la statistique et des études économiques (INSEE) propose ainsi une exploration des données censitaires françaises dans une cartographie interactive.

L'ensemble de ces changements survenus depuis quatre décennies pose problème à la discipline géographique: si la géographie est aujourd'hui dans toutes les poches, comment pouvons-nous, en tant que géographes, nous en servir pour rendre nos résultats lisibles à un plus grand nombre?

Les possibilités de représentation du numérique

Engagés depuis quinze ans dans la recherche universitaire, et travaillant sur des questions spatiales, qui recourent massivement à la cartographie, nous avons tenté d'utiliser les nouveaux outils mis à notre disposition pour rendre nos recherches plus accessibles. L'expérience a permis de mettre ces technologies à l'épreuve, de réfléchir à ce qu'elles pouvaient apporter, et de mieux cerner leurs limites. Deux expériences permettront d'illustrer les potentialités du monde numérique et notamment la variété des supports (CD-ROM, Internet), le contexte de réalisation (solutions techniques disponibles) et les choix effectués (interactivité ou non). Le but est d'éclairer le lecteur sur les objectifs envisagés (publics visés par exemple) et les contraintes techniques qui ont guidé le développement.

L'Institut français de Pondichéry a développé la première application en Inde, à la fin des années 1990. À cette époque, un programme de recherche avait développé une base de données géographiques incluant les données du

dernier recensement à l'échelle des quatre États méridionaux du pays. Ce programme de recherche, dirigé par Christophe Z. Guilmoto, était intitulé South India Fertility Project. Il se proposait d'étudier la baisse de la fécondité en Inde du Sud (http://www.demographie.net/sifp/).

On avait ainsi géolocalisé quelque 75 000 villages et 850 unités urbaines, regroupant 220 millions d'habitants. Si la base de données était avant tout destinée aux chercheurs du programme, il nous a rapidement semblé nécessaire d'en faire bénéficier un plus grand nombre, y compris en dehors du monde universitaire. Nous avons donc cherché une manière de rendre ces données accessibles. La mise à disposition sur Internet n'était pas envisageable pour des raisons d'ordre pratique et technique. Premièrement, la lecture de ces bases de données nécessitait un logiciel spécifique, que les utilisateurs de l'époque étaient peu nombreux à posséder et à maîtriser. Autrement dit, l'intérêt de rendre disponible une base de données qui n'aurait pu être utilisée que par quelques chercheurs maîtrisant les SIG nous semblait nul. La seconde raison était qu'au début du XXIe siècle en Inde, l'accès à Internet était encore trop peu répandu et trop lent pour constituer un support intéressant. On a choisi la diffusion sur CD-ROM car elle permettait de distribuer la base de données en l'accompagnant d'un logiciel spécifique. C'est ainsi qu'est né le projet SIPIS (South Indian Population Information System: http://www.ifpindia.org/content/south-indian-population-information-system-sipis-vol-1-tamil-nadu-and-pondicherry-cd-rom).

L'idée majeure était de montrer que l'on pouvait partager les données de la recherche au-delà du cercle des spécialistes du champ concerné. Il fallait développer pour l'occasion un SIG qui soit accessible sans formation particulière. Nous avons sous-traité cette opération à une entreprise indienne, qui a produit un SIG basique et convivial, reposant sur les technologies d'Esri (une

adaptation du logiciel Arc Explorer). Si la sous-traitance s'est révélée nécessaire, faute de compétence en programmation, elle n'est pas pour autant une solution de facilité. Le chercheur, devenu chef de projet informatique, doit s'accorder avec un informaticien, par définition éloigné de ses problématiques. La principale difficulté a été de savoir ce qu'il était techniquement possible d'obtenir, sans pour autant se laisser enfermer dans les schémas préconçus des développeurs. En outre, le développement était fait en Inde: à la différence culturelle liée aux métiers s'ajoutait la différence culturelle liée à nos origines respectives.

Bénéficiant d'un financement du Fonds des Nations unies pour la population (FNUAP), nous avons mené à bien ce projet pour l'État du Tamil Nadu. Le CD-ROM qui en résulte propose un SIG adapté aux néophytes et des cartes interactives permettant des interrogations sur les objets affichés (16 085 villages) pour avoir accès aux données censitaires. L'établissement par les utilisateurs d'une cartographie des valeurs est aussi possible. De plus, notre équipe a préparé quelques cartes. L'ensemble des résultats est exportable et imprimable. Nous avons ensuite organisé la distribution de ce CD-ROM auprès des administrations régionales et des principaux centres de recherches locaux, car la simple création du CD-ROM ne suffisait pas à garantir sa diffusion.

Si cette opération de valorisation a atteint son but, elle n'en comporte pas moins des limites. Le coût, tant financier qu'en matière de temps de travail, n'a rendu possible la création d'un CD-ROM que pour un des quatre États concernés. Le chercheur qui s'engage dans ce type d'opération doit trouver des financements spécifiques et y consacrer un temps important, alors que l'objet final ne sera que modérément reconnu dans le monde universitaire. D'autre part, nous avons dû constater, avec le temps, la faible durabilité de ce type de support. Développé pour être utilisé avec les systèmes d'exploitation Windows 95 ou

Windows 98, le logiciel ne fonctionne plus aujourd'hui. Enfin, les utilisateurs sont seuls devant l'interface proposée, sans explication sur les phénomènes qu'ils peuvent observer. Ils doivent eux-mêmes construire une lecture des cartes qu'ils créent, ce qui n'est pas naturel et requiert au contraire un apprentissage spécifique.

Instruits par cette première expérience et ses conclusions, nous avons décidé d'envisager une autre forme de valorisation de la base de données. Nous souhaitions nous appuyer sur Internet et proposer aux lecteurs une exploration maîtrisée de cette base de données. C'est naturellement la forme de l'atlas en ligne qui s'est rapidement imposée. Nous avons donc complété le précédent projet et commencé à explorer de nouvelles solutions.

Les objectifs de ce projet étaient de fournir pour les quatre États de l'Inde du Sud des séries de cartes construites et commentées par des spécialistes et facilement disponibles. Nous avons choisi de publier un atlas dans la revue *Cybergeo*. Tout d'abord, la revue est en libre accès en ligne. Ensuite, il s'agit d'une revue universitaire, et le système de *peer review* offre une garantie de scientificité. Enfin, elle offre un hébergement pérenne. Nous avons choisi d'utiliser non pas des cartes interactives (bien que le survol des villes fasse apparaître leur nom) mais des cartes préconstruites, accompagnées de commentaires guidant le lecteur. En revanche, nous avons utilisé des liens hypertextes pour que le lecteur puisse passer d'un État à l'autre lorsqu'il explore un thème, ou d'un thème à l'autre lorsqu'il explore un État. Cette forme de valorisation pallie une partie des manques observés dans le projet précédent: elle ne dépend pas d'un système d'exploitation et ne laisse pas le lecteur désarmé devant la carte. La principale limite de cette approche repose sur la nécessité de simplifier l'accès aux données: le nombre de variables retenues (15 variables déclinées pour chaque État et pour l'ensemble du sud de l'Inde) est faible

en comparaison avec la centaine que comprend la base de données initiale. De même, les données sont passées au filtre du cartographe, et l'utilisateur final est donc le lecteur passif d'une information préconstruite.

À la suite de ces deux valorisations, nous avons entamé en 2005 un nouveau projet visant à rendre disponible une centaine de variables issues des données censitaires récemment publiée pour l'ensemble des 5 468 sous-districts indiens. Il s'agissait pour nous d'explorer les capacités des outils de cartographie interactive en ligne. Après une exploration des solutions disponibles, nous avons fait le choix de nous appuyer sur Géoclip. Ce logiciel offrait plusieurs atouts. Le premier était d'offrir une interface conviviale et esthétique autant pour le constructeur de l'atlas que pour son lecteur. Ainsi, Anne-Claire Couïc a développé le projet dans le cadre de son stage de master 2 en géomatique. Ensuite, l'interface était développée avec la technologie Flash d'Adobe, qui rendait l'atlas lisible sur tous les PC. Le résultat n'a pas fait l'objet d'une valorisation particulière, faute de temps, ce qui reste bien dommage, mais il est accessible en ligne (http://www.demographie.net/atlas2001/). L'atlas propose au lecteur d'explorer les résultats à l'échelle indienne, pour toute la population ou uniquement pour les populations urbaines ou rurales. Une description des données accompagne l'atlas, ainsi qu'une petite aide pour utiliser l'interface.

Ces trois expériences nous ont amené à réfléchir sur ce qui semble être la quadrature du cercle: comment laisser un lecteur libre de travailler les données tout en le guidant? Guider le lecteur vise à ce qu'il ne se perde pas d'une part, et qu'il utilise correctement les données d'autre part. Il y a dix ans, nous concluions une intervention auprès de jeunes chercheurs de la sorte: «L'avenir est certainement à la mise en ligne libre et au mélange des genres. Ainsi, les atlas interactifs en ligne permettront au lecteur de participer à la

création de la carte, d'en examiner les données directement tout en ayant à disposition des cartes déjà finies et commentées» (Oliveau, 2006).

Lorsque l'on compare des grandes caractéristiques des trois projets, on note que l'interactivité (SIPIS et Atlas de l'Inde) semble s'opposer aux commentaires des cartes (Atlas de l'Inde du Sud). Il est en effet difficile de commenter des cartes qui ne sont pas encore produites et que le lecteur créera en fonction de ses questions.

Après avoir tenté d'autres expériences, nous travaillons actuellement à un nouveau projet de représentation cartographique des données de population, dans l'espace méditerranéen. Les apprentissages passés nous permettent d'envisager un projet plus complexe, qui tente de construire une approche où l'utilisateur est tout à la fois libre de travailler les données lui-même et accompagné dans son exploration, autant sur le plan thématique que méthodologique.

Un projet multidisciplinaire de représentation: DemoMed

L'ambition de ce projet, intégré dans un programme plus général d'observatoire démographique de la Méditerranée – DemoMed –, est de proposer une visualisation graphique et une cartographie interactive de données démographiques collectées dans les pays riverains de la Méditerranée. L'ambition est large: il s'agit de proposer un outil de connaissance et d'exploration de la situation démographique méditerranéenne à un public qui dépasse le monde universitaire. On vise aussi bien les décideurs politiques et les journalistes que la société civile. Dans ce cadre, plusieurs défis sont à relever. Celui de la pluridisciplinarité d'abord, puisque la base de données doit répondre aux attentes des démographes comme à celles des géographes, dont on découvre rapidement qu'elles ne sont pas les mêmes. Celui de l'harmonisation des données

ensuite, puisque les sources sont diverses (recensements nationaux, états civils, registres de population, enquêtes, etc.) et les territoires administratifs variés. Celui des méthodes enfin: tout en nous adressant à un public plus large, nous souhaitons préserver la qualité scientifique des productions issues de la base de données. Les utilisateurs doivent pouvoir construire eux-mêmes des cartes et des graphiques, mais il faut cependant que l'outil les contraigne à respecter les usages scientifiques établis (le respect de la sémiologie graphique pour les cartes par exemple: Bertin, 1967). Ainsi, lors de l'élaboration de la base de données, il nous a fallu définir un mode de classification des variables qui satisfasse à la fois les représentations graphiques des démographes (courbes ou diagrammes en bâtons) et celle des géographes (cercles proportionnels ou aplats de couleurs). Chaque variable, lors de son introduction dans la base de données, se voit attribuer un mode de représentation spécifique en fonction de sa nature (qualitative ou quantitative) et de son type (nominale, ordinale, discrète, continue issue de mesures, issue de calculs). L'utilisateur, même novice, produira donc des représentations qui respectent les codes scientifiques actuels.

Cet objectif de respect des normes est le résultat de la mise en place d'un cahier des charges très précis, issu d'un dialogue pluridisciplinaire parfois difficile. Il a fallu ensuite trouver les financements pour externaliser la création de l'interface Web qui permettra d'une part de charger la base de données et d'autre part de l'examiner.

Le module de téléchargement est réservé aux membres du réseau de recherche. Il se caractérise par un contrôle strict des conditions de saisie, afin que les métadonnées nécessaires soient présentées et leur qualité contrôlée. En outre, un formatage rigoureux des variables est aussi indispensable pour pouvoir harmoniser les données issues de sources hétérogènes. Cette saisie exige la validation d'un

administrateur, qui contrôle sa cohérence. De cette manière, les données téléchargées ont une qualité labellisée par le projet. Dans un contexte où l'accès aux données est de moins en moins un problème, c'est la question de leur qualité et de la confiance qu'on peut leur accorder qui devient l'enjeu central de l'information. C'est justement là que les chercheurs doivent intervenir. Un apport majeur de notre projet réside dans ce contrôle de la qualité.

L'autre aspect important est celui de la diffusion. Le développement d'une interface d'interrogation spécifique a été nécessaire, qui permette à la fois une recherche via une carte interactive à l'échelle de la Méditerranée, mais aussi dans des requêtes de type SQL (*Structured Query Language*) traduites en langage courant. Les résultats proposés le sont sous différentes formes (tableaux, cartes, graphiques), au choix de l'utilisateur et en fonction de la donnée. Les variables ainsi présentées sont, dans la mesure du possible, enrichies de ressources extérieures (définitions, références bibliographiques).

Cette base de données est conçue pour être durable. On a pensé dès le début de sa réalisation la possibilité de faire évoluer sa structure. De même, on a résolu la question de sa pérennité fonctionnelle. Si le développement a été externalisé, nous gardons la propriété et le contrôle des codes informatiques. Enfin, la plate-forme Huma-Num (http://www.huma-num.fr) héberge l'ensemble du projet. Cette infrastructure permet d'envisager sereinement la sauvegarde des projets développés, en offrant des conditions professionnelles de stockage, de duplication et de protection des données.

La numérisation de la géographie est une entreprise ancienne qui atteint aujourd'hui sa maturité. La discipline a largement investi ce champ, accompagnant la prise en main par le grand public d'outils destinés d'abord aux experts. C'est une bonne nouvelle. L'entrée dans le numérique a

révolutionné l’usage des informations produites, augmentant leur disponibilité, accélérant leur circulation, améliorant les exploitations que l’on pouvait en faire.

Il reste néanmoins un certain nombre de questions, auxquelles une partie de la communauté scientifique s’attèle à répondre actuellement. La pérennisation des données et de leur support en est une. La dématérialisation pose en effet la question de l’archivage, même si celle-ci existait auparavant. La mise à jour des données est un second enjeu: il faut concevoir des outils suffisamment souples pour autoriser l’actualisation dans l’avenir.

Plus généralement, la question principale aujourd’hui est celle de la documentation des ressources numériques mises à disposition. Le travail de qualification des données (les métadonnées) fait désormais partie intégrante de la production et de la diffusion de données. Ce travail sera un élément essentiel à l’avenir, permettant le développement de ce que l’on nomme aujourd’hui le Web sémantique.

CHAPITRE 11

Gestion des données, partage et conservation pérenne avec le Data Management Plan

AURORE CARTIER, MAGALIE MOYSAN
ET NATHALIE REYMONET

Nous verrons les problématiques de gestion et de partage des données de la recherche, ainsi que leurs freins et leurs leviers, à partir d'outils développés au sein de la COMUE Université Sorbonne Paris Cité (USPC). S'il n'existe pas encore de législation française pour le libre accès (Open Access), le projet de loi pour une République numérique prévoit l'ouverture et le partage des publications comme des données. Pour y contribuer, les Universités Paris Descartes et Paris Diderot ont créé un modèle de plan de gestion de données pour accompagner les chercheurs dans la gestion, la diffusion et la conservation des données tout au long du projet de recherche. Ce modèle est conçu comme un guide pratique qui définit le rôle des acteurs de la gestion des données de la recherche (chercheurs, ingénieurs projet, professionnels de l'information scientifique et technique).

Lire la suite sur
www.parcoursnumeriques-pum.ca/gestion-des-donnees-partage-et-conversation-perenne-avec-le

CHAPITRE 12

La Fonte Gaia: bibliothèque numérique et blogue scientifique

FILIPPO FONIO ET CLAIRE MOURABY

Le projet Fonte Gaia, porté par le Consortium franco-italien COBNIF, réunit chercheurs et bibliothécaires autour d'une ambition commune: fédérer les italianistes européens. Cette réflexion les mène à la conception d'une bibliothèque numérique scientifique commentée (Fonte Gaia Bib) et d'un blog scientifique (Fonte Gaia Blog). Né de la volonté de rendre accessible le patrimoine italien des bibliothèques françaises, Fonte Gaia est aujourd'hui une plateforme d'innovation pour une pratique renouvelée par les humanités numériques de la philologie, de la pédagogie de la littérature et de la traduction. Le partenariat chercheurs-bibliothécaires est la raison d'être et le moteur de Fonte Gaia; le libre accès en est la pierre angulaire; les réseaux sociaux scientifiques en sont le principal canal de valorisation. Depuis 2009, entre difficultés et réussites, Fonte Gaia tente d'ouvrir la voie d'une méthode de coopération au caractère exemplaire.

Lire la suite sur
www.parcoursnumeriques-pum.ca/la-fonte-gaia-bibliotheque-numerique-et-blogue-scientifique

Autres titres de la collection «Parcours numériques»

1. Sous la direction de Michaël E. SINATRA et Marcello VITALI-ROSATI, *Pratiques de l'édition numérique*
2. Maurizio FERRARIS, *Âme et iPad*
3. Matteo TRELEANI, *Mémoires audiovisuelles. Les archives en ligne ont-elles un sens?*
4. Ollivier DYENS, *Virus, parasites et ordinateurs: le troisième hémisphère du cerveau*
5. Sophie LIMARE, *Surveiller et sourire. Les artistes et le regard numérique*
6. Pablo ACHARD, *Les MOOCs. Cours en ligne et transformations des universités*
7. Jocelyn LACHANCE, *Adophobie. Le piège des images*
8. Sophie LIMARE, Annick GIRARD et Anaïs GUILET, *Tous artistes! Les pratiques (ré)créatives du Web*

Les Presses de l'Université de Montréal
www.pum.umontreal.ca

Table of Contents