Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire

Francisca Maria Cabrera

CONSERVATOIRE NATIONAL DES ARTS ET METIERS

Ecole Management et Société-Département

CITS INTD

MEMOIRE pour obtenir le
titre professionnel "Chef de projet en ingénierie documentaire"
INTD RNCP niveau I

Présenté et soutenu par

Francisca Cabrera

le 10 décembre 2014

Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques

Enquête exploratoire

Jury :
CHARTRON Ghislaine, professeur CNAM, Directrice de l'INTD
MINON Marc, Directeur de Cairn
PARISOT Thomas, Responsable des relations institutionnelles de Cairn

Promotion 44

A Anita

Remerciements

Je tiens à remercier très chaleureusement :

Les chercheurs qui ont bien voulu répondre à mes questions et discuter des points de vues divers et passionnants. Ils m'ont tous accordé volontiers de leur temps et très souvent autour d'un café ou entourés de leurs livres (et ordinateurs…). Je tire de ces rencontres bien plus que ce que la seule restitution des entretiens peut laisser transparaître. Elles/ils ont parlé en détail de leurs recherches, du contexte de la recherche aujourd'hui en France, des leurs expériences et des difficultés rencontrées. Parfois de grandes questions épistémologiques passionnantes ont surgi m'apprenant beaucoup sur des domaines dont je ne connaissais pas grand-chose, comme, par exemple, l'histoire sociale de la statistique. Je leur suis donc reconnaissante non seulement d'avoir répondu à mes questions, mais d'en avoir évoqué tant d'autres, élargissant ainsi l'horizon de ma pensée sur le sujet.

Thomas Parisot (Cairn.info) pour son aide précieuse, pour ses conseils avisés et pour sa grande disponibilité.

Ghislaine Chartron, (Directrice de l'INTD et Professeur titulaire au Cnam) pour m'avoir donné l'occasion de réaliser ce travail, pour l'intérêt qu'elle en a manifesté et pour son encadrement pédagogique.

Je remercie Laetitia Cammas de son soutien amical sans faille pendant ces deux années d'INTD, de ses observations intelligentes et de sa bonne humeur constante.

Je remercie enfin Guillermo Cabrera, Lucila Moreira, Daniela Cabrera, Paula Pino, Javiera Moreira, Sylvain Granjon et Helène Ventoura, pour leur soutien inconditionnel sans lequel il aurait été impossible de conclure ce travail.

Notice

CABRERA Francisca. Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques. Enquête exploratoire. 2014. P. 238. Mémoire Titre professionnel de niveau 1, CNAM-INTD.2014.

La réflexion sur les enjeux épistémologiques, culturels, politiques et économiques des données de la recherche en SHS mobilisent actuellement une variété d'acteurs du milieu de la recherche : institutions, chercheurs, éditeurs scientifiques, documentalistes et bibliothécaires qui abordent de différents points de vue les perspectives ouvertes aux SHS par le numérique à des fins d'exposition, valorisation et réutilisation des données de la recherche. Ce mémoire propose, en premier lieu, d'apporter des éléments pour une compréhension synthétique du contexte général dans lequel se pose la question des données de la recherche pour les SHS en France. Dans un deuxième temps, il présente l'analyse de cinquante-trois entretiens avec des chercheurs en SHS qui font part de leurs propres pratiques de recherche à l'heure du numérique. « Qu'est-ce qu'une donnée de la recherche en SHS » ? « Quelles données partager et dans quel but » ? « Quelles conditions, motivations et quels freins au partage » ? « Comment les chercheurs souhaitent-ils valoriser ces données ? » – sont quelques questions parmi d'autres traitées dans ce document. Nous montrerons en dernier lieu comment les questions et problèmes principaux qui s'y dégagent, se projettent en tant que tendances ou sont déjà objets d'initiatives et programmes en France et au plan international, notamment au Royaume-Uni.

DESCRIPTEURS : DONNEE DE LA RECHERCHE, OPEN DATA, RECHERCHE SCIENTIFIQUE, ENTRETIEN, SCIENCES HUMAINES, SCIENCES SOCIALES, FRANCE

The reflection on the epistemological, cultural, political and economic issues of data research in the field of Human and Social Sciences are mobilizing a variety of actors: institutions, researchers, scientists, publishers and librarians are called to improve, enhance and professionalize their research data management skills to meet the challenge of news perspectives open to HSS by the digital turn in these disciplines. In a first moment this papers proposes to provide elements for a synthetic understanding of the overall context in which the question of research data in HSS arises in France. In a second time, based on the analysis of 53 qualitative interviews with researchers in HSS, it aims at contributing to the understanding of current production practices, management, sharing and reuse of research data. As a result it should allow to identify the main problems or peculiarities of social science research in the French context, but also provide the basis for further development, and inclusion in a wider investigation. Finally, Appendices give the reader access to interviews (anonymous) and to various resources that illustrate some initiatives and development programs on Data in France and also in the UK.

Liste d'abréviations

ANR : Agence nationale de la recherche

CDSP : Centre de Données Socio-Politiques

CNRS : Centre national de recherche scientifique

DMP : Data Management Plan

ESFRI : European Strategy Forum on Research Infrastructures

IST : Information scientifique et technique

MESR : Ministère de l'Enseignement Supérieur et de la Recherche

OA : Open Access

OCDE : Organisation de Coopération et de Développement Économiques

SHS : Sciences Humaines et sociales

SIC : Sciences de l'information et de la communication

TDM : Text and Data Mining

TEI : Text Encoding Initiative

TGIR : Très Grande Infrastructure de recherche

TICs ou nTICs : Technologies de l'information et de la communication, nouvelles technologies de l'information et de la communication

Table des matières

Liste des figures

Liste des tableaux et schémas

Introduction

« *Over 50 years ago, Watson and Crick (1953) published the structure of the DNA in a single page article in* Nature*, with no raw data to underpin their findings. Recently, The 1000 Genomes Project Consortium (2010) accompanied their publication in* Nature *with 4.9 terabases of DNA sequences available through the project website and deposited in dbSNP, the datta base of single nucleotid polymorphisms. »* [1]

« Comme l'affirmait Alain Desrosières, la démarche statistique ne peut pas faire l'économie de la construction d'une vraie problématique comme point de départ. La production de données doit être critique, réfléchie et permettre la prise en compte de facteurs d'erreur. Cette école de pensée a permis des travaux aussi importants que ceux de Bourdieu, Boltanski, Chiapello, Furet, etc. Le recours au Data Mining dans le contexte de massification des données disponibles prend, en revanche, le contrepied de cette époque d'utilisation construite et réfléchie des données en sciences sociales. Dans ce sens sa montée en puissance serait le reflet d'une véritable crise épistémologique dans les SHS, le signe que les chercheurs ne seraient plus capables de construire de véritables objets de recherche. Rien ne prouve aujourd'hui qu'en sciences humaines et sociales, l'analyse de grands volumes de données, quelles que soient les performances des outils utilisés, ne débouche sur un niveau de compréhension supérieur des problématiques abordées. »[2]

[1] CORTI L. et autres auteurs. *Managing and Sharing Research Data* – A guide to good practice. V. Bibliographie, p. 1.
[2] Entretien n° 6, V. Annexes.

Les Sciences Humaines et Sociales (SHS) ne seront certainement plus les mêmes dans quelques années, comme elles ne le sont plus par rapport aux années 90, pourtant des années qui ont marqué nos souvenirs par la multiplication des premières plateformes de sources numérisées et textuelles sur Internet et les premières apparitions des OPAC dans l'environnement Web, rendant possible des recherches bibliographiques dans les bibliothèques les plus renommées du monde. Cette phase « hypertextuelle » du Web qui a duré bien une dizaine d'années cède sa place à l'ère des données numériques de l'hyperdata qui permet maintenant de lier les données elles-mêmes. Il s'agit sans doute d'un passage d'ordre technologique, mais les technologies en elles-mêmes ne sont porteuses de sens que si elles font partie intégrante d'un système de valeurs ; or, celles-ci sont essentiellement d'ordre culturel, social et politique.

« L'ère numérique présente des nouvelles opportunités pour les SHS ». Phrase qui ne cesse d'être répétée sans qu'on puisse savoir en quoi consiste cette opportunité : changement de paradigmes ? Nouveaux objets d'étude ? Fin du clivage entre technique et culture ? Possibilité de se hausser côte à côte avec les sciences dites « dures » ? Ces réponses me semblent laisser de côté l'essentiel, à savoir, que les données numériques ouvrent pour les SHS une autre manière de travailler avec le sens qui, paradoxalement, est plus proche des éléments simples du texte que ne l'était leur l'approche interprétative traditionnelle. En effet, les SHS sont aujourd'hui plus textuelles que jamais mais dans un sens très élémentaire où le texte est aussi de la donnée. Le changement de paradigme, s'il s'est déjà réalisé, vient du fait que les données les plus élémentaires (on pourrait dire « brutes ») bénéficient aujourd'hui d'une possibilité inédite de migrer vers d'autres univers, de s'enrichir de nouveaux sens à travers d'autres opérations intellectuelles. Et ici l'autre point essentiel de cette « révolution », celle-ci culturelle et sociale : le numérique, dans son modèle Open Data, permet une forme nouvelle de coopération scientifique qui exclut le modèle de travail des sociétés savantes hermétiques et exclusives.

S'il faut remonter à l'origine de cette étude, il est possible d'évoquer deux expériences qui ont sensiblement compté pour mon orientation vers le sujet de ce mémoire, à savoir, les données de la recherche en SHS.

Le stage réalisé en 2013 à l'AHICF au pôle Archives Orales a été le premier contact concret avec des problèmes liés aux données et, plus précisément, à leur traitement et valorisation. L'équipe projet, composée d'un archiviste, des trois chercheurs et d'une chargée de communication, réalisait la collecte de témoignages auprès de cheminots ayant été en poste lors de la Seconde guerre mondiale. La réflexion menée collectivement par l'équipe portait sur des problèmes divers : restrictions budgétaires limitant le choix des solutions techniques,

enjeux liés au type de données et modes de diffusion et de valorisation envisageables alors qu'il s'agissait de témoignages de personnes encore en vie.

Expérience très riche démontrant la dynamique très positive qui peut s'établir entre les professionnels de l'information et les chercheurs travaillant sur un même projet.

La deuxième expérience n'a été que la suite souhaitée de la première : en 2014, un stage à Cairn.info, plate-forme d'édition scientifique, portant sur le sujet des données de la recherche en SHS. La mission fixée était la réalisation d'entretiens avec des chercheurs en SHS afin de comprendre l'état actuel des pratiques et attentes des chercheurs touchant les données de la recherche.

Les étapes de préparation consistèrent en une veille ciblée, le suivi de quelques manifestations (journées d'études, réunions GFII, séminaires) et la découverte d'un domaine en ébullition et plein de promesses et expectatives de la part des chercheurs et ingénieurs.

Réaliser des entretiens était inédit pour moi. Rencontrer des chercheurs[3] en SHS ne l'était pas. Ils m'ont reçu ou accepté d'échanger par mail ou téléphone, et quelques-uns ont été des « maîtres » à certains moments, révélant la complexité du sujet sur lequel j'avais commencé à avancer de façon un peu hâtive. La richesse qui en est apparue ainsi à certains moments va bien au-delà des limites de la présente étude, qui reste ancrée dans le «maintenant » de l'ère numérique. Cette richesse concerne en effet l'histoire qui sous-tend les différents courants et disciplines des SHS qui ont des approches épistémologiques et des méthodologies très diverses s'opposant, parfois, de façon frontale. Cela concerne aussi l'histoire politique des institutions gérant les données en France, comme par exemple l'INSEE. Ces sujets, bien que fascinants, ne trouvent pas ici une place de développement.

A côté de ce vaste paysage, l'objectif de ce mémoire est donc bien plus modeste et se décline en trois temps :

- Contribuer à identifier les éléments généraux du débat (Première partie) sans lesquels il est impossible de saisir l'importance de l'enjeu des données de la recherche. Ces éléments peuvent et doivent être suivis dans une veille permanente par les documentalistes car ils sont en constante évolution. Ils sont riches en débats politiques et sociétaux, abordant souvent des problèmes épistémologiques qui attendent d'être approfondis par des chercheurs et professionnels en SIC. A titre d'exemple, la situation en France concernant la gestion des données qualitatives dont seulement le contexte général sera présenté ici, et qui mériterait une étude

[3] La forme masculine au pluriel a été finalement adoptée ici pour simplifier la lecture du texte et comprend bien évidemment chercheuses et chercheurs.

réunissant l'aspect « données de la recherche », l'aspect technologique et les politiques des systèmes d'information.

- Présenter un cas d'étude pratique (Deuxième partie) permettant une vision partielle mais représentative de la communauté des chercheurs et des préoccupations en cours en France. Les éléments clés qui se dégagent de ces 53 entretiens pourront être à la base d'une constitution de grilles pour des enquêtes plus larges ou alors devront être suivis par une veille.

- Fournir un cadre pour le développement des pratiques au sein des laboratoires ou centres de recherche en SHS (Annexes), indiquant un nombre de ressources en ligne qui peuvent intéresser aussi bien chercheurs que documentalistes souhaitant avoir des éléments pour constituer un cahier de charges ou un plan de gestion des données.

- Proposer des entrées bibliographique pertinentes à ceux ou celles qui souhaiteraient approfondir les thématiques évoquées.

Première partie :
Autour des données de la recherche

1 Enjeux et débats autour des données

A la date de début des recherches exploratoires qui ont orienté cette étude, c'est-à-dire mars 2014, la question des données se trouvait au cœur de l'actualité : les *Big data*, croissance exponentielle des données numériques dans le Web à côte de l'Open Data comme orientation politique et programme pour les données publiques. A la même période, un mémoire qui est aujourd'hui une référence sur le sujet de l'ouverture des données de la recherche, *De l'Open Data à l'Open Research Data: quelle (s) politique (s) pour les données de recherche ?* circulait dans un cercle réduit mais confirmé de personnes intéressées contribuant à la réflexion sur les données de la recherche. A la fin du mois de mars les débats et controverses autour des politiques de fouille de texte et Data Mining des grands éditeurs scientifiques atteignirent un point d'orgue lors des négociations entre Couperin et Elsevier, amplement médiatisées sur le Web.

A côte de ce contexte en ébullition, les manifestations en SHS concernant les nouvelles approches au numérique se multiplient, prolongeant les réflexions sur divers courants de l'actualité : possibilités du Web sémantique, qui commence à faire ses preuves dans des projets comme *Europeana*, pour les SHS ; la multiplication des projets de valorisation des données de la recherche, les nTICs qui rendent possibles des traitements nouveaux et de présentation nouvelles des résultats (data visualisation). Cette liste n'est pas exhaustive…

Les données de la recherche en SHS sont ainsi un sujet qui se retrouve au carrefour de toutes ces questions plus larges.

Pour être à même de déceler un contexte plus précis et centré sur problématique nous avons réalisé toutes les étapes menant à une veille ciblée : un sourcing, la mise en place de deux outils, le premier destiné à collecter l'information et le deuxième à en réaliser la curation :

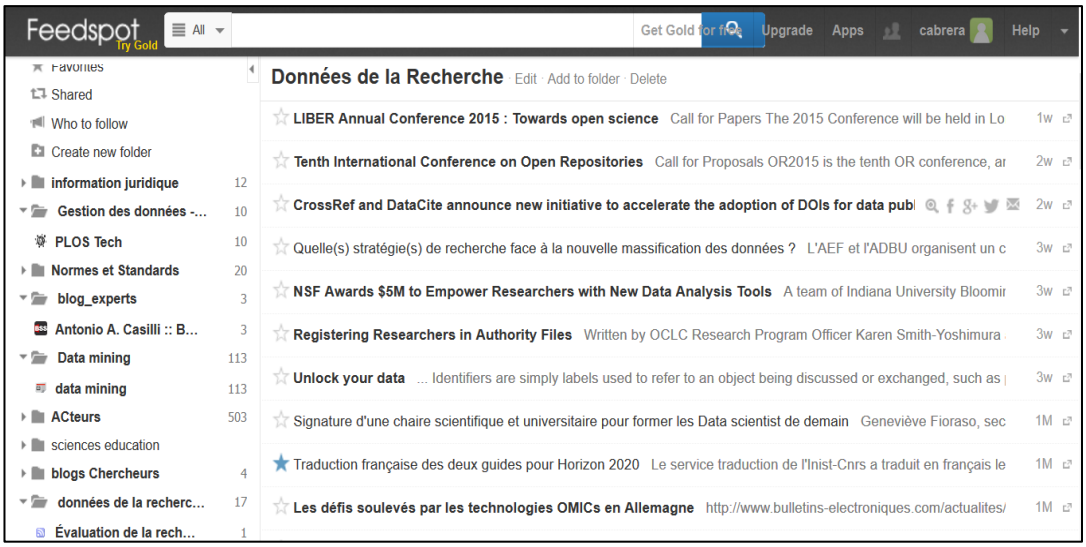


Figure 1 - Plateforme de veille Feedspot

La plate-forme de veille réalisée sur Feedspot pourrait être ouverte au public prochainement. Les sujets des données de la recherche et son contexte étant en constante évolution, nous avons estimé qu'elle pourrait offrir un outil aux personnes souhaitant s'y pencher.

Enfin, la plateforme de curation Scoop.it crée à la même occasion permettait de rendre visible les informations importantes sur le sujet.

Figure 2 - "Les données de la recherche à l'heure du numérique", http://www.scoop.it/t/donnees-de-la-recherche-en-shs-a-l-heure-du-numerique

1.1 *Big data*, sciences humaines et sociales

« Sixty years ago, digital computers made information readable. Twenty years ago, the Internet made it reachable. Ten years ago, the first search engine crawlers made it a single database. Now Google and like-minded companies are sifting through the most measured age in history, treating this massive corpus as a laboratory of the human condition. They are the children of the Petabyte Age. » [4]

> *« Big data is an all-encompassing term for any collection of_data sets so large and complex that it becomes difficult to process using traditional data processing applications. »*[5]

"The methodological danger is that the flood of data generated by our innumerable measuring devices may convince us that data is enough, that there is nothing beyond the microarray paradigm, and that opaque, enormous, data-driven models are the privileged way to approach phenomena, even though they become so similar to the famous map of Borges […], that was useless, since it was as big as the geography it was supposed to describe." [6]

L'impact économique des *Big data* à l'horizon 2020 en France est estimé à 9 milliards d'euros avec création de 130 000 emplois. Face à la montée en puissance de ces enjeux scientifiques, technologiques, économiques et de souveraineté qui sont aujourd'hui les Big data, Geneviève Fioraso, secrétaire d'Etat en charge de l'Enseignement supérieur et de la Recherche a parrainé, mercredi 15 octobre 2014, la signature de la chaire "Data Scientist" créée par l'école Polytechnique, les entreprises Keyrus, Orange et Thales et portée par la Fondation de l'X. L'objectif est d'encourager la formation des professionnels à profil « Data scientist » ayant des capacités à traiter et analyser ce type de données.[7]

Cet investissement d'envergure, signale une reconnaissance institutionnelle et politique, insérant le Big data dans les champs de la recherche de manière durable.

---

[4] ANDERSON C. « The End of Theory: The Data Deluge Makes the Scientific Method Obsolete », Wired Magazine, 06.23.08, en ligne : <http://archive.wired.com/science/discoveries/magazine/16-07/pb_theory>

[5] Wikipedia, the free encyclopedia, "Big Data", en ligne : <http://en.wikipedia.org/wiki/Big_data>

[6] NAPOLETANI D. ; PANZA M. ; STRUPPA D."Is big data enough? A reflection on the changing role of mathematics in applications", 2014, 61 (5), pp.485-490. <halshs-00984828> [Consulté le 1 janvier 2015]

[7] « Signature d'une chaire scientifique et universitaire pour former les Data scientist de demain », en ligne : <http://www.enseignementsup-recherche.gouv.fr/cid83013/signature-d-une-chaire-scientifique-et-universitaire-pour-former-les-data-scientist-de-demain.html>

1.1.1 Pourquoi les Big data intéressent les SHS?

En France et à l'international le débat sur les enjeux et opportunités des Big data pour les SHS perdure depuis plusieurs années et semble loin de s'achever[8].

Michel Wieviorka, à l'occasion de la sortie du livre «L'impératif numérique »[9] a souligné les transformations décisives provoquées par le tournant numérique dans le champ des SHS.

Une première période de cette transformation, dont les évolutions sont observables depuis déjà une quinzaine d'années, concerne l'accès aux sources qui sont aujourd'hui foisonnantes sur le Web, le temps et l'espace de la recherche se trouvant par-là profondément modifiés.

A présent, une deuxième période commence pour les SHS marquée, d'une part, par la production croissante des données numériques par les chercheurs mais aussi par l'ouverture d'un champ expérimental nouveau pour la recherche et particulièrement pour les Sciences sociales : cette grande masse de données du Web, qu'on appelle couramment de Big Data, et qui peut être explorée à l'aide d'outils performants de collecte et d'analyse.

Une question de quantité mais aussi de qualité: des données à granularité fine et mettant en évidence le phénomène de « réseau »

Concernant les Big data, il est très courant de mettre l'accent prioritairement sur leur aspect quantitatif et la performance technologique des nouveaux outils spécialisés à interroger ces données. L'aspect qualitatif joue néanmoins un rôle de première importance dans

[8] Seulement pour l'année 2014 on recense en France plusieurs journées d'études et colloques, par exemple : *Les enjeux éthiques du Big Data Opportunités et risques*. Journée du 22 mai 2014 à la Société Française de Statistique. [Interventions et programme] en ligne : <http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=1799>; Big data, entreprises et sciences sociales - Usages et partages des données numériques de masse, Journée d'étude Edu lundi 2 juin 2014 au Collège de France [programme]. En ligne : <http://www.college-de-france.fr/media/pierre-michel menger/UPL2275351989395172789_Menger_Colloque_2014.pdf>; Mastodons International Workshop on Big Data Management and Crowd Sourcing towards Scientific Data Lundi 30 juin 2014, IBC & LIRMM (UM2, CNRS-Mastodons, INRIA), en ligne : <https://www.lirmm.fr/actualites/workshop-mastodons-big-data-management-and-crowd-sourcing-towards-scientific-data>; « Quelle(s) stratégie(s) de recherche face à la nouvelle massification des données ? », Colloque organisé par l'AEF et l'ADBU, mardi 2 décembre 2014, en ligne :< http://adbu.fr/actualites/strategies-de-recherche-aef-et-ladbu-organisent-un-colloque-sur-les-enjeux-de-la-massification-des-donnees-paris-2-decembre-2014/>. L'année de 2015 s'ouvre également avec colloque du DEFI MASTODONS sur la gestion, l'analyse et l'exploitation des très grandes masses de données scientifiques qui aura lieu le jeudi 22 janvier et le vendredi 23 janvier 2015 au CNRS, en ligne : <http://www.cnrs.fr/mi/spip.php?article631>

[9] *L'impératif numérique*, Paris : CNRS éditions, 2013, 64 p.

l'argumentaire en faveur de l'aspect « révolutionnaire » du Big data pour les sciences humaines.

Il s'agit du constat d'une opportunité inédite d'exploitation d'autres types de données que celles, « classiques », issues d'enquêtes et collectées par les institutions publiques d'études. Les données du Big data seraient plus « proches » aux individus, à granularité tellement fine permettant l' « irrésistible » développement d'algorithmes de prédiction de comportements ou d'analyse de situations instantanées (déplacements des individus, géolocalisation)[10]. Il s'agit, en outre, de données dont les liaisons sont signifiantes et qui peuvent être objet d'analyses et traitements complexes.

Les réseaux sociaux par exemple, regorgent de données qui peuvent être exploitées à des fins d'études d'opinion, alors que les traces numériques laissées par les internautes constituent un terrain riche pour des études d'usages sociaux du Web ou par des études sur l'usage des TICs dans différents cercles de population. « Explorer » et « découvrir » des rapprochements inattendus grâce à des logiciels mining et analyse de données, les restituer à travers des visualisations et cartographies grâce à des outils de dernière génération de *datavisualisation*[11], est un mouvement puissant à l'intérieur des sciences en général et aussi des sciences humaines[12].

De plus, ces données semblent pouvoir satisfaire ou conforter deux préoccupations majeures des méthodologies de collectes de données en Sciences sociales appliquées à certaines études: réussir une enquête de terrain du type observation non participative avec une granularité pourtant assez fine et pouvoir collecter une grande quantité des données en très peu de temps.

Combler le "vide de réalité" des sciences statistiques par la performance technologique du calcul ?

Or, plus précisément, ces données semblent pouvoir « coller » objectivement avec la réalité, privilégiant les individus à la structure, alors que les procédés statistiques s'appuient sur l'élaboration des modèles, représentations idéalisées des sociétés, à travers des observations

[10] NOYER J.-M.; CARMES M. L'irrésistible montée de l'algorithmique : méthodes et concepts en SHS. 2013. <sic_00911858>

[11] Le développement d'outils de visualisation des données parallèlement à leur analyse est une activité centrale de Medialab à Sciences Po, en ligne : <http://www.medialab.sciences-po.fr/projets/teaching-controversy-mapping/>. A noter également la thématique du ThatCamp de cette année : Les "datavisualisations" au cœur des sociétés numériques : former et (s') informer sur/par la visualisation des données », en ligne : <http://thatcamp69.hypotheses.org/>

[12] DACOS M. ; MOUNIER P. Humanités numériques – État des lieux et positionnement de la recherche française dans le contexte international, p. 16. En ligne : <http://www.institutfrancais.com/sites/default/files/if_humanites-numeriques.pdf>

d'une parcelle limitée de ses membres (le principe même de constitution d'un échantillon est de « dézoomer pour mieux comprendre »). Le mouvement de l'open data (qui sont des données de l'ordre du Big data), comme le note Dominique Cardon, « refuse les techniques d'échantillonnages au profit d'une visée d'exhaustivité complète et de granularité la plus fine possible des données. »[13]

Découvrir ce qui se cache dans les chiffres

Par ailleurs, ces données, traitées et analysées, peuvent également apporter des éléments nouveaux et des corrélations inédites qui n'étaient pas prévus au départ de la recherche.

Le « εὕρηκα » d'Archimède n'est plus réservé qu'aux sciences dures... Ces « vérités » peuvent engendrer des modèles a posteriori et construire des nouveaux objets pour les sciences sociales, qu'elles pourront analyser, comparer et interpréter. Ce qui suggère le passage à des niveaux supérieurs de connaissances des réalités sociales.

La physique sociale de Pentland: "l'extraction de la réalité" (reality mining) va rendre possible la modélisation mathématique de la société...

Directeur du Laboratoire de Dynamique humaine du MIT Alex Pentland compte sans doute parmi les plus enthousiastes de ces nouvelles possibilités apportées par l'analyse des données du Big data. Son concept de "physique sociale" repose sur l'idée d'une modélisation mathématique de la société orientée à des finalités de management de la société: connaître le fonctionnement pour pouvoir avoir un impact sur la société.

Cette possibilité de connaissance est offerte par les technologies d'extraction et d'analyse des données liées aux comportements de personnes mais aussi aux flux d'informations qui circulent : *"La physique sociale est une science sociale quantitative qui décrit de manière mathématique l'efficacité des connexions entre l'information et le flot d'idées d'un côté et le comportement des gens de l'autre."*[14]

1.1.2 Un enthousiasme à nuancer selon certains sociologues

Les possibilités inédites offertes par le Big data suscitent un enthousiasme qu'il faut pourtant nuancer par une approche plus critique, renouvelant ainsi avec l'esprit propre aux SHS, qui

[13] CARDON D. "Zoomer ou dézoomer? Les enjeux politiques des données ouvertes" in OWNI, 21 février 2011. En ligne : <http://owni.fr/2011/02/21/zoomer-ou-dezoomer-les-enjeux-politiques-des-donnees-ouvertes/>

[14] « Big Data : vers l'ingénierie sociale ? » InternetActu.net, article paru le 20 mai 2014. En ligne : <http://www.internetactu.net/2014/05/20/big-data-vers-lingenierie-sociale/>

se doit d'interroger les importants changements des paradigmes scientifiques qui découlent des procédés d'extraction et d'analyse automatique des masses de données.

Parmi les critiques les plus relayés et citées sur le Web entre 2011 et 2012, apparaît le texte rédigé par deux sociologues américaines Danah Boyd et Kate Crawford : "Six provocations à propos des *Big data*"15. A côte de ce regard critique, l'approche du sociologue Dominique Cardon contextualise le Big data dans leur rapport à l'histoire sociale des statistiques et interroge dans ce cadre le sens de "donnée brute" donné aux Open data.

Nous proposons ici un rapide aperçu de quelques thématiques et problématiques qui jalonnent ces débats et présentons en Annexe les principaux acteurs communiquant sur le sujet.

1.2 Le Text and Data Mining (TDM) et l'édition scientifique

1.2.1 Quelques définitions

Le Data Mining est actuellement un thème largement évoqué à travers des approches émanant de communautés très diverses. Depuis quelques années seulement ce terme apparaît – associé au Text mining – de façon plus récurrente dans les champs des SHS et dans celui des sciences de l'information où des profils de poste comme « Data analyst » commencent à monter en puissance. Malgré cet air de nouveauté, le data mining comme le text mining (le premier étant un prolongement du second) sont des termes qui apparaissent dans le milieu scientifique au début des années 80.

Quelle est la différence essentielle entre ces deux procédés ?

"In the early days, there was little agreement on what the term data mining encompassed, and it can be argued that in some sense this is still the case. Broadly, data mining can be defined as a set of mechanisms and techniques, realized in software, to extract hidden information from data. The word *hidden* in this definition is important; SQL style querying, however sophisticated, is not data mining. In addition, the term information should be interpreted in its widest sense[16]. »

[15] BOYD D. et CRAWFORD K. « Six provocations à propos des big data ». in Mounier, Pierre, (dir.) Read/Write Book 2 : Une introduction aux humanités numériques. OpenEdition Press, 2012. En ligne : <http://books.openedition.org/oep/273>

[16] COENEN Frans. Data mining : past, present and future. The Knowledge Engineering Review, Vol. 26:1, 25–29.& Cambridge University Press, 2011, En ligne : <doi:10.1017/S0269888910000378>

En sciences informatiques, le data mining s'inscrit comme un procédé à l'intérieur d'une démarche plus large, le *Knowledge Discovery in Data* (KDD). Comme le souligne l'auteur en citant la définition de Fayyad[17], les « informations cachées » dont il est question sont alors la découverte des « patterns » valables, nouveaux, potentiellement utiles et compréhensibles.

Alors que le data Mining travaille en général sur des données structurées et stockées dans des bases de données, le text mining est un procédé qui opère sur des données textuelles non structurées. De ce fait, il procède à travers des moyens techniques spécifiques qui diffèrent de ceux du data mining, car il s'agit de traiter automatiquement le langage écrit. Le text mining peut être dans le prolongement d'un processus d'extraction de connaissances des données du data mining.

1.2.2 Text mining dans les SHS

L'histoire du développement des outils informatiques de *text mining* (fouille de textes ou analyse textuelle) remonte aux années 1970 et prend un nouvel essor dans les années 80 avec des applications dans le champ de la linguistique et plus tard, dans les années 90, dans celui des sciences sociales. L'exemple emblématique d'une application des logiciels de text mining en sciences sociales reste sûrement celui de l'analyse terminologique de corpus de textes de management des années 60 et des années 90 pratiquée avec le logiciel Prospéro[18], sur laquelle l'œuvre « Le Nouvel esprit du capitalisme » (1999) de Luc Boltanski et Ève Chiapello s'est bâtie.

Il n'est pas inutile d'évoquer cette expérience dans les termes des auteurs:

« Les deux corpus [de textes de management des années 60 et 90] ont été traités à l'aide du logiciel Prospero@ (…) qui combine une approche lexicographique et une approche herméneutique permettant la codification et la construction interactive de catégories (personnages, êtres collectifs, objets, actions, etc.) et l'élaboration de représentations adaptées à la fois aux textes concernés et à la problématique de recherche. (…) Ce logiciel nous a permis de comparer de manière systématique les deux corpus et de valider que notre analyse de leur contenu exposée au chapitre I était un reflet assez fidèle et non le résultat d'un biais d'interprétation. »[19]

Avec les développements rapides des technologies de *mining* les pratiques de recherche en SHS sont susceptibles d'évoluer très significativement ces prochaines années. Le *text mining*

[17] FAYYAD U., Piatetsky-Shapiro, H. & Smyth, P. 1996. The KDD process for extracting useful knowledge from volumes of data. Communications of the ACM 39 (11), 27–34.
[18] Logiciel développé dans les années 1990 par Francis Chateauraynaud et Jean Pierre Charriaud ; site : <http://prosperologie.org/?sit=22>
[19] BOLTANSKI L. ; CHIAPELLO E. Le nouvel esprit du capitalisme (1999), Gallimard, 2011, Appendices, Annexe 3, p.727-728.

et la classification automatique deviennent des technologies de première importance pour structurer les données non structurées du Big data et en extraire des connaissances nouvelles ou fournir des outils d'aide à la décision.

Ces remarques très générales permettent de comprendre l'origine des quelques critiques aux traitements des données via le *Data mining* et la fouille de textes dans les SHS.

En effet, très tôt le problème de la quantification s'est posé aux Sciences sociales. Les sujets de controverse d'aujourd'hui se portent plutôt sur les données elles-mêmes, sur la pertinence de leur utilisation massive et sur les limites des interprétations des résultats des traitements automatiques qui caractérisent le mining[20].

1.2.3 Quelle situation dans l'actualité pour les SHS ?

Nous présentons à titre d'hypothèses quelques observations issues de nos recherches sur le Web visant à comprendre le TDM en SHS.

- L'utilisation de ces technologies induit des orientations épistémologiques et crée des champs disciplinaires : travailler avec des grands volumes de données, rend possible des approches prédictives, identification des tendances, réalisation d'analyses de controverses, etc. Certaines disciplines semblent particulièrement concernées par ces méthodes, comme les sciences politiques et la sociologie ;

- Un discours critique sur le data mining se développe d'un point de vue épistémologique notamment (v. **1.1** ; 1.1.2) mais en parallèle il existe une multiplication des projets ayant pour objet le développement des outils de traitement automatique des données dans le monde scientifique. L'intérêt suscité par le *data mining* va de pair avec le développement des techniques de visualisation des données traitées (Ex : travaux du Medialab à Sciences Po), et de leur «cartographie»[21] ;

- La définition d'un cadre juridique pertinent pour le traitement des données est un point clé, qui a donné lieu, dans certains cas, à la dénonciation des politiques des grands éditeurs scientifiques propriétaires des bases de données hébergeant ces données (cf. ci-dessous, 1.2.5)

[20] CHARTRON G. « Open Access et SHS : Controverses », ArchiveSIC, version pré-print déposé le 24/03/2014. En ligne : <http://archivesic.ccsd.cnrs.fr/docs/00/96/52/72/PDF/Ghislaine-Chartron_RSS-fA_vrier2014-preprint.pdf>

[21] Voir par exemple les data visualisations réalisées avec le logiciel Gephi <http://www.iramuteq.org/captures-decrans/resultats/exporter-dans-gephi>

1.2.4 Exemple de projets et logiciels utilisés dans les recherches en SHS

- Sigma, ScienceScape, ANTA, Hyphe : projets développés par le Medialab : http://www.medialab.sciences-po.fr/fr/projets/
- Prospéro (PROgramme de Sociologie Pragmatique, Expérimentale et Réflexive sur Ordinateur) : http://prosperologie.org
- Base textuelle FRANTEXT, ATILF/CNRS et Université de Lorraine. Site internet : http://www.frantext.fr
- Applications logiciels : Alceste, Spad, Lexico, packages logiciel **R** (tm, R.TéMiS, IraMuTeQ)
- Le logiciel open-source de textométrie TXM, dont le développement est coordonné par le laboratoire ICAR (à Lyon, sur le site de l'ENS)
- Calliope : Logiciel développé depuis 1997 par Mathilde de Saint-Léger ; Modyco – Paris Ouest /CNRS http://www.calliope-textmining.com/
- Projet Tirésias http://prosperologie.org/?sit=31
- Webtatlas http://webatlas.fr/wp/
- Ressources et méthodes d'analyse textuel publiés diffusées sur le site de l'INED https://www.ined.fr/fr/ressources-methodes/methodes-analyses-statistiques/analyse-textuelle/

1.2.5 La position des éditeurs scientifiques et des institutions sur le Data mining

La puissance des opportunités offertes à la recherche à travers les usages du data mining et du text mining appliqués à l'extraction d'informations de corpus d'articles scientifiques est devenue célèbre grâce au projet text2genome. Ce projet a réalisé l'extraction des séquences de DNA d'environ 3 millions de publications scientifiques pour produire une cartographie en ligne dans laquelle chaque région est liée pertinemment aux articles[22].

[22] Pour cet exemple et pour d'autres exemples d'application du Text mining à la recherche en sciences exactes, v. Van Noorden R. « Text-mining spat heats up ». In *Nature*, 20 march 2013. En ligne: <http://www.nature.com/news/text-mining-spat-heats-up-1.12636>

Le text mining produit de la valeur par l'extraction de nouvelles connaissances de la masse grandissante d'articles produits tous les jours par les communautés scientifiques. En outre, le text mining permettrait également de réduire la redondance dans la recherche scientifique, identifiant des axes de collaboration potentielle à travers la diversité de domaines scientifiques.

Dans le champ des SHS, ces procédés possèdent des applications intéressant un nombre croissant de chercheurs travaillant sur l'exploration statistique de nombreux corpus des textes, indexés et hébergés dans les grandes bases de données des éditeurs scientifiques.

Ces procédés nécessitent toutefois l'extraction automatisée de corpus d'articles des bases de données des éditeurs et donc de la possibilité de le faire sans restrictions techniques ou légales. Puisque les pratiques de TDM n'étaient pas prévues au départ par les licences négociées avec les institutions et bibliothèques, les grands éditeurs scientifiques ont souhaité imposer des nouvelles licences et conditions techniques (leurs propres APIs) à ces procédés, ajoutant des coûts nouveaux à ceux déjà très élevés pris en charge par les institutions.

Au début de 2014 les discussions portant sur l'accès aux données hébergées dans les plateformes des grands éditeurs scientifiques en vue des pratiques de data mining et de text mining, sont devenues tendues sur le plan européen. Les jeux de forces entre les grands éditeurs scientifiques, soucieux de protéger leurs intérêts commerciaux et les institutions de l'enseignement supérieur et de la recherche, ont été objet des travaux d'un *working groupe* de la Commission Européenne[23].

En France, divers acteurs se sont impliqués dans les discussions comme, par exemple, Liber (Ligue des Bibliothèques Européennes de Recherche), Couperin ou le CSPLA (Conseil supérieur de la propriété littéraire et artistique). Les points controversés de ces discussions se situent aux croisements de plusieurs sujets : régimes de la propriété intellectuelle, droit de l'information, prérogatives de l'Open data Research et de l'Open Access, mettant en ébullition des sujets sensibles en France concernant l'accès aux publications et données produites par des recherches financées sur fonds publics.

Du point de vue juridique, le data mining se heurte ainsi aux cadres juridiques suivants :

- Droit d'auteur, protégeant les contenus de ces bases de données en tant qu' « œuvres de l'esprit » ;

[23] Text and Data Mining Working Group <http://ec.europa.eu/licences-for-europe-dialogue/en/content/text-and-data-mining-working-group-wg4>

- Droit d'auteur protégeant la base de données elle-même en tant que création intellectuelle (Code de la propriété intellectuelle - Article L112-3), une fois attestée l'existence d'une empreinte de la personnalité de l'auteur / producteur de cette base, à travers un agencement ou une architecture de l'information originale, par exemple ;

- Droit spécifique (droit *sui generis)* protégeant les producteurs des bases de données en tant que personnes ayant pris l'initiative et le risque d'investissements financiers, matériel ou humain (article L.341-1 al.1 du Code de la propriété intellectuelle)[24]

Alors que les négociations entre les éditeurs scientifiques et les divers représentants institutionnels en France (CSPLA, Couperin) et en Europe (Liber) appelés à trouver des solutions adaptées à la tension entre protection et exploitation des données semblent très difficiles, survient, en avril 2014, la nouvelle concernant la mise en œuvre d'une réforme du droit d'auteur au Royaume-Uni. Cette réforme juridique, visant un alignement avec certaines exceptions au droit d'auteur prévues par loi dans d'autres pays européens ne s'est pas contenté de rattraper le retard par rapport aux pays voisins, mais a également innové en introduisant l'exception en faveur du Data mining et du Text mining. Sous l'influence de l'exemple britannique, le contexte des débats semble à ce moment gagner en intensité (Fig. 3) et viser des objectifs plus ambitieux qu'une simple négociation sur les types et périmètres des licences imposés par les éditeurs. C'est le droit d'auteur en général et la nécessité d'une réforme en France qui sont visés par les voix les plus critiques et, dans l'immédiat, le statut de la donnée « brute » textuel, qui échapperait au concept d' « œuvre d'esprit » protégée par le code de la propriété intellectuelle.

[24] Les-infostrateges.com. Le droit des bases de données, 20 mai 2007, en ligne : <http://www.les-infostrateges.com/article/0705296/le-droit-des-bases-de-donnees>; Delaporte B. « La protection juridique des bases de données illustrées par les dernières jurisprudences », in JDN, 24 juin 2010, en ligne : <http://www.journaldunet.com/solutions/expert/47489/la-protection-juridique-des-bases-de-donnees-illustree-par-les-dernieres-jurisprudences.shtml>; Legifrance. Code de la propriété intellectuelle, en ligne : <http://www.legifrance.gouv.fr/affichCode.do;jsessionid=B0F5CDFC4EC740C4E18381DE7DB3C354.tpdjo05v_1?cidTexte=LEGITEXT000006069414&dateTexte=20141229>

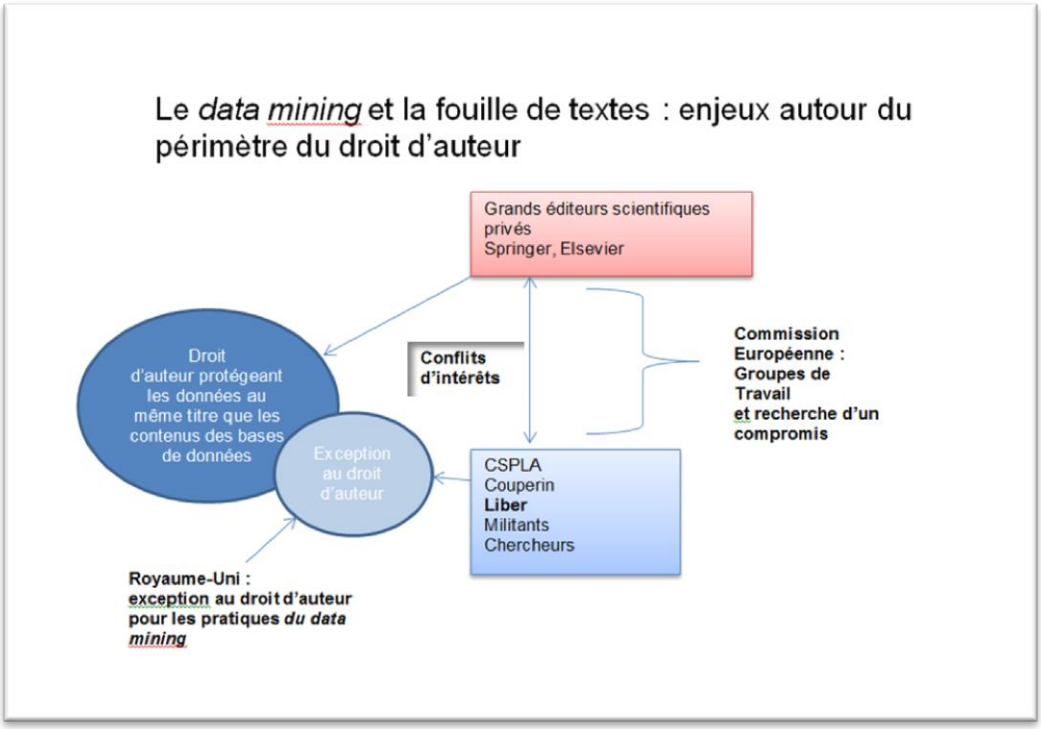


Figure 3 – Enjeux juridiques impliqués dans le *Text* et *Data mining*

Les tableaux à suivre résument les principaux éléments des débats en France à la date d'avril 2014[25] :

Tableau 1 – Termes des conditions imposées par les éditeurs et de contestation de Couperin et du CSPLA

QUE PROPOSENT LES EDITEURS ?	QUE CONTESTENT LES ACTEURS INSTITUTIONNELS ?
L'extraction et le traitement d'un grand volume de données doit être négocié via **une licence** payante car ces données sont protégées par le code de la Propriété Intellectuelle	La voie contractuelle introduit des dispositifs qui menacent l'indépendance de la recherche et constitue un abus financier.

[25] A la date de finalisation de ce mémoire le débat a connu des évolutions en France avec notamment la marche arrière réalisé par le CSPLA en relation aux positions critiques de Couperin et de l'ADBU face aux politiques de grands éditeurs et aux verrous juridiques de la législationk française relative au droit d'auteur. V. rapport rendu public en juillet 2014 en ligne et billets de blog de Pierre Carl Langlais du 29 octobre : Text Mining vers un nouvel accord avec Elsevier.

Sciences Directes (Elsevier) : limite de 10 000 articles par semaine	Cette condition méconnait entièrement le réel besoin des chercheurs en matière de fouille de données : analyse croisée et simultanée d'une hétérogénéité de données ou jeux de données, et non pas analyse successive des différents data sets provenant des différentes bases de données)
Elsevier impose de publier les résultats du TDM en CC-BY-NC	Les résultats d'un TDM sont de faits ou des données qui ne sont pas couverts par le droit d'auteur. Ils relèvent de l' « information » brute et ne seraient pas susceptibles d'être placé sous licences. En plus, la clause « no-commercial uses » peut mettre dans l'embarras des chercheurs travaillant sous financements publics-privés ou privés.
Les citations de ces résultats ne doivent pas excéder 200 mots et doivent inclure les liens vers les contenus originaux	Ces limites sont arbitraires et ne pourront jamais garantir une représentativité suffisante de la recherche et du corpus fouillé.
Springer : les chercheurs doivent remplir un formulaire de demande d'accès en détaillant leur projet de recherche après quoi ils pourront se voir accorder une clé d'accès à leur API	- Les éditeurs ne doivent pas avoir un pouvoir de décision sur la pertinence d'un projet de recherche. - Ce processus établit un droit de regard sur le *data mining* et la recherche en cours. En plus, il facilite la constitution de vastes corpus de métadonnées de la recherche. Springer sait désormais exactement qui étudie quoi avec son corpus.
Elsevier permet la pratique de TDM uniquement via leur API : - Tout autre traitement automatique (crawlers, spider, robots) est proscrit. - Elsevier se réserve le droit d'interrompre l'accès à l'API dans le cas où le serveur serait surchargé.	Imposer un API représente une restriction : - à la liberté du chercheur - à l'utilisation d'autres outils et a un impact sur la conduite d'une recherche et sur les objectifs recherchés.

Tableau 2 - Les arguments techniques et juridiques des éditeurs et les réponses des acteurs Couperin, CSPLA, Liber et militants pour l'ajout d'une exception au droit d'auteur pour le TDM

Argument des éditeurs	Que contestent les acteurs institutionnels ?
Un nouvel usage implique des nouveaux droits.	La pratique des TDM ne consiste pas pour les chercheurs à exercer un nouveau droit mais à poursuivre par des moyens technologiques modernes une activité ancienne de lecture savante et herméneutique des textes.
- Les contenus des bases de données sont sous droits et leur utilisation relève ainsi du régime du Code de la propriété intellectuelle. - Les bases de données sont protégées par un droit voisin au droit d'auteur (code L112 – 3 de la propriété intellectuelle).	- Cela présuppose que les données issues d'un TDM sont protégées par le droit d'auteur au même titre que leur source d'origine (article, ouvrage). - Pourtant, le TDM ne s'intéresse nullement à l'œuvre (agencement de mots selon une forme), mais seulement aux « mots » qui n'appartiennent à personne. - Les TDM dissolvent l' « œuvre » et ne se placent donc pas dans le périmètre couvert par cette législation. - Les bases de données sont protégées en tant que produits d'une activité de création (forme, structure, nature des contenus). Ce droit ne serait menacé qu'avec une tentative de sa republication de manière substantielle.
Les pratiques de TDM surchargent les serveurs	Des éditeurs Open Access, comme PLOS Opens, dont l'infrastructure est bien plus exposée à tous types d'usages par les internautes ont rapporté que l'impact des extractions automatiques par robots en vue de la fouille de données est négligeable et une augmentation future de la demande serait facile à gérer. Pour un éditeur comme Elsevier dont l'accès est soumis à des abonnements d'institutions cela devrait être encore plus facile.

Quelques outils proposés par les éditeurs en Open Acess

- Crossref Text and Data mining | http://www.crossref.org/tdm/index.html
- Outil de text-mining Bilbo sur Revues.org| http://leo.hypotheses.org/11655

1.3 L'*Open Data research* à l'Horizon 2020

Depuis la déclaration de Berlin sur le Libre Accès à la Connaissance en sciences exactes, sciences de la vie et sciences humaines et sociales, le mouvement pour l'ouverture des données scientifiques est l'objet des groupes de travail au sein de la Commission Européenne et au sein des institutions (centres de recherche, universités, bibliothèques, etc.). Ce groupe de travail tente de définir le cadre juridique et normatif, les infrastructures nécessaires, les politiques contraignantes ou incitatives de la mise en place de l'ouverture des données « brutes » de la recherche financée par des fonds publics.

En Europe, la Commission européenne lance le projet pilote du 16 décembre 2013[26], « Horizon 2020 » pour le libre accès aux données de la recherche issues des financements publics. Ce projet pilote est transversal aux principaux piliers du programme général Horizon 2020 développant des actions dans les domaines suivants 27 : technologies futures et émergentes, infrastructures de recherche, recherche et développement en TICs, science avec et pour la société.

Il s'agit d'une « expérience de libre accès de ces données » dans le cadre des nombreux projets financés par l'UE. « Pour la période 2014-2015, les travaux de recherche relevant du projet pilote seront financés à hauteur de 3 milliards d'euros. Fonctionnant par appel à propositions[28], l'initiative pilote vise à obtenir des informations concernant les pratiques existantes et les difficultés rencontrées et réunir des éléments suffisants permettant de décider des lignes directrices à suivre.

Le texte descriptif de ce projet est clair quant à la relation de ce mouvement avec le contexte général de l'ouverture des données publiques :

[26] CE, Communiqué de presse du 16 décembre 2013. En ligne : <http://europa.eu/rapid/press-release_IP-13-1257_fr.htm >
[27] Les différentes actions son décrites dans el portail français Horizon 2020. En ligne : <http://www.horizon2020.gouv.fr/cid81981/les-differents-instruments-regimes-financement.html>
[28] Tableau des appels à propositions d'Horizon 2020. En ligne : <http://www.horizon2020.gouv.fr/cid77090/tableau-des-appels-propositions-horizon-2020.html >

« Ce projet pilote «Horizon 2020» sur le libre accès aux données issues de la recherche est l'équivalent, pour l'information scientifique, de la stratégie en matière de libre accès aux données mise en œuvre pour les informations du secteur public: elle vise à élargir et à améliorer l'accès aux données générées par les projets, et leur réutilisation, au profit de la société et de l'économie. »

CADRE GENERAL DE L'INITIATIVE « PROJET PILOTE OPEN RESEARCH DATA » DU PROGRAMME EUROPEEN[29]	
Quelles sont les données concernées ?	- données et métadonnées nécessaires à la validation des publications (obligatoire) - autres données et métadonnées que le bénéficiaire a choisi de diffuser en accès ouvert : spécifiées dans le plan de gestion des données ou DMP - "Data Management Plan". - Certaines données peuvent être exclues du programme sous justificative explicité dans le DMP.
Quels domaines d'application ?	Les domaines définis par les groupes de travaux de la CE (Work Program) sont consultables dans les adresses : EC Participant Portal http://ec.europa.eu/research/participants/portal/desktop/en/home.html; et le Portail français du programme européen pour la recherche et l'innovation http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html
Quels objectifs ?	Déposer les données **dès que possible** dans des bases de données de recherche et permettre un accès gratuit et sans restriction à tout public pour les opérations suivantes : extraction, exploitation, reproduction et dissémination
Qui est concerné ?	- Projets validés par le programme - Projets soumis sur la base du volontariat (« opt in »)
Livrables attendus	- Proposition de projet - Plan de gestion ou DMP livré dans le six premiers mois de vie du projet

Plusieurs dispositifs ont été créés pour informer et apporter de l'assistance aux chercheurs et communautés porteurs de projets désirant répondre aux appels à propositions du programme européen. Par exemple :

[29] Source : Le libre accès aux publications et aux données de recherche. En ligne <http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html>

- Point de contact national ERC (European Research Council) en France sur le Portail français Horizon 2020[30] ;
- Portail du participant[31] de la CE ;
- « Information aux déposants »[32] sur le site de l'ANR.

Des initiatives institutionnels sont également en cours et devront se multiplier. Par exemple, des formations[33] destinées à préparer les équipes et chercheurs porteurs de projets à monter un projet européen dans le cadre des programmes Horizon 2020.

Le paysage général des politiques incitatives et de soutien à l'ouverture des données publiques et de la recherche a été traité de manière exhaustive dans l'étude réalisé par Rémi Gaillard, « *De l'Open Data à l'Open Research Data: quelle (s) politique (s) pour les données de recherche ? »*[34].

1.4 Les Humanités numériques

Les Humanités numériques sont, avant d'être une discipline transversale, une manière stratégique de positionner les SHS face au phénomène du web, face auquel les chercheurs ne peuvent plus rester indifférents.

Déjà familiarisés avec l'informatique depuis les années 70 (Linguistique computationnelle, puis dans les années 80, les Humanités computationnelles), Internet apparait à la fois possibilité d'ouverture à un public plus large, source heuristique, interface de modélisation à travers des outils divers (data visualisation, etc.), possibilité de partage de données, collecte de données (pour la sociologie par exemple), etc. Les Humanités numériques donnent un cadre et inscrivent les initiatives de production de données (résultats de recherches, numérisations, etc.), exploitation et diffusion des données de la recherche.

Les Humanités numériques se définissent comme :

[30] < http://www.horizon2020.gouv.fr/cid73935/le-point-contact-national-erc.html>
[31] < http://ec.europa.eu/research/participants/portal/desktop/en/home.html>
[32] < http://www.agence-nationale-recherche.fr/financer-votre-projet/informations-aux-deposants/>
[33] Université Sorbonne Nouvelle (Paris 3), «Horizon 2020 comment monter un projet européen ? » En ligne : <http://www.univ-paris3.fr/horizon-2020-comment-monter-un-projet-europeen--281899.kjsp?RH=1207746285942>
[34] V. Bibliographie.

- engagement des chercheurs qui se sentent concernés dans une réflexion portant sur l'impact ou les effets de la technologie sur leurs pratiques : modalités de recherche mais aussi apparition des nouveaux objets de recherche;

- convergence de communautés scientifiques diverses, intéressées par les pratiques, outils, objets transversaux pouvant émerger du numérique : encodage de sources textuelles, systèmes d'information géographique, lexicométrie, numérisation du patrimoine culturel, scientifique et technique, cartographie du Web, fouille de données, 3D, archives orales, arts et littératures numériques et hypermédiatiques, etc.

- surgissement de projets de numérisation, exploitation et diffusion des résultats de la recherche

- appel pour l'accès libre aux données et aux métadonnées. Celles-ci doivent être documentées et interopérables, autant techniquement que conceptuellement.

1.4.1 Un phénomène qui tarde à se développer en France

Comme le remarque Pierre Mounier, en France, les Humanités numériques ne trouvent pas encore la structuration de leurs activités au sein des universités, contrairement aux Etats-Unis où les Digital Humanities bénéficient de centres au sein des universités, à cheval entre les départements disciplinaires et les bibliothèques, composés de chercheurs, ingénieurs et designers.

En effet, il semble exister en France des barrières culturelles, politiques et économiques qui retardent ou empêchent de suivre la voie américaine à ce niveau :

- le corporatisme des métiers ;

- le poids de la structuration des disciplines ;

- du point de vue économique, le manque de financements pouvant permettre la création des centres d'Humanités numériques et le développement d'infrastructures ou de logiciels.

1.4.2 Des critiques sévères du « phénomène » voire de cette « mode »

Les *Digital Humanities*, telles que pratiquées aux Etats-Unis, sont par ailleurs l'objet de critiques et de réticences en France. On peut citer celles émises par Éric Guichard**,** Maître de conférences à l'Enssib, responsable de l'équipe Réseaux, Savoirs & Territoires de l'ENS et directeur de programme au Collège international de philosophie.

Cet auteur dénonce notamment l'opportunisme qui entoure l'institutionnalisation des *Digital Humanities* aux Etats-Unis et la recherche d'affichage médiatique d'un certain groupe de disciplines en SHS (histoire et littérature notamment), en vue d'obtenir des financements :

- le terme « humanités » serait pour lui un fourre-tout qui viendrait se superposer à « Sciences humaines et sociales, Arts et lettres » dont la définition et son sens scientifique pose déjà un véritable problème ;

- les Humanités numériques traduiraient un changement d'attitude des chercheurs en SHS, devenus moins soucieux des avancées conceptuelles que de la forme prise par celles-ci, de la présentation de résultats connus, ou de la visibilité du chercheur / de l'institution à travers le recours à des outils de pointe et des questions techno-textuelles (encodage, portabilité, etc.) ;

- les soucis technologiques auxquels se heurtent ces travaux sont particulièrement chronophages.

Un article récent paru dans *Slate* dénonce par ailleurs le phénomène de « mode » des *Digital Humanities*, « discipline » qui :

- coûte parfois huit ans d'investissement aux étudiants avant l'obtention d'un doctorat ;

- brouille les frontières entre disciplines ;

- créé une obsession technologique des chercheurs qui s'affichent en tant que *digital humanists*.

1.4.3 Quelle place pour les données dans ce débat ?

Le rapport de l'Institut Français réalisé par Marin Dacos et Pierre Mounier suggère quelques convergences entre la proposition générale des Humanités Numériques et la problématique des données de la recherche en SHS :

- accès aux données : « L'exigence d'accessibilité aux résultats de la recherche doit être étendue aux données de la recherche et pas seulement aux publications » ;

- évolution des pratiques de la recherche : « À tous les niveaux et à toutes les étapes de son déroulement, la recherche en sciences humaines est désormais immergée dans le numérique et ses usages. »

- nature des données : « (…) quel que soit le sujet d'étude, la recherche s'accomplit désormais systématiquement sur des objets numériques, qu'ils soient le résultat d'opérations de numérisation ou qu'ils soient nativement numériques. »

- évolution des environnements : « (...) les chercheurs sont aujourd'hui conduits à concevoir, construire, manipuler de grandes bases de données, bibliographiques, quantitatives, textuelles, d'images ou de sons ; ils sont conduits à travailler au sein d'équipes multi-localisées et en réseau, souvent à l'échelle internationale ; ils sont conduits à publier les résultats de leurs travaux, que ce soit les données ou les interprétations de ces données, sur le Web ouvert.

1.5 Web sémantique et Web de données dans le contexte des SHS

1.5.1 Quelques définitions de base

On désigne généralement par le terme de *Web sémantique* un ensemble de standards et de technologies développé par le W3C (l'un des principaux organismes de normalisation du web) visant à faciliter l'exploitation des données structurées, notamment en permettant leur interprétation par des machines.

Le *web de données* (*Linked Data* en anglais) consiste à exposer ces données structurées sur le web et les relier entre elles, ce qui permet d'accroître leur visibilité et leur réutilisation. Les bibliothèques produisent depuis longtemps dans leurs catalogues des données structurées et

contrôlées qui ont une valeur sur le web. Elles pourraient devenir des acteurs majeurs du web de données[35].

1.5.2 Une démarche des bibliothèques

En France, l'intérêt par les technologies du Web sémantique commence au milieu des années 2000. Il s'agit au départ d'un questionnement des bibliothécaires et professionnels de l'information face à la nécessité de faire évoluer l'architecture du système de catalogage des bibliothèques vers le modèle architectural du Web, c'est-à-dire, un modèle entièrement conçu sur le principe de l'interopérabilité :

« Le Web c'est avant tout un ensemble de standards qui permettent une dissémination de technologies partagées par tous les acteurs (producteurs de contenus, logiciels…), et indépendant des environnements matériels et logiciels. Le principe nous est devenu tellement naturel que nous n'en avons plus conscience, mais si on observe le phénomène à l'aune de l'histoire de l'informatique, on peut dire qu'aujourd'hui le Web est l'environnement le plus interopérable qui soit. » [36]

Les catalogues informatisés ont été conçus pour des usages limités que le phénomène Web a complètement bouleversés et bâtis avec des technologies difficilement conciliables avec l'architecture de réseau du Web. Les enjeux du Web sémantique pour les bibliothèques, aujourd'hui, consistent à réconcilier les normes des bibliothèques avec celles du Web, intégrant les données des catalogues dans un espace d'information global.

Cette démarche est le **décloisonnement des données** qui se trouvent dans les bases de données des bibliothèques et dont l'accès n'est pas praticable pour les moteurs de recherche du Web, grâce à des barrières technologiques mais aussi de structuration de l'information.

La BNF est très impliquée dans cette démarche qui consiste non seulement à adopter des technologies du Web développées par le W3C mais à faire un gigantesque travail de gestion de l'hétérogénéité des formats des données (MARC, Dublin Core, EAD, TEI) et de structuration et partage des données, rendant possible leur interprétation par les machines.

Le Web sémantique possède deux aspects généraux :

- Technique : respect des standards du « Web sémantique » et ouverture sur le «Web de données » (« linked data ») définis par le W3C.

[35] Source : BNF, « Les enjeux du web de données en bibliothèque»

[36] BERMES E. Le Web sémantique en bibliothèque, v. Bibliographie.

- Juridique : les données sont placées sous Licence ouverte autorisant la réutilisation libre, y compris commerciale, avec mention de la source.

Le Web sémantique en bibliothèque a comme objectifs :

- l'ouverture des données et la possibilité de réutiliser ces données dans d'autres projets (bibliothèques numériques), V. par exemple les entrepôts de données OAI-PMH de la BNF
- un accès aux données à travers les moteurs de recherche du Web
- la structuration de l'information selon deux modèles dont l'un est l'expression de l'autre : le modèle FRBR préconisé par l'IFLA pour la description bibliographique des ressources et son expression interprétable par les machines, le modèle RDF (Resource Description Framework), standard du W3C et son langage de requête SPARQL.

1.5.3 Le Web sémantique et les SHS[37]

Les questions du Web sémantique, mûries dans le contexte des bibliothèques, pénètrent depuis peu l'environnement de la production scientifique selon des problématiques semblables :

- questions d'accessibilité
- question de la pérennité
- plus grande visibilité
- partage et réutilisation efficaces
- placement sous licences libres

Mais dans un contexte différent :

- augmentation des pratiques d'auto-archivage par les chercheurs
- mouvement de l'open acess de la publication scientifique

[37] Les remarques à suivre proviennent de mes notes personnelles lors de l'intervention de Stéphane Pouyllau à la demi-journée du GFII (30 avril 2014) : "Isidore, modèle d'architecture informationnelle et technologique ancré dans le Web sémantique".

- multiplication des plateformes ouvertes en ligne placées sous les responsabilités des chercheurs et laboratoires de recherche

- initiatives œuvrant pour l'explicitation de la production scientifique, rejoignant la question de l'ouverture des données de la recherche (projet d'Open Science)

1.5.4 Huma-num et la plateforme ISIDORE

Actuellement, en France, Huma-num est le porte-parole de cette initiative, ayant formalisé à plusieurs occasions les objectifs du web sémantique pour la recherche en SHS .

Huma-num est une infrastructure de recherche dont l'objectif principal est celui de proposer des services, des moyens et des solutions aux chercheurs qui développent des programmes de recherche en SHS dans l'environnement numérique. Ces services s'adressent aux chercheurs qui se sentent concernés par les questions sur l'accès, la diffusion, le stockage et la pérennité des données de la recherche.

Huma-num se propose, à travers ses services, à œuvrer pour :

- une meilleure qualité des métadonnées des chercheurs : cette question peut encore heurter les pratiques de chercheurs pour diverses raisons relevant souvent de l'approche technique exigée ;

- une meilleure compréhension des enjeux du Web sémantique et de l'intérêt à travailler avec Huma-Num pour la fabrication, la diffusion et la réutilisation des données de la recherche (interopérabilité).

- Huma-num est la « tête nationale » de DARIAH (Digital Research Infrastructure for the Arts and Humanities) et travaille en synergie avec cette infrastructure européenne.

ISIDORE, plateforme développée par Huma-num, est construite à partir du modèle d'architecture informationnelle et technologique du Web sémantique. C'est un ensemble de services de collecte des métadonnées et d'enrichissement sémantique de ces métadonnées via leur mise en relation avec des ontologies et référentiels préexistantes. En plus, ISIDORE est un outil de « moissonnage » des métadonnées, des notices, des textes en version intégrale, des bases de données, des corpus, accessibles sur le Web et produits dans des standards ouverts.

1.5.5 Enjeux pour les SHS

Pour Stéphane Pouyllau, la question de l'enrichissement des données est au cœur même de l'activité scientifique et met en évidence une démarche qui tend à s'amplifier dans les communautés scientifiques en SHS et de façon générale :

- explicitation du processus de la production scientifique lui-même
- explicitation de chaque étape de production et de collecte des données

Cette démarche est aussi question de l'**interopérabilité** des données. Celle-ci ne peut être garantie que par un travail de fond de la description des données.

- l'enjeu est de dépasser la logique du « web des documents » et du simple signalement des documents dans des bases de données
- il y a une prise de conscience de l'intérêt à dépasser la logique des constructions de bases de données pour aller vers un modèle où les métadonnées des documents, plus la structuration de l'information, permettent de contextualiser les données et rendent possible **de relier les informations entre elles et non seulement de les mettre à disposition.**
- Cela est possible via les protocoles W3C
- Il y a aussi une prise de conscience des chercheurs de leur responsabilité dans la publication des données.

Pour S. Pouyllau[38], la pénétration du Web sémantique dans le SHS se fait plus rapidement depuis au moins quatre ans. L'avancement des initiatives dépend plus des personnes qui sont porteuses de projets que des disciplines à proprement parler.

Il constate néanmoins certains freins dont le plus important est le manque d'ingénieurs concepteurs (professionnels de l'info-doc, par ex.), au sein des laboratoires, qui développent des démonstrateurs et des prototypes.

---

38 POUYLLAU S. Web de données, big data, open data, quels rôles pour les documentalistes? Manuscrit auteur, publié dans "Documentaliste - Sciences de l'Information Vol. 50 (2013) 32-33", ArchiveSIC, version pré-print déposé le 18/03/2014. En ligne : <http://archivesic.ccsd.cnrs.fr/docs/00/96/08/53/PDF/Doc-SI_50-3-1_pouyllau-stephane_V1.pdf>

Le plus souvent, il n'y a pas assez de recul pour mener une réflexion sur les gains et les bénéfices d'entreprendre un si gros travail sur les données, de là l'importance des démonstrateurs qui peuvent les rendre sensibles.

1.6 Infrastructures de la recherche en SHS en France et Europe

Qu'est-ce qu'une infrastructure de la recherche ? Par cette notion sont compris : des installations, ressources, services dont la communauté scientifique a besoin pour effectuer des travaux de recherche dans tous les domaines scientifiques et technologiques.

Au-delà sa définition matérielle, une infrastructure de recherche se définit également par un ensemble de principes fondateurs qui caractérisent son rôle et sa situation à l'intérieur d'un système. Ainsi la feuille de route du MESR « Stratégie nationale : Infrastructures de recherche 2012-2020 »[39] définit les principes innervant les infrastructures comme :

- « Outil (ou dispositif) possédant des caractéristiques uniques identifiées par la communauté scientifique utilisatrice comme requises pour la conduite d'activités de recherche de haut niveau (…) »

- « L'infrastructure peut conduire une recherche propre, et/ou fournir des services à une communauté d'utilisateurs (…) »

- « L'infrastructure doit disposer d'une gouvernance identifiée, centralisée et effective et d'instances de pilotage scientifique. »

- « L'infrastructure doit être ouverte, accessible sur la base de l'excellence scientifique évaluée par les pairs au plan international ; elle doit donc disposer d'instances d'évaluation adéquates. »

- « L'infrastructure dispose d'un plan de financement et doit être en mesure de produire un budget consolidé. »

Concrètement, une variété de formes[40] est aujourd'hui reconnue par la communauté scientifique comme infrastructure de la recherche :

- infrastructure localisée, réseau de plateformes ;

[39] MESR, octobre 2012, p. 10. En ligne : <http://cache.media.enseignementsup-recherche.gouv.fr/file/TGIR/29/6/infras_def3_243296.pdf>
[40] Pour des exemples concrets consulter le document cité ci-dessus.

- infrastructure de recherche virtuelle et base de données ;

- collection, archives, bibliothèques scientifiques ;

- infrastructure à la base de réseaux humains de très haut niveau scientifique nécessaires, par exemple, pour la coordination d'un ensemble d'actions menées sur le terrain en lien avec des partenaires européens dans les domaines des Sciences biologiques et médicales.

Actuellement la question des infrastructures de recherche en France se pose à l'intérieur du programme plus large de création d'un espace européen de la recherche sous de critères définis d'excellence scientifique. Depuis une dizaine d'années, La France développe de manière soutenue une quantité d'instruments partagés entre les différents acteurs sur son territoire et participe également à des nombreux programmes internationaux européens et internationaux. Comme pour l'ouverture des données de la recherche, objet d'une réflexion visant à instaurer des dynamiques entre ces espaces nationaux et internationaux, il existe actuellement en France un effort d'intégration des programmes d'infrastructures nationaux et internationaux dans une même stratégie répondant aux critères supranationaux d'excellence.

Le cadre de collaboration entre la France et les réseaux d'infrastructures au plan européen et international est garanti par l'ESFRI [41] (European Strategy Forum on Research Infrastructures), dont la mission principal est de soutenir et guider les initiatives de normalisation entre ces infrastructures[42]. Ainsi, à titre d'exemple, pour les SHS, ESFRI a inscrit dans sa feuille de route le cadre de collaboration avec plusieurs projets européens: DARIAH[43] (Digital Research Infrastructure for the Arts and Humanities), CLARIN[44] (Common Language Resources and Technology Infrastructure), CESSDA[45] (Consortium of European Social Science Data), ESS[46] (European Social Survey) et SHARE[47] (Survey of Health, Ageing and Retirement in Europe).

Dans le contexte national les SHS comptent avec deux importants TGIR (Très Grands Équipements de recherche) : Huma-Num (déjà évoqué en 1.5.3), né en 2013 de la fusion des TGE (Très Grand Equipement) d'ADONIS et CORPUS, et PROGEDO[48] (PROduction et

[41] < https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri>
[42] ESFRI, Research & Innovation, *Infrastructures*, en ligne :
<https://ec.europa.eu/research/infrastructures/index_en.cfm?pg=esfri >
[43] < http://www.dariah.fr/>
[44] < http://clarin.eu/>
[45] < http://www.cessda.net/about/members.html>
[46] < http://ess.sciencespo.com/>
[47] < http://www.share-project.org/>
[48] < http://www.progedo.fr/spip.php?rubrique1>

GEstion des DOnnées) TGIR des sciences sociales. Ces infrastructures sont numériques et fonctionnent avec des politiques orientées vers la construction européenne d'ESFRI. Ainsi, actuellement, PROGEDO se charge du volet français de CESSDA, ESS et SHARE et Huma-Num de celui de DARIAH et CLARIN.

Deux autres infrastructures réseaux ont été reconnues sur le plan national : le réseau des MSH (Maison des Sciences de l'Homme), destiné à consolider des pôles territoriaux d'excellence, et NEFIAS[49] (Network for Internationalizing Advanced Science) destiné à créer des conditions d'échanges transversaux entre les institutions de recherche et réseaux internationaux.

Le Réseau Quetelet est une infrastructure à niveau national, membre du CESSDA, et une des composantes de PROGEDO. Il coordonne les activités d'archivage, de documentation et de diffusion des données en sciences humaines et sociales. Il s'organise en quatre unités de partenaires qui offrent l'accès à des multiples bases de données d'enquêtes, documents administratifs et données de gestion, provenant de la recherche mais aussi des organismes des statistiques publiques.

Ces quatre unités sont :

- Le CDSP (Centre de Données Socio-politiques) : enquêtes pré et post électorales et autres.
- CMH-ADISP (Centre Maurice Halbwachs- ADISP) : grandes enquêtes françaises.
- INED (Institut national d'études démographiques) : enquêtes produites par l'INED depuis 1945
- CASD (Centre d'accès sécurisé distant aux données) : données individuelles très détaillées de la statistique publique française (données sur les individus et ménages et données entreprises)

Les Sciences sociales comptent depuis 2011 avec l'EQUIPEX (Equipement d'excellente-*Investissement d'avenir*) DIME-SHS[50] (Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales). Il s'agit d'un équipement proposant trois instruments aux chercheurs pour produire, collecter et réutiliser des données. Trois types de données sont visés par ces instruments :

[49] <http://www.allianceathena.fr/nefias-network-internationalizing-advanced-science>
[50] < http://www.sciencespo.fr/dime-shs/>

⇒ Données quantitatives - DIME-SHS / Quanti : Un instrument pour les données quantitatives qui prend la forme d'un panel internet, ELIPSS (Etude Longitudinale par Internet pour les sciences humaines et sociales) ;

⇒ Données qualitatives - DIME-SHS / Quali : Un instrument pour les données qualitatives qui prend la forme d'un site web, BeQuali (banque d'enquêtes qualitatives) ;

⇒ Données du Web - DIME-SHS / Web : Un instrument pour les données du web qui offrira des outils pour constituer des corpus et pour les analyser.

1.7 Les données de la recherche

Les données sont aujourd'hui thématisées dans des univers différents et par des acteurs variés. Pour chacun de ces univers des problématiques différentes se posent, liées à leur objet et aux finalités d'usages prévus pour ces données. Ces univers contribuent aujourd'hui à la formalisation des enjeux relatifs aux données dans les secteurs de la recherche. Malgré la différence d'approche, des questions centrales communes sont partagée, notamment les questions de signalement et des modalités d'accès aux données.

Univers	**Type de données concerné**	**Principaux enjeux**
Institutions et communautés scientifiques	Données de la recherche, métadonnées descriptives	Plan de gestion de données, traitement des données, partage, publication
Bibliothèques et professionnels de l'information et documentation	Métadonnées, données de la recherche	Enrichissement sémantique (web sémantique) et décloisonnement des données ; protocoles d'échanges de données, p. ex. OAI-PMH (Open Archives Initiative - Protocol for Metadata Harvesting) ; multiplication des sources (intégration des données de

		chercheurs) et numérisation
Editeurs scientifiques	Données de la recherche	Modèles éditoriaux : revues « augmentées », *data journals*, standards de citation, périmètre d'ouverture

Qu'est-ce qu'une « donnée de la recherche » dans un sens et un périmètre bien précis ? Selon la définition, très souvent citée, de l'Organisation de Coopération et Développement Économiques (OCDE) il s'agit de :

« Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. »[51]

Data.bris Project, développé par JISC et l'Université de Bristol, propose une autre définition :

« Les données de la recherche sont des informations impliquées directement dans la recherche scientifique, qu'elle soit ou non financée sur des fonds publics. Les données de la recherche sont souvent agencées et formatées visant leur viabilité à des fins de communication, interprétation et traitements. Dit d'une manière plus simple, les données de la recherche sont toutes les informations que vous utilisez dans votre recherche comme parties constitutives de celle-ci. »[52]

En croisant ces deux définitions, les données de la recherche seraient :

- Les sources principales ou constitutives de la recherche ;
- des informations ;
- des enregistrements factuels.

[51] OCDE, Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics, http://www.oecd.org/fr/science/sci-tech/principesetlignesdirectricesdelocdepourlaccesauxdonneesdelarecherchefinanceesurfondspublics.htm

[52] *What's count as research data ?*: « Research data is information that is involved directly in funded or unfunded research activities. Research data is often arranged or formatted in a such a way as to make it suitable for communication, interpretation and processing. Put more simply, research data is all of the information that you use as an integral part of your research. » ; http://data.bris.ac.uk/research/bootcamp/data/

Les données de la recherche serviraient à des fins de validation de la recherche, de communication scientifique, d'interprétation et de traitements (par exemple, traitements automatiques).

Deuxième Partie : entretiens exploratoires

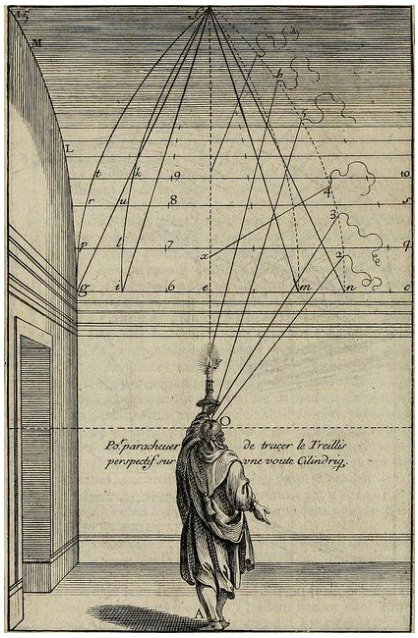


Figure 3 - "Moyen universel de pratiquer la perspective sur les tableaux, ou surfaces irregulieres : ensemble quelques particularitez concernant cet art, & celuy de la graueure en taille-douce" (p.79) par A. Bosse (1653) – Source : Flickr The Commons

Présentation

Pour nous servir d'une métaphore picturale, nous dirions qu'une fois constitué le cadre thématique *macroscopique* de cette étude sur les données de la recherche, il s'agit à présent, dans cette deuxième partie, de le mettre en perspective en l'adaptant au regard du chercheur.

En effet, une des constatations principales de l'étape de veille qui a précédé les entretiens a été la difficulté de composer un tableau clair, à dimension réduite, qui permette de saisir concrètement la manière dont les chercheurs sont en train de répondre à ces nouveaux contextes et environnements de travail et aux défis imposés par la production grandissante des données numériques.

L'avènement du numérique comme contexte généralisé de travail dans les SHS représente l'ouverture à des multiples possibilités d'usages et enrichissement des travaux de recherche. Il ne s'agit plus simplement d'un accès optimal aux formes de présentation de ces travaux (résultats de la recherche) à travers la numérisation des publications scientifiques ou, comme c'est le cas aujourd'hui, à travers le format numérique natif de la communication scientifique.

Les nouvelles opportunités offertes par le numérique concernent, pour emprunter un terme utilisé par un des chercheurs interrogé, *les coulisses* de la recherche dont les données constituent les pièces fondamentales et recèlent une valeur (épistémologique, économique) qu'il importe de conserver à long terme afin d'ouvrir des nouveaux usages à la communauté scientifique.

Dans ce contexte, l'étude présente s'intéresse à examiner dans les entretiens les questions suivantes :

- Quelle est la réceptivité des chercheurs en SHS aux enjeux relatifs à ces nouvelles opportunités pour la recherche, ouvertes par le numérique ?
- Quel rôle les données peuvent-elles jouer à côté des formes courantes de contribution et communication scientifique ?
- Que font les chercheurs aujourd'hui pour préserver et valoriser leurs données ?
- Quelles sont leurs attentes et craintes ?

Les entretiens exploratoires réalisés auprès de 53 chercheur(e)s dans 18 disciplines[53] ont été l'opportunité de connaître de près le point de vue et les pratiques de ces chercheur(e)s concernant les données de la recherche. Ces échanges, menés comme des conversations assez libres (semi-dirigées), ont apporté un contenu concret aux problématiques évoquées dans la première partie en même temps qu'ils ont permis de relativiser leur portée et de les comprendre à l'intérieur d'un contexte pratique. A travers le regard plus fin de ces entretiens il a été possible de dégager des définitions, des exemples de production, d'utilisation et de partage de données, des problèmes particuliers liés à un certain type de données et les motivations ou freins à leur diffusion. Ces entretiens ont été aussi l'occasion d'évoquer des problèmes épistémologiques surgissant par les transformations des méthodes et outils (les nTICs) des SHS comme, par exemple, le Data Mining.

L'élaboration et la structuration de cette analyse a été en grande partie conçue en parallèle à la conduite des entretiens. Nous avons pu compter a posteriori sur l'exemple et le cadre de deux travaux fondamentaux, tous deux anglo-saxons, *Managing and Sharing Research Data – A Guide to Good Practice* et le rapport de 2008 produit par Alma Swam *To Share or not to Share:Publication and Quality Assurance of Research Data Outputs* basé sur 100 entretiens avec des chercheurs tous domaines et disciplines confondues[54].

☛ **Nota bene : tout au long de ce mémoire les chiffres entre [ ] remettent aux numéros des entretiens consultables dans l'annexe 5 à la fin du document.**

[53] Trois entretiens ont été réalisés dans le domaine des Sciences de l'information et de la communication (SIC) qui ne sont pas, traditionnellement, rangées parmi les SHS. Il nous a paru important d'inclure la réflexion de ces chercheur(e)s, d'une part, en raison du cadre de cette étude (INTD) et d'autre part en raison de la croissante interaction des SICs avec les SHS (SHS), comprise dans ce qu'on appelle aujourd'hui les Humanités Numériques.
[54] Pour la référence complète de ces documents, v. Bibliographie à la fin de ce mémoire.

1 Définir et classer les données de la recherche

Cette étude exploratoire a posé comme objectif d'approcher le contexte actuel des pratiques des chercheurs en SHS sous l'optique particulière du sujet de ce mémoire, à savoir, les données de la recherche. Il importe donc, avant tout examen de ces pratiques, de bien convenir sur le sens de ce terme, « donnée de la recherche », et de chercher à en définir les contours. Compte tenu du caractère empirique de la démarche d'analyse d'entretiens, une attention particulière est portée pour éviter de superposer des notions générales - dont on a pu prendre connaissance dans la première étape de cette étude - à une description plus factuelle des informations livrées par les chercheurs.

Dans les sections du chapitre présent, nous proposons une rentrée en matière et une confrontation avec la multiplicité de sens attribués aux données par les chercheurs, avançant progressivement vers la constitution d'une typologie des données et vers une identification des enjeux et problèmes relatifs à leur diffusion.
Les étapes qui jalonnent cette analyse, décrites ci-dessous, éclairent différemment, mais de façon complémentaire, la question des données de la recherche en SHS :

En premier lieu, nous analysons les données par groupe disciplinaire et essayons de mettre en lumière quelques spécificités liées aux pratiques de ces disciplines dans leur réceptivité au sujet. Cela nous permettra aussi d'avancer vers une typologie de ces données (1.2.3). Il faut préciser qu'il s'agit d'un exercice de simplification qui a l'inconvénient de passer outre à la transdisciplinarité, très présente pourtant dans ces entretiens, et d'opérer dans une optique généraliste des SHS[55]. L'avantage de se prêter à cette simplification est qu'elle permettra par la suite de dégager des éléments dépourvus d'ambiguïtés, écartant un certain nombre de problématiques qui ne relèvent pas de la sphère de notre sujet.

En deuxième lieu, nous essayerons de réajuster ces éléments à la réalité plus complexe des pratiques de recherche des chercheurs interrogés (v. 1.3). A travers quelques exemples de schémas[56] de processus de production et utilisation des données dans la recherche nous

[55] Nous avons dû, dès le départ fonctionner à travers les grandes nomenclatures des SHS tout en sachant que le champ d'activité des chercheurs s'inscrit dans la trans-ou inter-disciplinarité et que les nomenclatures sont aujourd'hui source de réflexion en vue d'un alignement à des standards internationaux (V. http://www.obs-ost.fr/sites/default/files/epubliOST_nomenclaturesSHS_disparitesNotables_NCR6_sept2014.pdf) . Mais au vu des limites et objectifs de cette étude et du nombre d'entretiens réalisés, il nous a paru peu pertinent d'atteindre cette granularité si fine.
56 Nous attirons l'attention du fait que les schémas ici présentés ne sont pas ce qu'on appelle couramment « cycles de vie des données », mais des simples modèles des processus de recherche. Des exemples de cycle de vie applicables aux préparations des plans de gestion des données sont présentés en Annexe.

essayerons de montrer que la question d'une diffusion ou de partage des données se pose différemment selon les moments de ces processus. Ces schémas sont constitués uniquement à partir d'éléments fournis par les entretiens.

Finalement, un dernier moment systématise et généralise les résultats de nos réflexions précédentes. Les différences disciplinaires seront alors laissées de côté et les données seront analysées en fonction des problématiques et des enjeux évoqués par les chercheurs au respect d'une éventuelle diffusion.

1.1 Une multiplicité de données et de sens

Interrogés sur ce qui peut être compris par « donnée de la recherche » dans le cadre de leurs pratiques, les chercheur(e)s se sont prononcé(e)s très différemment en fonction de leur discipline, de l'objet de leurs recherches, et méthodes employées pour celles-ci, ou de leur familiarité avec les nTICs appliquées à la recherche.
Le caractère à la fois *objectif et technologique* de cette notion rendait parfois difficile à quelques personnes interrogées de l'appliquer à leur mode de travail, considérant ce type de réflexion plus proche et « héritée » des préoccupations dominant les sciences de la vie ou les sciences exactes. C'est le cas des chercheurs en philosophie, par exemple, dont les étapes de travail sont inséparables au processus d'écriture traversé par une démarche originelle interprétative, critique ou créative.
La connaissance des enjeux et débats actuels concernant la communication scientifique a également joué un rôle important dans la manière d'approcher le sujet. En effet, nous avons pu constater que la question des données était très souvent ramenée au terrain des derniers débats, très médiatisés, sur les conditions d'accès aux publications scientifiques. Au point où, parfois, les données étaient identifiées aux publications et vice versa, révélant ainsi une certaine confusion[57] ou imprécision du terme « donnée »[58].
L'arborescence ci-dessous (fig.4) illustre cette hétérogénéité de réponses avant toute tentative de typologie. On y voit notamment la présence de données aussi différentes que « données statistiques » et « archives orales » ou, au contraire, des termes différents pour qualifier des données qui semblent être du même type : « données de terrain », « donnée brute », « observations de terrain ».

[57] En effet, les publications sont exclues des définitions « officielles » des données de la recherche. V. Première Partie, 2.
[58] Nous verrons que pour certaines disciplines, comme la philosophie, les questions touchant les publications en revue ou sous autre format éditorial supplantent une réflexion sur les données de la recherche.

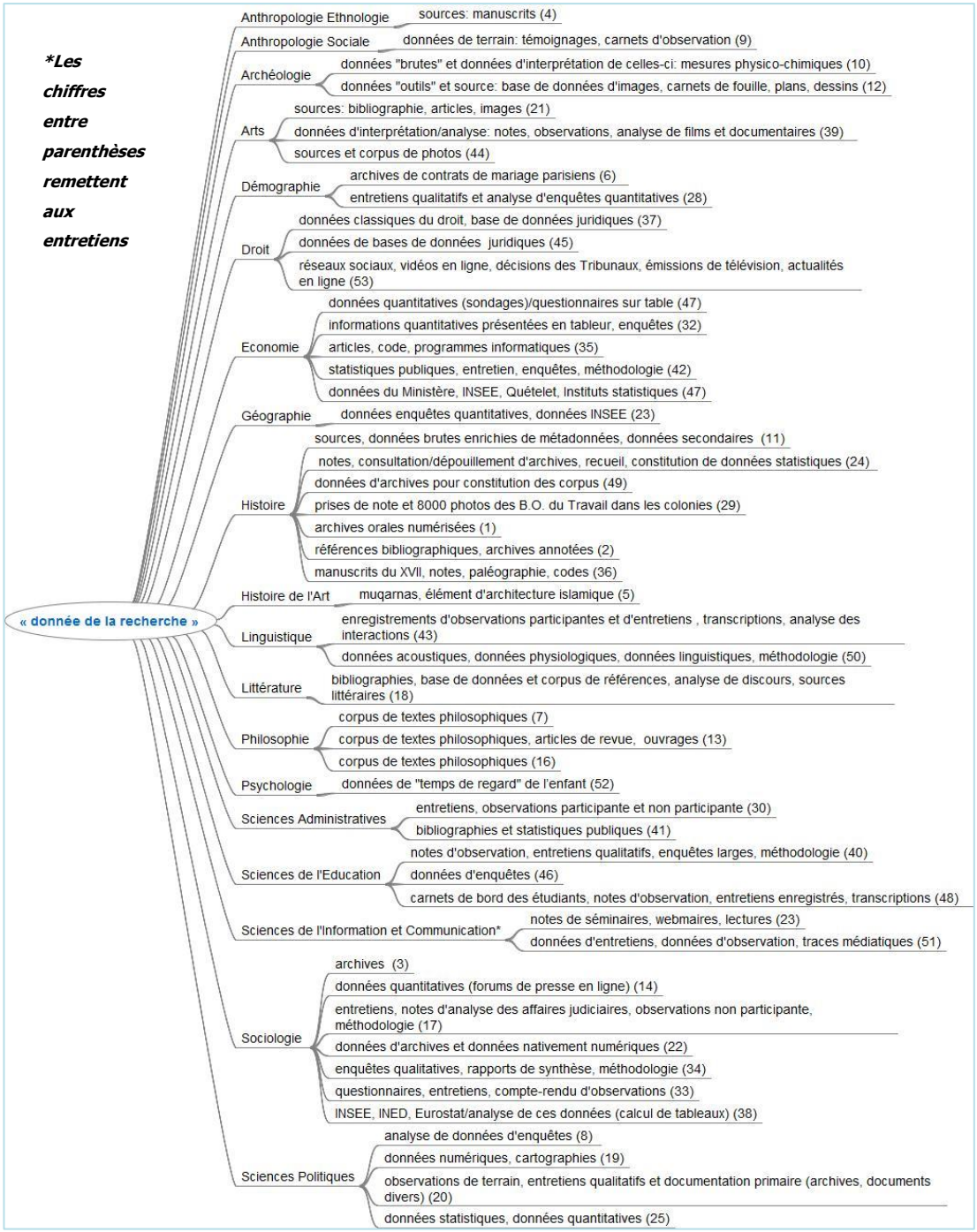


Figure 4 - Réponse à la question: « Dans le cadre de votre pratique, qu'est-ce qu'une donnée de la recherche[59]? »

Certaines données ainsi évoquées semblent présenter des points semblables - que ce soit d'un point de vue formel que de celui méthodologique - et pouvoir ainsi être qualifiées par « type » de donnée. Réaliser une typologie présente un double intérêt : être capable de comprendre l'approche particulière des disciplines face à la question des données de la

[59] Cette arborescence reprend les éléments du Tableau 1, Annexe 4.

recherche et en même temps avoir des bases concrètes pour aborder le genre de difficultés et enjeux qu'une diffusion de ces données entraînerait.

1.2 Des approches disciplinaires typiques à la question des données

Au cours de l'examen de ces entretiens nous avons pu repérer deux approches représentatives des différentes méthodes pratiquées en SHS et qui déterminent la manière dont les données seront thématisées par les chercheurs. Deux groupes ont été ainsi constitués (v. schéma en page suivante).

1.2.1. Disciplines à approche herméneutique et textuel : les sources, les bibliographies et les « outils » à la recherche

Le premier groupe de disciplines (V. schéma page suivante) auxquelles appartiennent les chercheurs interrogés se caractérise par des méthodes de travail d'interprétation et exégèse de textes (juridiques, historique ou philosophique).

Pour ces disciplines la question de ce que c'est qu'une donnée de la recherche ne va pas de soi car, dans l'essentiel de leur activité, les étapes de réflexion et construction d'un objet scientifique sont inséparables de la construction du *texte* et visent une forme aboutie du discours à travers la publication (l'ouvrage ou article).

En conséquence, les chercheurs de ces disciplines ont identifié les données de la recherche tout d'abord comme des données **matériaux** - sources, corpus de textes philosophiques, textes juridiques, articles et ouvrages scientifiques - sur lesquels s'*appuient* leurs recherches.

- **En Histoire : des sources, des notes, du matériel bibliographique et une réflexion concernant la pertinence du terme « donnée »**

La réflexion des historiens sur une définition de « données de la recherche » passe par leur qualification en tant que **sources**. Les historiens travaillent essentiellement avec ce type de données qu'ils annotent, décrivent, sélectionnent (création de corpus), expliquent et interprètent. Ces étapes de travail sur les sources, imprégnées de réflexions personnelles et très spécifiques au sujet traité ne sont pas envisagées comme des « données » et n'ont pas de vocation à être transmises à d'autres chercheurs (v. par ex. entretien 2 et 24).

« A son sens, la question des données de la recherche doit avant tout se comprendre comme celle des sources accessibles aux chercheurs. L'enjeu à ce niveau est très important : le nombre, la qualité et la

facilité d'accès aux sources déterminent grandement la qualité du travail des chercheurs (…). » (Histoire, entretien n°2)

Ce contexte général possède des situations particulières où d'autres types de données interviennent dans la recherche : ainsi pour une historienne des archives orales [1] et pour une historienne de l'art [5] utilisant des méthodes de modélisation, comme les SIG[60].

- **En Philosophie, une réceptivité faible à la question des données et forte à la question des publications en revue et d'accès aux sources**

Les chercheurs interrogés travaillant dans le domaine de la philosophie et esthétique ne se sentent pas concernés directement par les enjeux liés aux données, car l'essentiel de leur pratique de recherche n'est pas profondément modifié par l'avènement du numérique. Certes, le numérique a changé profondément les formes d'accès aux sources et les formes de diffusion de résultats, mais les chercheurs interviewés la production scientifique est toujours un travail d'élaboration conceptuelle qui garde son indépendance aux formes et supports technologiques.

Mais il existe également une impression de perdre de la vitesse par rapport à la réflexion sur les nouvelles TICs : ainsi de cette philosophe qui dit n'avoir pas les « compétences » ou de réaliser un travail « artisanal » [13].

Ce qui est intéressant de noter, et ceci est vrai pour les philosophes qui ont consenti à l'entretien, c'est qu'ils estiment produire autre chose que des « données » : du texte, des articles, des œuvres, des communications. Le fait que ces documents puissent devenir des données numériques reste contingent.

Nous sommes aux antipodes des Humanités Numériques et leur invitation à embrasser les nouvelles technologies non pas comme des simples « outils » appliqués à un objet ou à transformer un objet qui leur demeure étranger, mais comme possibilité pour les sciences humaines d'aboutir à un accord outil-méthode, cela impliquant bien entendu que la machine soit comme dit l'historienne de l'entretien n° 36 «des extensions du travail intellectuel du chercheur ».

Ces chercheurs ont donc identifié pour cette discipline deux enjeux principaux pour la recherche, tous les deux liés à la numérisation : l'accès aux corpus de textes philosophique et l'accès aux publications d'anciens numéros de revue.

[60] Systèmes d'information géographique.

1.2.2 Disciplines à approche de terrain et expérimentale

Le deuxième groupe réunit les disciplines travaillant surtout dans la production des données. Pour les chercheurs, ayant l' « habitude » des données, comme ceux qui travaillent avec des grandes enquêtes, les données constituent des informations brutes collectées suite à une problématique de départ.

- **En Anthropologie, Archéologie, Linguistique et Psychologie : des données de terrain et expérimentales**

  Les chercheurs de ces disciplines vont mentionner notamment l'importance du travail de terrain pour la réalisation de leurs recherches. Différents types de collecte sont mentionnés : paléographies, observations, enregistrements, prises de mesures. Les données ont un caractère très objectif, proches à celui des sciences « dures ».
- **En Sciences Administratives, Démographie, Economie, Géographie, Sociologie, Sciences Politiques, SIC et Sciences de l'Education : des données quantitatives, qualitatives et statistiques**

  En Economie, les chercheurs ne produisent que très rarement des données et utilisent les données statistiques et micro-données (les données à granularité plus fine utilisées par l'INSEE[61], par exemple) produites par des instituts de statistique nationaux et internationaux.

  C'est le cas également pour les Sciences administratives, mais les chercheurs de ces disciplines réalisent également des travaux de terrain et sont ainsi producteurs de données qualitatives.

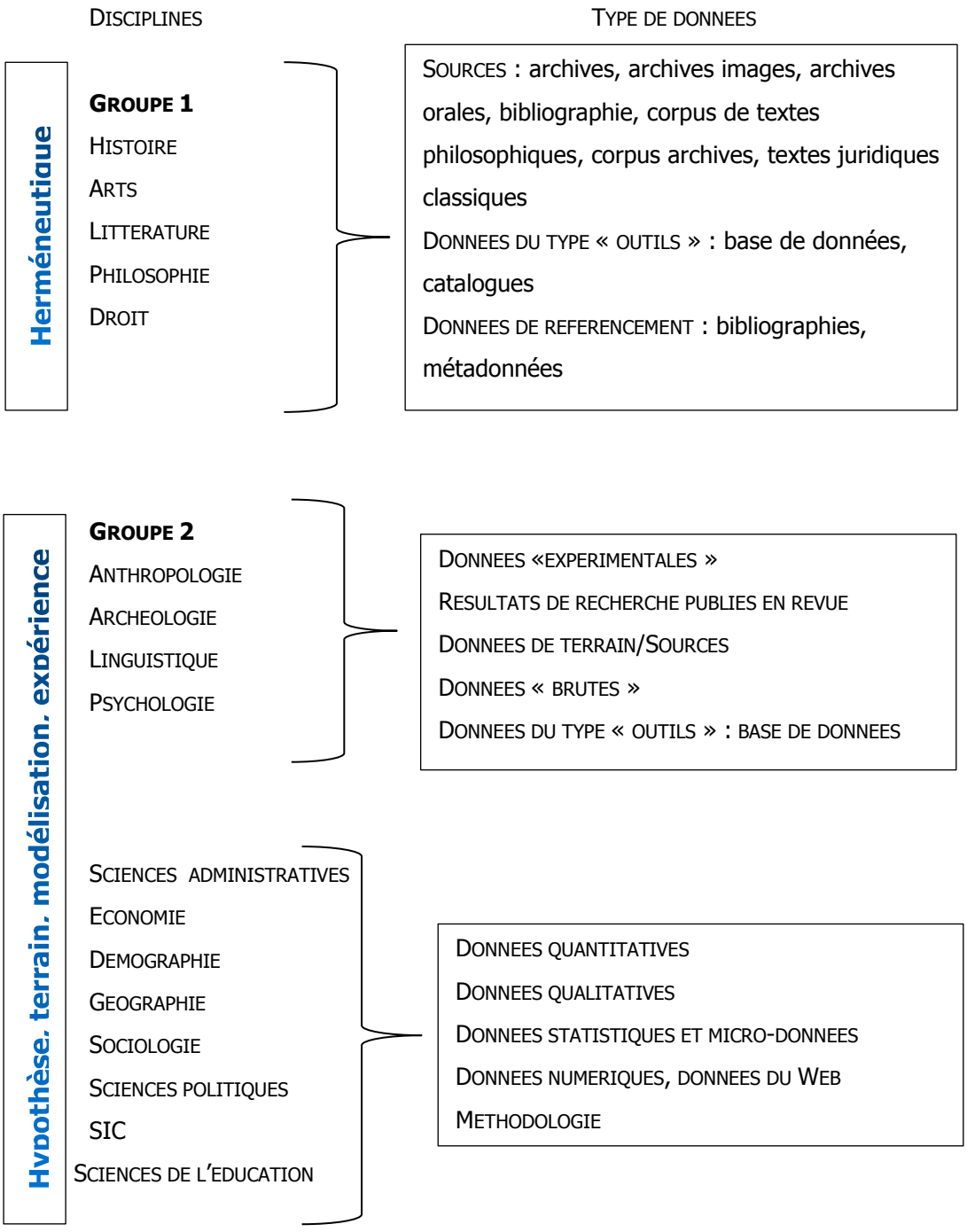
  En Démographie et en Géographie les chercheurs utilisent les données d'enquêtes quantitatives et qualitatives (en moindre mesure) des bases de données de l'INED mais, en Démographie, ils produisent également des données par élaboration d'enquêtes.

  En Sociologie, les chercheurs ont mentionné une grande variété de données, leurs domaines de recherche étant souvent transversaux à d'autres domaines. Ainsi par exemple de ce chercheur travaillant sur l'expertise psychiatrique judiciaire et mêlant une approche de terrain (entretiens) à une étude historique de l'expertise psychiatrique (archives).

  En Sciences politiques c'est sans doute les données d'enquêtes – quantitatives et qualitatives – en plus des données statistiques, qui constituent la matière principale aux recherches évoquées dans ces entretiens.

[61] INSEE, L'accès aux micro-données et la gestion de la confidentialité dans quelques INS européens <http://www.insee.fr/fr/themes/document.asp?reg_id=0&id=2151>

Les Groupe de disciplines par type de données



1.2.3 La distinction entre les données utilisées et les données produites

En nous appuyant sur les éléments mis en évidence à ce stade d'analyse, un premier classement des données du type « statique » devient possible selon deux critères : d'un côté, les chercheurs *produisent* des données au cours de la recherche, de l'autre ils *utilisent* des données de type « matériau » pour la recherche.

Sans aller trop loin dans la complexité de termes qui devraient caractériser une étude approfondie des étapes de la recherche, nous comprenons le classement « statique » comme une organisation formelle des données faisant abstraction du fait qu'une recherche implique le plus souvent une dynamique d'interaction entre phases de production et phases d'utilisation des données (nous y reviendrons en 1.3) et du fait qu'une donnée « produite » par un chercheur peut en constituer un « matériau » par un autre.

La distinction entre ces deux grandes catégories de données est présente dans un grand nombre d'entretiens :

En **Histoire** :

« La notion de « données de la recherche » peut s'entendre de deux manières au moins, soit qu'il s'agisse de nommer les données sur lesquelles s'appuie la recherche ou les données produites par la recherche. Dans le cadre du tournant numérique des SHS, les données produites par la recherche bénéficient d'une nouvelle exposition et de possibilité d'accessibilité, d'enrichissement et d'interopérabilité accrue. » (Entretien n° 11)

« Les données pour les historiens sont essentiellement les sources archives, des données qu'on pourrait appeler, dans un but de simplification, « objectives ». Il y a, aussi, les données produites par les historiens, à savoir des données d' « interprétation », transformations des premières par le travail herméneutique de l'historien. » (Entretien n° 49)

En **Arts**:

« Pour elle il y a toute de suite deux types de données de la recherche : d'une part les données qui lui servent pour la réalisation de sa recherche : références d'articles, travaux des collègues, images documentées, communications des colloques, dialogue avec les collègues. D'autre part, les données de son terrain de recherche, par exemple, un corpus de 6000 photos qui lui ont été transmises par un photographe amateur. Bien qu'elle ne les ait pas produites elle-même, elle considère qu'elles sont les données propres de sa recherche, matériaux sur lesquels elle va réaliser un travail intellectuel. (Entretien n°44) »

En **Sociologie** :

« Il comprend tout d'abord « données de la recherche » comme celles produites par les disciplines des SHS adoptant une méthodologie quantitative. » (Entretien n°14)

En **Économie :**

« Par « données » elle comprend tout d'abord des informations quantitatives, généralement présentées dans un tableur (individus en ligne et variables observées en colonne). Mais le terme données « de la recherche » sous-entend des données « produites », ou utilisées à titre secondaire, dans un but de production de savoir scientifique. Ces dernières impliquent dès lors des étapes d'analyse et d'interprétation. » (Entretien n° 32)

En **Sciences politiques** :

« D'un point de vue très précis, il ne produit pas de données mais utilise des bases de données fournies par l'Insee, l'Ined ou Eurostat, soit en accès libre, soit sur demande (par exemple via le réseau Quételet). Au sens large, il produit des données en analysant ces bases brutes, par exemple, en calculant des tableaux. » (Entretien n°38)

En **Droit** :

« En tant que juriste elle ne produit pas, à proprement parler, des données mais utilise celles mises à dispositions sur des bases de données juridiques, comme Legifrance.fr. Ces données sont essentiellement des textes juridiques, lois, conventions collectives. » (Entretien n° 45)

Nous avons examiné cas par cas les données produites et utilisées par les chercheurs et avons obtenu la cartographie suivante (Fig.5):

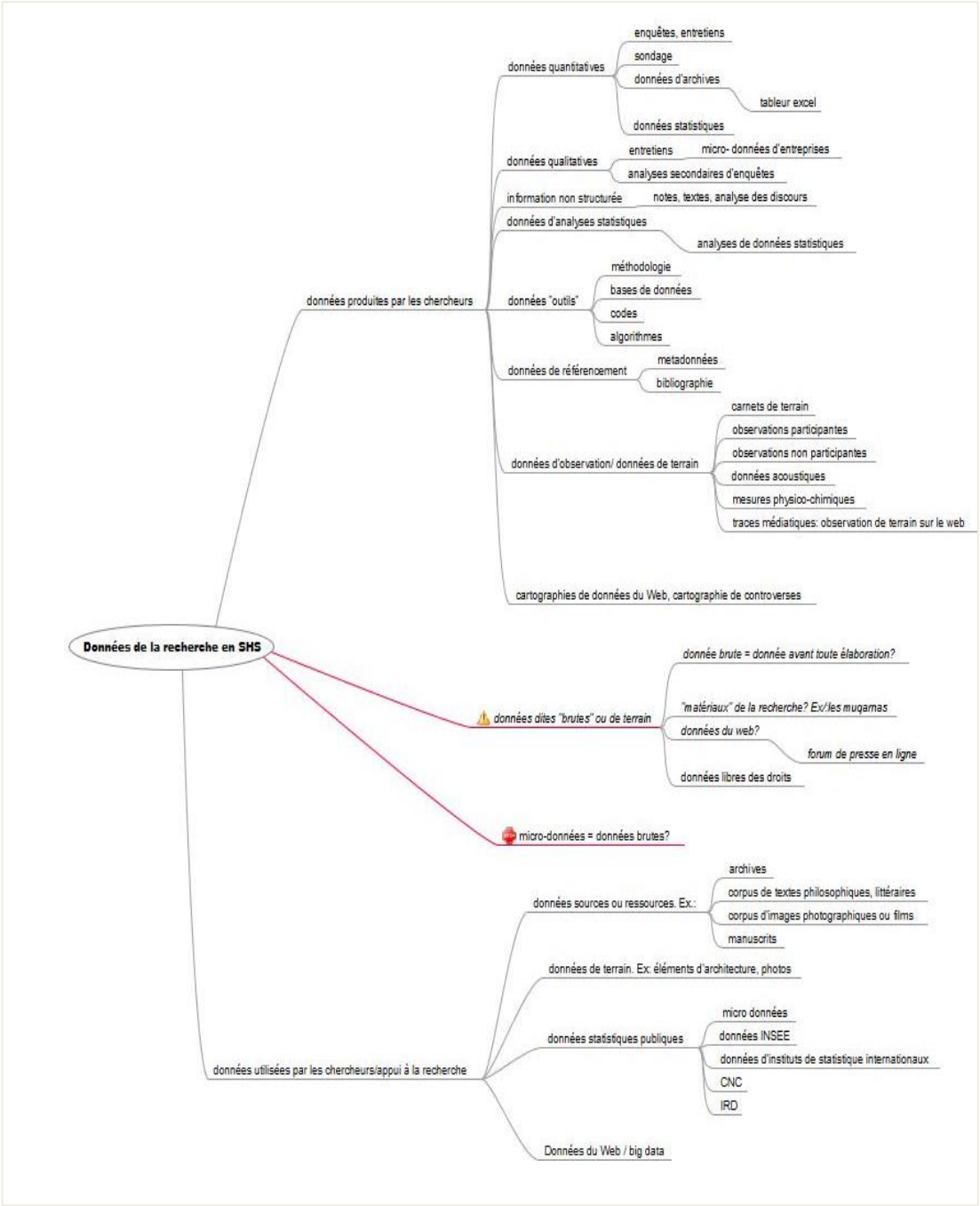


Figure 5 - Type de données produites et utilisées par les chercheurs interviewés

Suivant le sens suggéré par les chercheurs, les données « sources » ou « matériaux de la recherche » peuvent être définies comme :

- Des documents « sortis » du circuit des usages, relevant des archives [1, 3, 22, 24, 26, 49]

- Des documents faisant partie du circuit de la diffusion des résultats de la recherche à travers un format éditorial abouti plus ou moins visible : articles de revue, ouvrages, littérature grise [24, 33]

- Des sources matérielles ou primaires : corpus de textes philosophiques, objets, éléments d'architecture, photos, tweets [5, 7, 12, 16, 29]

- Des données de référencement : bibliographie [11, 18, 21]

- Des données statistiques produites par l'Institut national de statistiques et études économiques (INSEE) et internationaux et micro-données [17, 25, 38, 41, 42, 47]

Alors qu' « utiliser » des données est une notion qui bénéficie d'une clarté grâce à son caractère objectif (les archives *préexistent* à la recherche comme les données statistiques sont *accessibles* dans les bases de données d'INSEE), *produire* des données est un terme qu'exige quelques précisions. En effet, produire des données est une notion large qui se comprend d'au moins deux façons :

- Opérer des transformations sur des données « brutes » ou non structurées, par exemple, numériser des archives, les documenter et les enrichir sémantiquement en les structurant dans une base de données [11]. Un autre exemple, transformer (recalculer) des données statistiques brutes de l'INSEE [38] ou réaliser des modélisations sur des données du web à travers des cartographies [19].

- « Provoquer » et collecter des données à travers un travail de terrain, l'exemple typique étant la collecte de données dans la réalisation d'entretiens, cité dans plusieurs entretiens en Sciences politiques et Sciences de l'éducation, par exemple. Sont inclues ici également les données produites par démarches expérimentales, en psychologie, par exemple.

La question qui se pose est alors : quelles données peut-on considérer comme données de la recherche en SHS et qui seraient concernées par des programmes institutionnels d'ouverture ? Les données du type source sont-elles concernées ?

Pour rappel, la définition de l'OCDE en parallèle à celle de l'Université de Bristol, de « données de la recherche » est :

« Enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme **sources** principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour **valider** des résultats de recherche. »[62]

Selon cette définition, les données de la recherche ont valeur de **preuve** et de **justification** et doivent être communiquées ou rendues accessibles à des fins de vérification. Les corpus d'archives sur lequel un historien base sa recherche doivent donc pouvoir figurer en tant que données de la recherche au même titre que les données « brutes » d'enquêtes quantitatives. Mais la définition de l'Université de Bristol (voir Première partie, 1.7) apporte des nouvelles précisions,

« Les données de la recherche sont souvent agencées et formatées visant leur viabilité à des fins de communication, interprétation et traitements. »

Les données de la recherche en SHS correspondent aux sources et aux données produites dans la recherche dans la mesure où il est possible d'y identifier des actions visant une finalité de partage. Les sources et les données brutes ne seront ainsi des « vraies » données de la recherche que si elles sont agencées *délibérément* en fonction des fins de partage et utilisation poursuivies. Les données produites et les données sources doivent, toutes les deux, passer par des formes d'élaboration avant de pouvoir être considérées comme « données de la recherche », dans le sens compris par les définitions officielles qui orientent notre approche. Cette conclusion nous semble élargir le champ des potentielles données de la recherche en SHS, à la fois qu'elle permet de bien visualiser la taille de l'enjeu à venir pour les chercheurs et documentalistes.

1.2.4 Les données brutes

Les données « brutes » ont été qualifiées dans ces entretiens comme des données avant toute élaboration ou transformation intellectuelle et technique. Pour certains chercheurs les données brutes se distinguent des « données de la recherche », ces dernières considérées alors comme une élaboration des premières ou des étapes de pré-analyse de celles-ci. Comme le nom l'indique, « brute » sera à la fois la « chose » même de la recherche : un élément architectural d'une mosquée, par exemple, mais aussi un terrain d'observation ou un document d'archive.

[62] OCDE, Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics, http://www.oecd.org/fr/science/sci-tech/principesetlignesdirectricesdelocdepourlaccesauxdonneesdelarecherchefinanceesurfondspublics.htm

« Il importe toutefois de distinguer la « donnée brute », le texte d'archive lui-même, qu'il qualifierait plutôt de « données de terrain », des données de la recherche qui correspondent plus à un travail de pré-analyse. » (Histoire, entretien n°27)

Les données brutes sont aussi la quantification des faits observables, collectées ou provoquées (par expérience ou sondage).

« En ce moment, il travaille, conjointement à d'autres membres de son laboratoire, à la collecte de données issues des contrats de mariage parisiens auprès des archives notariales de Paris. Ce travail est quasiment manuel, la plupart de ces archives n'ayant pas été numérisée et se trouvant dans des microfiches. Les données sont ensuite intégrées dans une base de données simple, réalisée sur Excel, et serviront à la réalisation de graphiques et à l'élaboration d'analyses statistiques. » (Démographie, entretien n°6)

Les données brutes, ont également provoqué des réflexions épistémologiques très intéressantes, mettant en cause leur statut objectif qui permettrait à ces données d'être réutilisées pour d'autres recherches.

Par exemple, un historien travaillant sur des archives soulève la question d'une réutilisation possible des corpus qu'il a constitués dans sa recherche. Si la constitution d'un corpus est orienté par une idée de départ (un sujet, une question, un problème) la sélection de ces archives est déjà une interprétation, donc à la rigueur une **élaboration** des « données brutes » qui constituaient l'ensemble d'archives non dépouillés au départ.

« A supposer qu'il les mette en ligne dans une base de données, un chercheur ne pourrait plus les réutiliser en tant que « données brutes ». De là l'importance de bien documenter ces données, le contextualiser, avant de les diffuser, pour éviter une réutilisation inappropriée de ces sources. Toutefois, il est visible aussi par-là, qu'une telle réutilisation sera assez limitée[63]. » (Histoire, entretien n°49)

De même, en linguistique expérimentale, les données acoustiques collectées en laboratoire semblent ne pas pouvoir être séparées du contexte dans lequel elles ont été produites. S'agissant en plus d'une expérience, ces données sont à caractère unique.

« Il pense d'ailleurs qu'il est extrêmement difficile de parler des « données brutes » dans son domaine, car toute collecte est ciblée au départ par un objectif et une expérience qui va la provoquer. » (Linguistique, entretien n° 50)

[63] L'entretien n° 11 présente un développement semblable à la différence que ce chercheur voit dans la documentation et la qualité des métadonnées une possibilité de créer des véritables outils de la recherche. Nous traitons cela de plus près en 2.4.2

Les données brutes selon les chercheurs
- Des données collectées avant toute interprétation ou analyse (archéologie, droit)) - Des archives, photos, élément d'architecture, « données de terrain » ; avant toute description ou annotation (histoire, arts) - Des données quantitatives sans documentation (sociologie) - Des informations quantitatives (sociologie) - Des données avant tout traitement éditorial, structuration en base de données - Données statistiques INSEE, INED, Eurostat (sciences politiques) - Micro-données INSEE (sciences administratives, sciences économiques) - Des données web (sciences juridiques)

1.3 Les données dans le processus de recherche

A suivre, nous illustrerons les propos précédents à travers deux exemples tirés des entretiens, visant à montrer à quel moment des données « deviennent » des données de la recherche.

1.3.1 En Histoire : distinguer les données selon leur degré d'élaboration et transformations

« Sur la plateforme hypermédia conçue par lui et des coopérateurs, cette distinction [entre les données d' « appui » et les données « produites »] a été réalisée pour classer les données selon leur degré d'élaboration. Les « données brutes », annotées ou enrichies de métadonnées, sont classées dans les « sources ». Les données secondaires sont le produit d'une élaboration qui suppose – dans la plupart des cas - de déstructurer le document initial (agrégation, traitement, extraction sélective) et de s'affranchir de la représentation initiale du support pour élaboration d'une nouvelle visualisation des données. Ces données produites par la recherche sont plus proches d'« outils » ou d'instruments de recherche (typiquement, les « bases de données. » (Histoire, entretien n° 11)

« Dans le cadre de ses recherches en histoire, les données de la recherche sont avant tout des données recueillies en archives. Il importe toutefois de distinguer la « donnée brute », le texte

d'archive lui-même - qu'il qualifierait plutôt de « données de terrain »-, des données de la recherche qui correspondent plus à un travail de pré-analyse. » (Histoire, entretien n°27)

Ces deux historiens partagent l'idée selon laquelle les données « brutes », par exemple, des documents d'archives, ne constituent pas à proprement parler des « données de la recherche » dans les sens plus restrictif des « données produites » au cours de de la recherche. En effet une donnée de la recherche au sens propre serait une donnée conçue par une transformation appliquée aux sources primaires ou données brutes dans le but de produire du savoir scientifique. Il est par là donc évident que le rapport entretenu entre « données brutes » ou « sources » et « données secondaires ou de pré-analyse est loin d'être un rapport purement externe mais, selon l'expression du chercheur, suppose une déstructuration qui est déjà une interprétation des premières : sélection et agrégation de documents d'archive pour constitution d'un corpus sur un sujet particulier.

Mais cette transformation est comprise de deux façons différentes par ces deux chercheurs : pour le premier chercheur (entretien n°11), elle a un sens qui se traduit *techniquement* (par la réalisation d'un ensemble d'opérations), alors que pour le deuxième (entretien n° 27 et v. aussi n° 49), il s'agit d'une transformation opérant à travers les étapes d'un travail intellectuel sans contrepartie technique.

Dans le cas du premier extrait cité, l'intention du partage existe, il s'agit de proposer des instruments de recherche à d'autres chercheurs : des corpus annotés, documentés, enrichis de métadonnées descriptives. En effet, un des effets du tournant numérique des SHS réside dans la possibilité offerte de rendre visibles les corpus ainsi constitués, alors qu'avant le Web le chercheur avait des possibilités très réduites de le faire.

En conséquence les sources sont potentiellement des données de la recherche lorsqu'elles sont objet de transformations intellectuelles qui se traduisent par un ensemble d'opérations techniques visant à rendre ces données utilisables à d'autres chercheurs (Fig. 6).

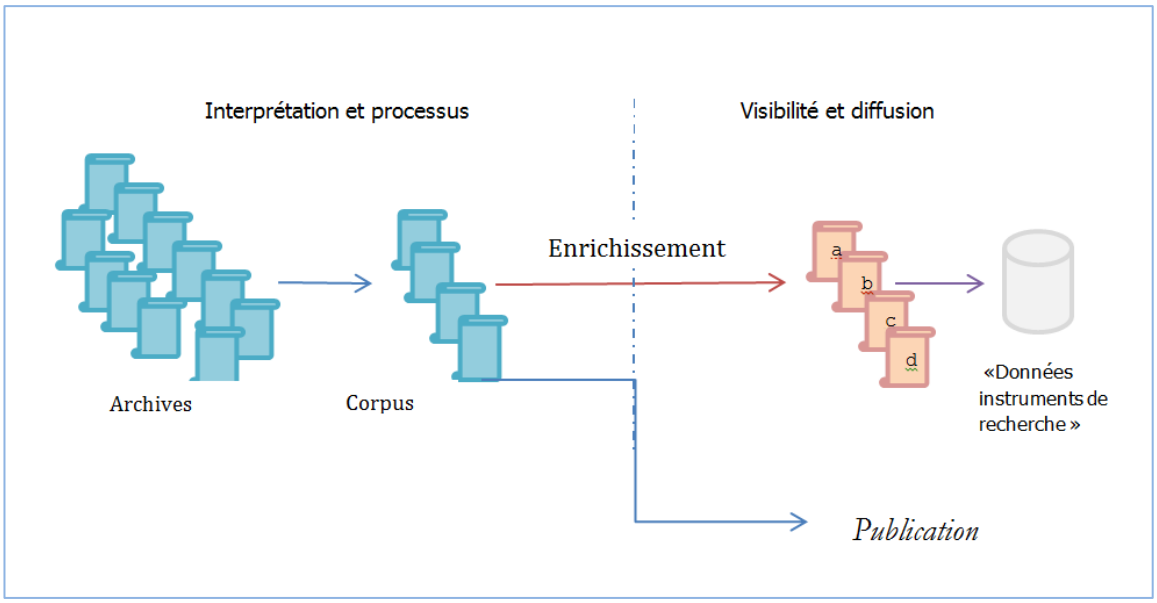


Figure 6 - Des sources aux données de la recherche

En Histoire, comme dans d'autres disciplines utilisant des données d'archives ou corpus de textes, un champ très large de réflexion sur les sources, outils et les finalités poursuivies est en train de se produire à travers quelques initiatives innovantes comme par exemple, Criminocorpus[64], plateforme d'édition scientifique pour l'histoire de la justice, des crimes et des peines, produit grâce à la coopération entre chercheurs, archivistes, documentalistes et collectionneurs.

1.3.2 Des données hétérogènes tout au long du processus de recherche

Tout au long de ce processus plusieurs types de données peuvent être produites, utilisées et réutilisées. Un exemple très caractéristique consiste à montrer toutes les types de données impliquées dans la démarche d'*enquête*. Dans l'entretien sur lequel nous basons ces remarques [Démographie, 28] il est possible de repérer les différentes étapes de cette démarche (fig.7) :

[64] V. < https://criminocorpus.org/>; une autre initiative d'envergure est le développement de la plateforme web sur l'Histoire des sciences du Centre Alexandre Koyré : <http://koyre.ehess.fr/index.php?936>

1. Entretiens qualitatifs selon une approche biographique auprès des immigrants de l'Afrique Sub-saharienne (données qualitatives, méthodologie)

2. À partir des données collectées, préparation d'une enquête pilote (méthodologie)

3. Enquête réalisé par un institut de sondage (données quantitatives)

4. Analyse de ces enquêtes (résultats de la recherche)

5. Publication

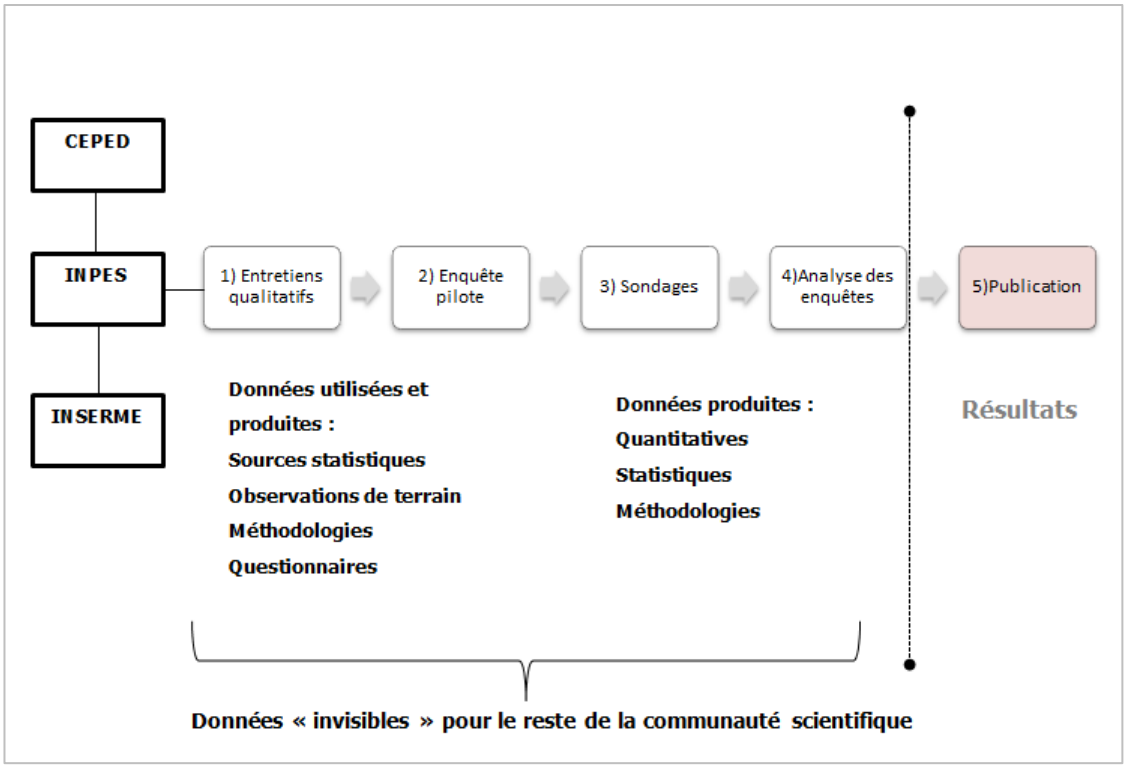


Figure 7 - Processus type d'une démarche d'enquêtes

Le contexte de cette recherche est collectif et interdisciplinaire : « Plusieurs équipes appartenant à des organismes divers (CEPED, INSERME, INPES) ont travaillé sur la préparation de ces enquêtes pendant plusieurs mois. »

Tout au long de la recherche les chercheurs produisent plusieurs jeux de données qui ne sont pas forcément mis en valeur dans les résultats diffusés. Il est naturel par exemple, que les données produites dans le but d'appuyer des résultats soient traitées et documentées de façon plus exhaustive à des fins de publication. Mais d'autres jeux de données produits en cours de chemin resteront moins visibles, inexploités et pas traités pour un partage.

La réutilisation des données produites par d'autres chercheurs en parallèle à la production de ses propres données peut également faire partie d'une démarche du type constitution d'enquêtes. Par exemple, la méthodologie utilisée pour constituer un échantillon, collecte et analyse des données peut être utilisée par un chercheur en vue de constituer sa propre enquête, sur un sujet divers.

« Elle estime que ces travaux préparatoires sont effectivement très riches et présentent un intérêt à être diffusés, notamment du point de vue méthodologique. Une difficulté se présente, néanmoins, originaire de la particularité de cette étude d'être à cheval sur deux disciplines, la démographie et l'épidémiologie, et des types différents de données utilisées, qualitatives et quantitatives. Il faudrait réfléchir à un moyen de présenter ces données «brutes » en les contextualisant, mais cela implique beaucoup de temps disponible pour leur préparation et les chercheurs ne sont souvent pas en mesure d'accomplir seuls ce travail. » (Démographie, entretien n° 28)

1.4 Quelles données partager ? Quelles problématiques ?

Un résultat qui se dégage rapidement de notre première étape d'analyse concerne les données « sources » et le constat que les problématiques associées à ce type de données appartiennent à un univers différent de celui des « données » produites dans les différentes étapes de la recherche dont nous avons pu ébaucher une définition dans le chapitre précédent.

Pour ces dernières, les chercheurs peuvent attribuer des valeurs plus ou moins grandes en termes de :

- Intérêt de réutilisation ou potentiel de nouvelles formes d'utilisation,
- Intérêt à être traitées, conservées, pérennisées en tant que patrimoine scientifique
- Intérêt à produire des outils pour la recherche,
- Intérêt à produire des nouvelles connaissances,
- Intérêt de permettre la reproduction ou la réplicabilité d'une recherche

Les données qualitatives intéressent particulièrement les chercheurs car elles incluent une diversité de matériaux très riches : notes d'observation de terrain, méthodologies

d'enquêtes, préparation de l'échantillon, préparation des questionnaires, choix des restitutions, etc[65].

TYPE DE DONNEES LE PLUS SOUVENT PARTAGEES	**PERIMETRE**	**ENJEUX DU PARTAGE**	**INTERET**
Données quantitatives	Equipe travaillant sur le même projet Autres équipes	Ces données doivent être correctement traitées et documentées pour être réutilisables	Vérification Réutilisation Réplicabilité/ reproductibilité
TYPE DE DONNEES RAREMENT PARTAGEES	**PERIMETRE**	**ENJEUX DU PARTAGE**	**INTERET**
Données qualitatives Données de terrain	Cercle réduit de chercheurs	Questions de déontologie Respect à la confidentialité	Enseignement Analyses secondaires Renouvellement des questions Preuve
TYPE DE DONNEES QUE LES CHERCHEURS SOUHAITERAIENT POUVOIR PARTAGER OU REUTILISER	**PERIMETRE ET FORMES DE PARTAGE SOUHAITEES**	**BENEFICE DU PARTAGE**	
Données qualitatives Méthodologies d'enquêtes et de traitement des données qualitatives	Un partage large est difficile d'envisager Cercle plus large de chercheurs, au moyen de plateformes	Le partage permettrait l'optimisation de ces matériaux : réalisation d'analyses secondaires, amélioration de l'enseignement de jeunes chercheurs, possibilités d'études comparatives basées sur la reproduction d'une enquête, enrichissement des méthodologies, transfert de savoir en matière de traitement des données qualitatives	

[65] Pour un aperçu de la richesse de ce matériel, consulter la rubrique « Ressources et méthodes » de l'INED : http://www.ined.fr/en/resources-methods/survey-methodology/les-demarches/.

	collaboratives	

Les principales difficultés ou obstacles évoqués à une diffusion des données sont principalement de deux types : questions de confidentialité et questions de temps et d'investissement à fournir.

PROBLEMATIQUES ET ENJEUX EVOQUEES

Données source/matériaux	*Données « produites »*
Accès	***Exploitation des données qualitatives anonymisées***
Littérature grise	
Données sur d'autres supports	***Confidentialité/éthique***
Anciens numéros de revue	
Sources numérisées	***Archivage et pérennité***
Droit d'auteur :	***Infrastructures insuffisantes***
Photos/images	
	Réutilisation
Sous exploitation des données statistiques :	***Documentation et contextualisation des données « brutes »***
Micro-données	
	Déontologie

2 Ouvrir et diffuser les données

Tout au long des entretiens les termes « ouvrir » « partager » « diffuser » et « publier », ont été utilisés de façon assez souple et parfois imprécise. Alors qu'entre « ouvrir » et « partager » la distinction peut paraître claire, celle entre « diffuser » et « ouvrir » peut être plus subtile.

Nous avons procédé à une distinction minimale de ces termes car les pratiques diffèrent sensiblement en termes de finalité, de périmètre d'ouverture ou de diffusion souhaités. Cette distinction s'avère utile aussi pour une réflexion sur les usages, car les enjeux y différent considérablement, notamment en termes de droits de réutilisation de ces données. En effet, diffuser les données en accès libre sur une plate-forme en ligne ne les empêche d'être protégées par les droits de la propriété intellectuelle. Alors que partager des données entre chercheurs peut permettre des réutilisations moins restrictives. Et enfin, l' « ouverture » des données de la recherche devrait idéalement garantir une réutilisation sans barrières juridiques[66].

Le lecteur trouvera ci-dessous, à titre indicatif, des définitions très générales[67] des termes que nous emploierons au long des sections suivantes où les entretiens seront analysés du point de vue des pratiques des chercheurs. C'est de façon intentionnelle que ces définitions ne soulignent pas les aspects juridiques de chaque pratique et fournissent un minimum d'éléments d'explication. Il sera possible ainsi d'éviter de superposer notre discours à la parole des chercheurs et à ce qu'ils/elles comprennent de ces pratiques.

- ***Partager*** **: transmettre volontairement de l'information de chercheur à chercheur ou de laboratoire à laboratoire.**

- ***Diffuser :*** **rendre accessible à une consultation large**

- ***Publier*** : **transmettre les données à une revue ou à une plateforme de publication de données sous des conditions prédéfinies par celles-ci.**

[66] The Open Definition : "Open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)." In http://opendefinition.org/

[67] Ces définitions s'inspirent du rapport de Knowledge Exchanges « The Value of Research Data - Metrics for datasets from a cultural and technical point of view » qui procède également à la définition des quatre concepts liées aux pratiques des chercheurs dans le domaine des données de la recherche. Compte tenu de l'objectif spécifique poursuivi par ce rapport - offrir un état de l'art domaine de la « data science metrics » - ces concepts ne coïncident pas tout à fait avec ceux déployés ici. V. Costas, R., Meijer, I., Zahedi, Z. and Wouters, P. (2013). The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange (KE) Report, available from <www.knowledge-exchange.info/datametrics>.

- ***Ouvrir :*** **rendre disponible pour réutilisation, partage et modification**

2.1 Partager des données entre chercheurs

Suivant la définition minimale donnée au début de ce chapitre, le terme « partage » sera compris ici comme l'acte volontaire de transmettre de l'information de chercheur(s) à chercheur(s) ou de laboratoire à laboratoire strictement dans le cadre d'une collaboration scientifique. Sont exclus donc d'autres acteurs ou d'autres finalités que celle de coopérer dans la recherche en mutualisant les outils et les données. Par exemple, les exigences épistémologiques d'ouverture des données ayant valeur de justification de la sociologie (v. entretien 34) ne sont pas considérées comme un partage, car ne sont pas motivées par une intention réfléchie de coopérer avec un ou des chercheurs en particulier (travaillant, en général sur des sujets proches).
Il est également sous-entendu que « partager » ses données implique une acceptation de l'éventualité de leur réutilisation.
Si le partage est considéré de manière générale très salutaire à la construction de l' « édifice commun de la recherche » (V. entretien n° 12), il peut rencontrer des obstacles à sa concrétisation selon le type de données, le mode de travail du chercheur et ses motivations au partage.
Des 53 interviewés, 19 ont affirmé pratiquer une certaine forme de partage ou être actuellement engagés dans une action allant dans cette direction.
Dans le cadre de cette étude les résultats « comptabilisés » pour chaque discipline ne sont pas significatifs. Néanmoins, quelques comparaisons peuvent être intéressantes parmi les disciplines le plus représentées :

- En Sociologie : seulement 2 chercheurs sur 7 affirment pratiquer une forme de partage dans un cercle restreint de chercheurs
- En Histoire : 5 chercheurs sur 8 partagent des données avec des collègues de la même équipe ou travaillant sur des sujets proches
- En Economie : 2 chercheurs sur 4 partagent sur internet (des codes) et au sein de l'équipe
- En Sciences de l'éducation, géographie, démographie, le partage des données fait partie des programmes collectifs développées au sein des laboratoires.

Parmi les chercheurs en droit, philosophie, arts et cinéma, le partage est un terme dépourvu d'une signification concrète, leur recherche n'ayant pas d'étapes de production de données indépendantes du travail d'élaboration textuelle dont l'objectif premier est la publication (sous forme d'article ou ouvrage).

2.1.1 Partager au sein de l'équipe

Le partage des données au sein de l'équipe du laboratoire est sans doute la forme de partage la plus pratiquée. Celle-ci peut éventuellement s'étendre à d'autres cercles de chercheur(e)s appartenant à d'autres laboratoires mais qui sont, en général, impliqué (e)s dans le même projet. Dans le cadre d'une recherche en équipe, ces données sont partagées immédiatement[68].

Les partages dans des cercles très réduits de chercheurs (membres d'une équipe travaillant sur le même sujet, dans le même laboratoire) concernent les données qualitatives et de terrain. Cela s'explique facilement, compte tenu de la nature confidentielle de ces données, collectées à travers d'entretiens qualitatifs sur la base d'un contrat oral de confiance passé entre le chercheur et la personne interviewée.

« Le partage des données se fait uniquement au sein de son équipe et une diffusion plus large semble difficilement envisageable. Tout d'abord parce que les données nécessitent d'un travail de contextualisation important pour avoir une valeur scientifique à part entière. Ensuite parce que ces données touchent de près la vie des personnes concernées et ne peuvent / doivent pas être diffusées sans leurs accords. » (Anthropologie, entretien n° 9)

Les données quantitatives sont, quant à elles, partagées dans un périmètre plus large surtout si elles ne posent pas de problème particulier et aussi du fait de bénéficier de la possibilité d'un dépôt institutionnel au CDSP par exemple ou autre centre de données.

« Ces tableaux sont partagés par les membres du laboratoire mais une diffusion à plus large échelle est envisagée. Son équipe travaille actuellement aux modalités de cette mise à disposition avec un ingénieur de recherche, spécialiste des problématiques de la « mise en données » des sources et à la constitution de bases de données pour la recherche. L'ouverture de ces données n'est pas particulièrement problématique dans ce cas, car il s'agit de données concernant des personnes disparues depuis longtemps. » (Démographie, entretien n°6) [1, 5, 23, 27, 32, 33, 40, 43, 46, 52]

2.1.2 Partage one-to-one selon demande

Certains chercheurs se réfèrent au partage entre « collègues » en fonction d'une demande. Il peut s'agir des données qui sont conservées à titre personnel et qui dépendent uniquement de la volonté du chercheur à être transmises.

[68] Le moment de mise à disposition des données par les chercheurs à des équipes ou chercheurs externes aux laboratoires n'a pas été éclairé par ces entretiens. Mais il n'est pas abusif d'imaginer qu'une partie des chercheurs souhaitent exploiter leurs données avant de les partager.

Dans d'autres cas, la demande peut concerner des données qui ont été déposées dans un centre de données et qui, pour des raisons diverses, sont difficilement accessibles dans celui-ci (démarches administratives, embargo).

« Les chercheurs peuvent, parfois, contourner ces dispositifs et faire le « one to one », c'est-à-dire se transmettre des données entre eux. A sa connaissance, ces pratiques restent assez limitées à des projets menés collectivement. » (Sciences Politiques, entretien n°8)

« (..) le partage des données selon d'autres formes, moins formelles, existe : de chercheur à chercheur ou de chercheur à étudiant, à l'occasion d'échanges dans les locaux de l'université. Si un chercheur est sollicité par d'autres chercheurs à partager certaines données, il le fera sans trop de difficultés, mais il le fera difficilement de sa propre initiative et sans un objectif très précis. » (Sociologie, entretien n°34)

2.1.3 Motivations et freins au partage

Même si une partie considérable des chercheurs que nous avons pu interviewer ne partagent pas leurs données avec d'autres chercheurs, ils ont été en général très disposés à reconnaitre que le partage est une source de bénéfices pour celui qui partage et pour celui à qui on autorise une réutilisation de ces données. De plus, une grande majorité serait disposée à le faire si les conditions existaient. Car ce qui peut apparaître comme simplement une transmission des fichiers envoyés par mail, révèle une réalité bien plus complexe. Si ces données n'ont pas été suffisamment et correctement traitées, décrites et documentées le partage s'avère impossible à défaut de prendre le temps pour le faire. Le « temps manquant» a été invoquée, effectivement, comme un des freins principaux rencontrés au partage ou à toute autre diffusion. En outre, plus la recherche est loin dans le temps, plus il devient difficile, en absence d'une documentation, de retrouver le contexte de production de ces données.

Principales motivations au partage :

- Mutualiser et échanger des données permet d'avancer plus vite et de créer un esprit de travail collectif parfois rare chez les chercheurs en SHS [23, 27, 40, 46,47]
- Le partage peut aider à contourner les dispositifs administratifs lourds des centres de données [8]
- Le partage contribue à la mention explicite de la source des données dans les publications contribuant à la réputation du chercheur ou du laboratoire [5, 32, 46]

- Proposer des méthodologies documentées et instruments de recherche ayant fait leurs preuves permet d'avancer plus vite, cela devrait être une pratique courante parmi les chercheurs [17, 28, 34, 36, 40, 42, 50]

- Partager ses données entre chercheurs permet d'élargir l'horizon de départ de la recherche et enrichir celle-ci avec des nouvelles données

Principaux freins au partage

- La préparation des données est chronophage et nécessite des conditions propices à sa réalisation, par exemple, un environnement de collaboration entre les chercheur(e)s, la reconnaissance et le soutien institutionnel (nouveaux recrutements, financements) [17, 32, 23, 29, 39]

« Les chercheurs en esthétique ne partagent pas leurs données autrement que dans les formes déjà mentionnées, probablement parce que leur manière de procéder est très littéraire et comporte une bonne partie de travail intellectuel individuel. Mais aussi par ce qu'une « vie de laboratoire » authentique fait défaut, il n'y a pas des locaux institutionnalisés qui favorisent un échange plus dynamique entre les chercheurs. Malgré cela, il pense assister à une évolution progressive vers des formes plus collectives de la recherche. » (Cinéma, entretien n°39)

- Les données brutes nécessitent d'être correctement documentées pour servir à d'autres recherches : ces données ont un intérêt limité si elles ne sont pas accompagnées d'une documentation et d'une description méthodologique. De plus il faut que le chercheur ait les compétences nécessaires pour traiter ces données, ce qui n'est pas toujours le cas.

« Les cartographies et l'article issus du travail sur ces données ont été partagés, mais pas la base de données elle-même. En effet, le partage de ces données « brutes » n'aurait pas été de grand intérêt en tant que source d'information. D'autre part, il y aurait certainement un grand intérêt à contextualiser, documenter et décrire ces données pour les rendre réutilisables dans de bonnes conditions. » (Sociologie, entretien n°14; v. aussi 17)

« Il y a toutefois un bémol pratique au partage des données : la compétence spécifique de celui qui les a recueillies. C'est, à ses yeux, plus difficile d'utiliser les données d'autre chercheur si celui-ci ne rend pas disponible ou ne communique pas la façon dont elles ont été recueillies et quelles en sont les limites. » (Histoire, entretien n° 27)

- Les chercheurs en France seraient divisés entre une culture du partage et une culture de la propriété : certains chercheurs estiment que les données appartiennent au chercheur qui les a produites et qu'un partage serait improbable avant une exploitation complète de ces données. Cela peut être motivé par des facteurs de

compétition mais aussi par des formes de travail plus individualistes ou impliquant une grande partie de subjectivité, comme dans la recherche littéraire ou en philosophie.

« (…) elle observe que les nouvelles technologies n'ont pas changé beaucoup les modes de travail des chercheurs, assez individualistes encore, ainsi que leur rapport aux données produites par leurs recherches, qui restent très majoritairement leur propriété. » (Sociologie, entretien n° 34)

« La question urgente à se poser et à laquelle il faut répondre est, à son avis : « A qui appartiennent les données ? ». Les chercheurs sont encore divisés entre une culture de la propriété intellectuelle et une culture du partage. » (Sociologie, entretien n° 17)

« Les chercheurs en SHS en France se considèrent très souvent « propriétaires des données ». Cela provient de la dimension « auteur/littérateur » que possède la recherche en SHS. » (Sociologue, entretien n° 22)

« Concernant les pratiques des chercheurs, le paysage en France est partagé : alors qu'on assiste à une ouverture progressive de ces données par des initiatives individuelles ou collectives, il est à la fois toujours facile de constater l'existence d'une culture de la propriété des données dans le milieu scientifique et des modes de travail plus individualistes. » (Sciences Politiques, entretien n°8)

- Les données impliquent des questions de confidentialité et éthique et les partager décontextualisées peut ôter une partie importante de leur valeur épistémologique.
- En SHS les chercheurs éprouvent une réticence au partage des méthodologies qui doivent accompagner leurs données [22, 36,40]
- Lorsqu'un chercheur ouvre ses données, il n'est pas toujours possible de garantir que les usages qui en seront faits respecteront les règles déontologiques.

☛ Points d'attention :

1) Les données qui suscitent le plus grand intérêt au partage sont les méthodologies, c'est-à-dire, la documentation qui accompagne les données « brutes » quantifiables d'une enquête, par exemple. Ces méthodologies sont-elles mêmes toute sorte de données du type qualitatif.

2) Les données qui posent le plus de difficultés au partage sont les données qualitatives

3) les données qui suscitent le moins d'intention de partage sont les données du type « interprétation », c'est-à-dire, étapes inhérentes au processus subjectif

4) Sous quelles conditions les chercheurs seraient-ils/elles disposés au partage et réutilisation des données ?

2.2 Diffusion et publication des données

« Publier » et « diffuser » des données sont des termes utilisés souvent de manière interchangeable et pas rigoureuse par les chercheurs.

« Publier » peut signifier une publication en revue, comme celle sur un site internet personnel ou une plate-forme en ligne. A son tour, « diffuser » des données peut également vouloir dire une publication en revue pour certains chercheurs alors que pour d'autres cela revêt le sens d'une mise à disposition plus large des données à des fins de consultation, par exemple, dans une base de données ou plateforme scientifique ouverte (à des publics non scientifiques, par exemple).

Dans l'un comme dans l'autre cas, la **réutilisation** de ces données publiées ou diffusées n'a pas été mise en exergue ou suffisamment définie pour qu'on puisse établir un cadre clair des périmètres d'ouverture attribués à chacune de ces formes d'exposition des données.

Nous avons ainsi établi la distinction entre « diffuser » et « publier » pour des raisons de clarté, comprenant ce dernier terme exclusivement comme l'exposition ou la mise en accès des données dans une plateforme éditoriale ou revue.

« Diffusion » sera, à son tour, compris de façon beaucoup plus large comme l'action d'exposer des données sur le web en vue d'une consultation large, par exemple, en amont d'une publication ou indépendamment d'une publication[69]. Ce qui distingue une « diffusion » du « partage », tel que nous avons pu le définir plus haut, est que la diffusion ne vise pas un chercheur ou une équipe en particulier mais se destine à un public plus large.

A ce stade, ces termes sont définis ici indépendamment de leur périmètre d'ouverture ou conditions de réutilisation de ces données.

Dans les sections suivantes, nous présentons les modes de publication et diffusion pratiqués ou préconisés, les motivations et les freins posés, soit par des raisons inhérentes au type de donnée, soit par d'autres raisons plus subjectives.

[69] Le rapport de RIN (v. supra) de 2008 constate une imprécision assez semblable dans le sens attribué à « publishing » parfois en hybridation aussi avec « sharing ».

2.2.1 Dépôt institutionnel

Le dépôt dans les centres de données est pratiqué par les chercheurs produisant des données quantitatives. Deux dépôts ont été cités :

- Le réseau Quetelet
- Les linguistes déposent parfois leurs corpus dans la plateforme ELRA (European Language Resources Association) dont l'accès est en partie payant.

2.2.2 Diffuser dans les plateformes scientifiques sur le Web

Les données des chercheurs d'un laboratoire peuvent être mises à disposition aux chercheurs à travers des initiatives de création de plateformes accessibles via Internet.

Très citées dans les entretiens, seulement dans trois cas il s'agit de plateformes déjà existantes en Histoire, Arts et Anthropologie [11, 29, 44, 4]. Les motivations au partage dans ces plateformes sont principalement :

- Ouverture immédiate des données à d'autres chercheurs ;
- Une plus grande visibilité permet de créer des réseaux scientifiques nationaux et internationaux importants ;
- Création de véritables outils pour la recherche.

Mais les plateformes suscitent l'intérêt auprès des autres chercheurs pour plusieurs raisons ;

- Possibilité de publier en libre accès les données (Histoire, entretien n° 5) ;
- La démultiplication des plateformes expérimentales pour les données est importante pour permettre d'évaluer leurs formes d'utilisation dans la communauté scientifique (Sciences politiques; entretien n°19)
- Les plateformes pourraient suffire à promouvoir la diffusion des connaissances, aidant les chercheurs à s'affranchir du système d'évaluation par publication (Economie, entretien n° 31) ;

- Des plateformes d'échange de méthodologies seraient souhaitables (Démographie, entretien n°28)

La question de la diffusion des données se pose ici indépendamment d'une éventuelle publication des résultats a posteriori, fonctionnant comme une sorte de « pré-print » qui pourrait engendrer un circuit différent pour ces données parmi le réseau des chercheurs. (V. entretiens 4 et 5) mais qui pourraient être également valorisées dans une publication en revue (soit par un lien, soit par une citation).

2.2.3 Publier en revue

Quel rôle les données peuvent ou pourront jouer en enrichissant les publications des revues ou inaugurant des nouvelles formes de communication scientifique à travers, par exemple, les *data journals* (v.5.1.2) ?

Le rapport des chercheurs à l'édition scientifique peut être objet d'un aperçu en fonction de la discipline d'appartenance. Il existe, en effet, un certain nombre de traits remarquables qui caractérisent certaines disciplines dans leur approche à la communication scientifique. Même lorsque cette approche ne se rapporte pas directement aux données, elle éclaire à contre-jour les raisons pour lesquelles certaines disciplines semblent moins impactées par les sujets relatifs aux données de la recherche. Nous traiterons en parallèle les formes de publication des données préconisées par les chercheurs[70] .

En Droit : le rôle fondamental des revues et l'attachement au format papier

Les publications dans cette discipline n'ont pas lieu sous le même régime que dans d'autres disciplines. En premier lieu, les chercheurs et juristes sont rémunérés par publication. Deuxièmement, les maisons d'édition en Droit assurent des positions historiquement dominantes sur le marché de l'édition scientifique. En conséquence de ce contexte français, les chercheurs ne subissent pas la pression d'une « publication intensive » comme dans les autres disciplines des SHS. La question de publier des données ne se pose aucunement, car les juristes ne produisent pas de données mais des textes juridiques ou de texte d'interprétation. En outre, l'attachement au format papier est une caractéristique de cette discipline.

[70] Car en concret, à l'exclusion des économistes aucun chercheur interviewé ne pratique une forme quelconque de publication en revue.

En Philosophie et Arts (esthétique) les revues représentent les domaines et les réseaux de chercheurs

La publication en revue dans ces disciplines remplit le rôle important de signaler l'appartenance des chercheurs à certaines écoles de pensée. Dans ce sens la ligne éditorial et le renom de la revue dans des domaines spécialisés sont déterminants. La revue comme support de communication scientifique est loin de disparaître en Philosophie, bien au contraire, l'enjeu majeur pour cette discipline est l'accessibilité élargie à des anciens numéros difficiles à consulter et non encore numérisés. Dans le domaine de l'Esthétique (Cinéma), la publication des actes des colloques et conférences jouent un rôle très important. Les données de la recherche n'occupent pas une place importante dans la réflexion des chercheurs de ces disciplines, mais des pratiques nouvelles apparaissent chez les chercheurs plus jeunes :

« Elle diffuse les étapes intermédiaires de sa recherche dans son blog personnel hébergé dans une plateforme collective de chercheurs et pense que la plupart de jeunes chercheurs en font autant. Mais cela reste incompris par quelques-uns de ses collègues qui préfèrent les canaux traditionnels de diffusion des résultats de leurs recherches, comme les revues. Parmi les raisons invoquées de ces préférences, le plagiat revient de forme récurrente. » (Arts/Photographie, entretien n° 44)

En Linguistique et Psychologie, les chercheurs dépendent beaucoup de la publication en revues spécialisées et actes des conférences et publient encore très peu en Open Access.

- **Type de publication préconisée pour les données :** Le domaine de la linguistique étant compétitif, une publication des données n'est envisageable qu'en aval de la publication des résultats.

En Économie les chercheurs doivent publier en anglais et transmettre les données en fonction des exigences de la revue. Il y a une insatisfaction à l'égard du système de publication dominante qui est déséquilibré : d'un côté il existe une forte pression du domaine poussant à publier dans des revues prestigieuses (l'auteur doit payer au moment de soumettre son article), d'autre côté une incitation institutionnelle à déposer des articles dans des archives ouvertes. Mais en définitive ce qui pèse dans la balance pour la réputation du chercheur est la première alternative.

- **Type de publication préconisée pour les données :** en parallèle et en aval des résultats, les chercheurs étant très compétitifs dans ce domaine.

En Histoire, il n'a pas été possible de retrouver des points de convergence très marqués entre les avis des chercheurs interrogés. En même temps que l'attachement au format papier est évoqué comme caractéristique des historiens par un des chercheurs, d'autres

chercheurs semblent avoir accompagné le mouvement de numérisation croissante des ressources disponibles sur Internet et promouvoir la publication des données dans des plateformes numériques. Un format éditorial unique pour ces données est peu concevable, mais « celui proposé par Robert Darnton peut être considéré comme théoriquement idéal : le chercheur présente un résultat en donnant accès à ses sources, qui constituent alors des éléments de preuves et de vérification. » (Histoire, entretien n°11)

D'autre part, ce foisonnement de ressources disponibles et publications Open Access est également abordé comme un problème par un chercheur. Le manque de temps pour faire le tour complet des publications, les lire, les trier, est ainsi évoqué comme une des raisons pour laquelle la publication de données paraît superflue.

« (…)il considère que l'objectif des publications de qualité est précisément de permettre aux chercheurs/lecteurs un accès à des formes abouties de la recherche, résultats d'un processus mené à terme par les chercheurs. Publier des données ou des étapes intermédiaires de la recherche aurait-il donc un sens? (…) Au moment où les chercheurs manquent de temps pour l'essentiel des lectures dans leur propre domaine, il est peu réaliste d'imaginer qu'ils se précipiteront dans l'exploration des données produites par d'autres chercheurs. » [24]

- **Type de publication préconisée pour les données :** en parallèle aux résultats, modèle « Darnton ».

En **Sociologie,** la publication des données devra prendre en compte la possible réutilisation de celles-ci. Si les données sont convenablement traitées, elles pourraient être publiées en parallèle aux résultats. En revanche, il est important de dissocier les deux démarches et ne pas privilégier la publication des données au détriment de leur interprétation. Le risque existe car la préparation des données est hautement chronophage [14].

« Il ne faut pas oublier que le travail du chercheur n'est pas seulement de produire des données, mais aussi de réaliser des comptes rendus qui synthétisent et interprètent ces données. [14] »

- **Type de publication préconisée pour les données** : en parallèle aux résultats.

En **Sciences Politiques**, la publication des données en parallèle aux résultats aurait plusieurs finalités assez importantes : fonctionner comme preuve, permettre la reproduction des résultats, permettre le prolongement de la recherche dans un autre projet.

« La publication devrait aussi permettre à un autre chercheur de prolonger une analyse déjà publiée sans avoir à repartir de zéro. Une diffusion potentiellement intéressante pour ces données, à son avis, se ferait en parallèle ou rapidement après la publication d'articles et ouvrages car une fois que le chercheur change de projet, il est difficile de s'y remettre et de préparer des données correctement. » [38]

- **Type de publication préconisée pour les données :** en parallèle aux résultats à la façon Piketty.

☞ De manière générale, la forme de publication des données en parallèle aux résultats de la recherche semble être la plus largement souhaitée par les chercheurs.

Principales Motivations pour publier les données en revue :

- Les données commencent à être de plus en plus demandées par les revues, en Economie c'est déjà le cas ;
- Les données sont de justificatives méthodologiques et fonctionnent comme preuves et doivent pouvoir être soumises à vérifications et être citées correctement ;
- Contribution à la réputation du chercheur/du laboratoire qui les a produites ;
- Les données peuvent être source et outils à d'autres projets de recherche et contribuent ainsi à accroître la dimension coopérative des SHS.

3 Réutiliser des données

« Ouvrir » et « réutiliser » sont les deux faces de la même monnaie ou, du moins, au sens où l'on pourrait parler d'une ouverture comme elle a été définie selon les principes de l'Open Definition[71] .

A la lecture et examen de ces entretiens, il a été possible de constater que les chercheurs, bien que favorables à l'ouverture de certaines données, ne semblent pas être, eux-mêmes, des grands réutilisateurs de données produites par d'autres chercheurs. Il n'y a pas eu, non plus d'allusion sur les conditions de réutilisation qu'ils préconiseraient pour leurs données. Ils ne sont pas non plus très nombreux à réutiliser des données du type qualitatif ou quantitatif produites par d'autres chercheurs au cours de leurs recherches. La réutilisation la plus évoquée concerne les données statistiques et les données du type instrument de recherche pour un accès aux sources.

Point d'attention :

☛ Pourquoi, alors qu'on est favorable de manière général au partage ou à l'ouverture des données de la recherche (et des données « brutes » si documentées et traitées), réutiliser des données semble non seulement une pratique peu répandue mais, aussi, peu envisagée ?

A suivre, nous essayerons :

- d'indiquer les différents cadres de réutilisation et intention de réutilisation identifiées dans les entretiens ;

- d'identifier les données plus susceptibles d'être réutilisées et la finalité/ ainsi que les données qui ne peuvent pas être réutilisées selon les chercheurs ;

3.1 Finalités de réutilisation

Réutiliser des données ne doit pas être compris comme « citer » ou « signaler » des données. La réutilisation implique que les données seront partie intégrante d'un nouveau projet de recherche. Par exemple, un corpus d'archives en histoire, mis à disposition par un

[71] The Open Definition : « Open data and content can be **freely used, modified, and shared** by **anyone** for **any purpose**" <http://opendefinition.org/>. V. aussi ci-dessous chapitre 9.

chercheur, sera utilisé comme source par un autre chercheur, dans une recherche complémentaire ou tout à fait diverse. Or, ce type de réutilisation semble peu probable pour un nombre de chercheurs en Histoire et Sociologie. Les données produites par une recherche ne peuvent pas, selon ces chercheurs, en être séparées.

« Dans le domaine de la sociologie, un chercheur a besoin de produire ses propres données dans le cadre de ses travaux, c'est un processus inhérent à son activité de recherche. La réutilisation des données produites par d'autres est une problématique qui concerne plutôt des disciplines qui travaillent avec des informations quantifiables. » (Sociologie, entretien n°3)

Pourtant en l'Economie et en sciences sociales les questions de vérification et de réplicabilité sont centrales.

3.2 Quelles données réutiliser ? Quelles conditions ?

Les données qui ont suscité le plus de mentions explicites quant à leur intérêt à être réutilisées sont certainement les données du type « méthodologies » d'enquêtes (en plus des données brutes de ces enquêtes). Celles-ci, comme les données quantitatives, doivent remplir un certain nombre de conditions :

- Etre fiables et avoir un « label » de qualité : la question de la provenance et la garantie de la qualité est essentielle et pour cette raison les données faisant partie des programmes institutionnelles ont plus de chance de susciter des demandes de réutilisation que celles mises à disposition par des chercheurs à titre individuel.
- Etre documentées de façon adéquate ;
- Respecter les règles de confidentialité.

3.3 Principaux intérêts d'une réutilisation

- les méthodes peuvent être confrontées et réutilisées, ainsi que les données ;
- en Sciences sociales et en Economie, la notion de réplicabilité est fondamentale, « dans ce sens il ne suffit pas de mettre à disposition les seules « données brutes », mais il faut aussi rendre disponible la méthodologie documentée de leur collecte en parallèle aux résultats de la recherche qui sont un travail d'interprétation » [Economie, 42] ;

- Les éditeurs scientifiques se positionnent de plus en plus stratégiquement vis-à-vis des exigences épistémologiques de la « réplicabilité » de l'expérience et commencent à exiger (en économie par exemple) les données en parallèle aux articles ;

3.4 Principales obstacles à une réutilisation

- les différentes formes de réutilisation des données d'enquête souhaitées par les chercheurs ne sont pas prévues en amont à la préparation de ces enquêtes, ce qui rendra les utilisations possibles assez limitées ;
- l'anonymisation et décontextualisation rendent difficiles l'utilisation des données qualitatives ;
- les conditions de réutilisation sont souvent en deçà des exigences épistémologiques ;
- Les données qualitatives restent largement confidentielles et soumises à des problèmes de déontologie et respect à la confidentialité ;
- Le risque de plagiat peut être plus grande ;
- Le chercheur doit produire ses propres données.

3.5 Questions juridiques

Un chercheur en Droit résume ainsi la problématique sur laquelle les chercheurs travaillent en ce moment en sciences dures :

« Les chercheurs souhaitent pouvoir trancher sur des questions juridiques touchant la réutilisation des données, en tant que producteurs et en tant qu'utilisateurs. Pour faire la part de ces questions, une distinction importante doit être réalisée : les données brutes et les données ayant subi un traitement (éditorial dans une publication, intégration en bases de données). Les premières sont des données libres de droit et pourraient être utilisées par les chercheurs sans demande d'autorisation. La difficulté réside dans la définition des « données brutes », car dès que ses données subissent un traitement quelconque elles ne seraient plus tout à fait brutes. » (Droit, entretien n°37)

4 Valoriser et préserver les données

Valoriser les données de la recherche est une opération qui vise à augmenter leur visibilité, les rendre plus facilement accessibles et exploitables à des fins d'utilisation, de publication, justification et de réutilisation par d'autres chercheurs.

Indifféremment de l'objectif visé, les données numériques, natives ou générées par la numérisation, posent toujours le problème de leur conservation pendant et après utilisation, de pérennisation et d'interopérabilité. En d'autres mots, garantir que des années après leur collecte ou production elles puissent être retrouvées et exploitées indépendamment des évolutions technologiques. Un défi qui dépasse le périmètre des seuls laboratoires et invite ceux-ci à chercher des solutions stables en s'adressant aux infrastructures (notamment les TGIR Huma-Num et le Réseaux Quetelet) qui les proposent[72].

Si en 2011 Jean-Luc Pinol remarquait que le TGE Adonis (dont la fusion avec Progedo et Corpus IR résulte dans le TGIR Huma-Num) pâtissait d'une certaine méconnaissance dans les SHS[73], peut-on dire que le scénario est complètement changé aujourd'hui ?

Pourtant à y regarder de près, on constate que cette question ne se pose pas seulement comme un problème « matériel », c'est-à-dire, offrir aux chercheurs les moyens de créer, gérer et conserver les données produites. Les infrastructures existent, peut-être beaucoup moins nombreuses que pour les sciences dures, mais offrant un impact déjà mesurables aux SHS au plan européen.

C'est qui est en jeu également se joue dans les diverses pratiques parfois enracinées dans les différentes cultures propres aux disciplines.

Dans ces entretiens, nous avons constaté quatre grandes orientations des pratiques de chercheurs exposées ci-dessous.

[72] Huma-num. Les services de conservation des données. En ligne : <http://www.huma-num.fr/sites/default/files/ressourcesdoc/dossier-thematique-mai2014.pdf>; Réseau Quetelet <http://www.reseau-quetelet.cnrs.fr/spip/rubrique.php3?id_rubrique=2>

[73] Pinol Jean-Luc, « Une infrastructure pour les SHS : le TGE Adonis », Revue d'histoire moderne et contemporaine 5/ 2011 (n° 58-4bis), p. 90-100. En ligne : <www.cairn.info/revue-d-histoire-moderne-et-contemporaine-2011-5-page-90.htm.>

4.1 Assurer la pérennité à travers le dépôt dans les centres de données

Si les chercheurs produisant des données quantitatives (données d'enquête, sondages) semblent plus sensibilisés aux questions liées à la valorisation, cela tient plus à l'injonction institutionnelle qui les pousse fortement au dépôt de ces données dans Quetelet, par exemple, qu'à une réelle motivation de procéder à la valorisation des autres types de données produites avant et pendant la réalisation des enquêtes quantitatives (qui seront, finalement, du matériel qualitatif).

Dans les disciplines ayant affaire aux données quantitatives, comme en sciences sociales et en économie, des pratiques de gestion classiques ont lieu. Dans un premier moment, la collecte des données, stockage de ces données en local et mise à disposition à l'équipe du projet ensuite. Celle-ci pourra, par exemple, réaliser des analyses statistiques sur ces données, les résultats faisant objet de publications a posteriori.

Les centres des données, comme par exemple, le CDSP, procède au traitement (« nettoyage » technique, vérifications et enrichissement des métadonnées). Les chercheurs sont très minoritaires à vouloir ou à avoir les capacités de réaliser ce travail qui est une activité à part entière.

« Les données collectées par ces enquêtes quantitatives en face à face seront accessibles à la seule équipe pendant 3 ou 4 ans et ensuite déposées dans le réseau Quetelet, comme cela se passe toujours dans son domaine. » (Démographie, entretien n°28) ».

4.2 Stocker des données qualitatives

Certains chercheurs produisant des données qualitatives se confrontent à des problèmes particuliers liés à la confidentialité de ces données dont une anonymisation excessive les rendrait inexploitables.

D'autres chercheurs considèrent que certaines données qualitatives sont « expérimentales », c'est-à-dire, issues d'une recherche très spécifique menée dans un temps et lieu uniques, n'offrant aucune réutilisation possible. Dans un cas comme dans l'autre, nous avons constaté que, la plupart du temps, aucune prise en charge de ces problèmes n'avait lieu et que des pratiques de stockage en local étaient la procédure ordinaire.

4.2.1 Les données sont stockées en local sans suite

Stocker des données dans des postes locaux et parfois dans des ordinateurs personnels est encore une pratique courante chez beaucoup de chercheurs et découle d'une production de données pour lesquelles aucune forme de réutilisation n'a été prévue en amont.

« Les données collectées sont confidentielles, il ne peut les partager qu'avec les chercheurs qui travaillent sur le même projet. Ils créent ainsi des bases de données simples sur Excel, exploitent ces données le temps de la recherche et ensuite stockent localement ces données. Aucun système de réutilisation de ces données n'est envisagé car la grande difficulté est d'instaurer un système qui respecte la confidentialité. L'anonymisation excessive rendrait, d'un autre côté, ces données inexploitables. « (R.B., Sciences administratives, entretien n°41)

4.2.2 Les formes de valorisation a posteriori sont plus difficiles

Lorsqu'aucune forme de réutilisation ou diffusion n'a été prévue en amont, les données semblent être plus difficilement objet d'une valorisation particulière, à plus forte raison si une nouvelle recherche a déjà été engagée.

« Il a partagé ces données uniquement avec des collègues et s'est posé récemment la question d'une diffusion plus large, mais le travail de mise en forme et de documentation de ces données est complexe et exigerait peut-être d'y consacrer beaucoup de temps. » (Sociologie, entretien n° 17) [48, 51, 42, 20]

Un chercheur en linguistique remarque que, si les nouvelles TICs sont une opportunité de rendre visibles les données des chercheurs, cela n'est pas encore le cas dans son laboratoire où les supports autres que numériques s'accumulent depuis deux décennies au moins. Sinon, la pratique courante est de stocker sur les postes locaux les données numérisées, sans aucune véritable prévision d'une utilité future. C'est aussi que ces données dites « brutes » ne le sont pas au sens strict et qu'une réutilisation n'est pas prévue lors de leur collecte : il s'agit en effet des données à portée « unique » c'est-à-dire collectées lors de l'observation d'un phénomène dans un moment ou lieu spécifique. [50]

4.3 Engager une réflexion collective sur la gestion des données

Les pratiques décrites ci-dessus, menées bien souvent à titre individuel ou dans un cercle réduit de chercheurs, vont se distinguer des initiatives menées en collectivité qui bénéficient d'une reconnaissance institutionnelle.

4.3.1 Valoriser avant et pendant la collecte des données

Les données peuvent être objet d'une initiative très réfléchie par une collectivité de chercheurs au sein d'un programme à portée nationale (ANR) ou par des équipes ayant le soutien de leur institution. Ces initiatives ne sont plus rares aujourd'hui et doivent prendre une ampleur dans les années à venir, bien qu'il reste difficile à imaginer une participation massive des équipes à taille réduite qui éprouvent beaucoup de peine, à cause du manque de ressources humaines et matérielles, à être porteurs de projets candidats aux « Labex ».

Mais, parfois, le simple recrutement d'ingénieurs de recherches à profil « Humanités numériques » permet au laboratoire de s'engager dans la voie de la valorisation. Dans ces cas, une démarche à deux temps oriente souvent ces initiatives, si le laboratoire porteur du projet a, au préalable, des données stockées qu'il souhaite valoriser. Une démarche visant ces données « anciennes » et une démarche visant l'élaboration d'un plan de gestion pour les données en cours de création ou à être créés. Dans ces cas les données sont valorisées à travers une structuration en bases de données dans des plateformes élaborées pour permettre un accès et des recherches faciles.

« Au sein de son laboratoire ils sont plusieurs chercheurs à s'intéresser aux modalités d'exploitation des données produites par les chercheurs du laboratoire. Depuis 2005, une base de données recense tous les travaux des chercheurs du laboratoire et des quelques chercheurs à l'international travaillant sur la circulation migratoire, avec l'objectif de donner un accès intégral à un maximum de ressources. La réflexion se transpose actuellement aux « données de la recherche ». L'activité scientifique du laboratoire se développe actuellement dans quatre axes de recherche, réunissant différentes compétences interdisciplinaires. La création d'un cinquième axe de recherche intitulée «axe méthodologique » est en cours de discussion au sein du laboratoire et verra certainement le jour. L'objectif de sa création est d'intégrer à la réflexion méthodologique des chercheurs un volet sur la mutualisation et le partage des données au sein du laboratoire. Des groupes travaillent actuellement pour définir les besoins, les contraintes et les objectifs de ce projet visant, à terme, la mise en place d'un dispositif qui puisse concentrer et fédérer des données hétérogènes : quantitatives, qualitatives, observations de terrain. » (Géographie, entretien n° 23) [29, 40, 32, 33]

4.3.2 Enrichir les données

Intégrer des données dans des bases de données implique une attribution de métadonnées descriptives et permet différentes sortes d'enrichissements contextuels. Ces derniers ne sont pas à confondre avec la méthodologie de collecte et analyse de ces données dont la documentation est destinée à permettre leur correcte réutilisation.

« Un de ses travaux de recherche a impliqué la collecte d'une grande quantité de données dans les forums de presse en ligne. Ces données ont été ensuite intégrées à une base de données et enrichies d'autres données contextuelles (données sur les personnes, contexte de la collecte, données de l'INSEE, etc.). » (Sociologue, entretien n° 14) [11]

Des « bonnes » métadonnées sont la condition sine qua non d'une visibilité, accès, utilisation et curation des données. Les métadonnées deviennent objet central des préoccupations des équipes constituant des bases de données et des plateformes multimédias pour les exposer. Cette réflexion fait appel au travail collaboratif entre chercheurs, professionnels de l'information, bibliothécaires et informaticiens pour créer des métadonnées de bonne qualité si possible à partir de référentiels fiables déjà existants.

« Par la même occasion, deux ingénieurs de recherche spécialisés en humanités numériques sont recrutés et mettront en place un outil collaboratif de dépôt et partage des données (…) Les ingénieurs de recherche se chargeront de garantir la qualité des métadonnées descriptives, l'accessibilité des documents et leur diffusion au sein des laboratoires, mais chaque chercheur est libre de déposer ses données. » (Histoire, entretien n°29) [11]

4.3.3 Sensibiliser les chercheurs à traiter les données en amont

Les chercheurs peuvent être réticents à changer leurs modes de travail qui relèvent parfois d'approches assez personnalisées aux données produites, surtout dans des disciplines moins tenues de dévoiler leur méthodologie, comme l'histoire, par exemple. Ainsi, comme l'évoquent plusieurs chercheurs, les transformations des pratiques qui sont en cours grâce au numérique et aux technologies du Web doivent être objet d'un programme de sensibilisation capable de montrer les bénéfices découlant d'un travail en amont sur les données produites par la recherche. Ce programme pourrait se traduire par la formalisation et adoption des politiques d'archivage et gestion des données à l'intérieur du laboratoire.

« Actuellement, ils développent un projet de sensibilisation des collègues à la question du traitement des données : étapes et techniques d'anonymisation, contextualisation de ces données, enjeux liés au

partage et méthodologie. Ce projet a un caractère de « recommandations » non coercitives. » Sciences de l'Education, entretien n° 40) [29]

« Dans son laboratoire, elle mène, auprès des collègues, une campagne pour l'adoption des politiques d'archivage en vue de travailler conjointement les formats des données en garantissant leur pérennité. Mais la question est loin d'être simple et il y a, à son sens, énormément de développements de politiques d'archivage et de sensibilisation à être réalisés en France. » (SIC, entretien n° 26)

« Elle développe actuellement un projet de numérisation d'archives orales, dont les résultats ne sont pas encore publics. Elle et le reste de cette équipe se sentent fortement concernés par la question des données de la recherche et de leur ouverture, même si cette question est loin d'être simple à appréhender.[…]Constituer et diffuser des fonds comme celui que cette équipe a pu constituer pose des questions d'ordre éthique, mais aussi d'ordre juridique, notamment vis-à-vis des règles fixées par la CNIL. S'ajoute à cela le problème financier, la numérisation d'archives orales étant une activité en général assez couteuse et les financements publics étant souvent difficiles à obtenir pour ce type d'initiative. » (Histoire, entretien n° 1)

4.3.4 Les principaux problèmes évoqués

⇒ Accès aux données déposées dans les centres de données nationaux

⇒ Accès aux micro-données des instituts de statistiques

⇒ Infrastructures insuffisantes

⇒ Politiques de gestion et archivage pas assez développées en France

En France les chercheurs produisant des données qualitatives manquent d'un véritable programme pour la gestion de ce type de données

5 Evolutions et perspectives

Nous proposons à présent de nous pencher sur quelques questions se détachant avec une force particulière dans ces entretiens et qui devront gagner en importance dans SHS dans les prochaines années. En effet, les principaux points de vue, situations et problèmes évoqués par les chercheurs concernant les données, permettent à présent d'engager une recherche active sur des réponses et tendances existantes, soit dans le contexte des SHS soit dans des contextes plus larges.

- Les problèmes liés aux données qualitatives qui constituent des pans entiers de matériaux à être réutilisés et/ou valorisés : comment, par exemple, répond-on aux difficultés d'une diffusion des données qualitatives en France ?
- Les formes possibles de publication des données : comment publie-t-on des données actuellement ?
- Les citations des jeux de données et les data metrics comme systèmes de mesure de l'impact des citations : Comment citer un jeu de données ou mesurer l'impact des citations ?
- Les questions juridiques impliquées dans le partage et réutilisation des données : comment commence-t-on à formuler les termes du cadre juridique des données de la recherche ?
- Traiter et documenter les données : comment préparer un plan de gestion ?
- Les quatre premiers points seront traités dans un premier temps (1), le cinquième dans un deuxième temps (2)

5.1 Scénarios possibles d'évolution à la question des données de la recherche en SHS

Le cadre de cette étude ne permet pas en effet d'apporter un degré d'approfondissement sur tous les points qu'on pourrait y dégager. Par exemple, il y aurait été extrêmement utile et profitable aux chercheurs et professionnels consultant ce document de savoir comment deux pays, les EUA et le Royaume-Uni répondent aux défis de la révolution numérique depuis presque une décennie. Toutefois, ce sujet dépasse le périmètre et le calendrier établis au préalable pour cette étude et nous ne pourrions pas traiter ce sujet de façon adéquate. Le lecteur trouvera néanmoins l'indication de quelques ressources importantes concernant ces deux pays dans les Annexes. Nous proposons dans la suite de nous pencher sur quelques questions qui ont émergé avec une force particulière dans ces entretiens et qui devront gagner en importance dans SHS dans les prochaines années.

5.1.1 Les données qualitatives en France : l'initiative beQuali

Les données qualitatives sont-elles vouées à l'enclavement ? Hormis la forme de partage dans un cercle réduit de chercheurs impliqués dans un même projet, les entretiens analysés montrent l'existence de nombreux verrous qui empêcheraient à l'heure actuelle une mise à disposition de ce type de données à d'autres usages. Pourtant pour les Sciences sociales et les Sciences politiques en particulier, fermer la porte aux données qualitatives signifie priver les recherches actuelles et futures des véritables pans de matériaux non exploités ou non exhaustivement exploités. En effet, les données quantitatives, dont la visibilité est très développée en comparaison à celle des données qualitatives, proviennent très souvent de démarches qualitatives préalables: entretiens, observations de terrain, réflexions et mise en place de méthodologies, contextualisations, etc.

En France, le projet de banque d'enquêtes qualitatives beQuali[74], créé en 2010 essaie de répondre aux défis de créer un outil à la mesure des besoins de chercheurs, voire en élargissant les usages de celui-ci.

L'objectif de l'initiative de beQuali est double : «(…) d'une part, que les données d'une recherche et leurs conditions de production puissent être rendues visibles au plus près de la publication des résultats; et d'autre part, que ces matériaux puissent être partagés i.e. réutilisés par des collègues. »[75]

Sophie Duchesne conceptrice et première coordinatrice de ce projet[76], définissait dans un article co-écrit avec Guillaume Garcia, les questions décisives posées d'entrée de jeu par le projet beQuali :

- Quelle orientation choisir pour cet outil ?

  Instrument scientifique ou orientation archivistique ? L'objectif est de sauvegarder des enquêtes pour la postérité et au plus près de leur originalité, organisés selon les principes de l'archivistique ? Ou construire un instrument réunissant une sélection des documents organisés de façon à faciliter leur analyse scientifique ?
- Quels objectifs ?

[74]< http://www.bequali.fr/bequali/>

[75] DUCHESNE S., GARCIA G. « beQuali, une archive qualitative au service des sciences sociales » version pré-print, en ligne : <halshs-00922690>

[76] Ce projet est maintenant coordonné par une équipe permanente d'experts du CDSP : une chargée d'études, une archiviste, un chercheur, un ingénieur d'études. Le projet est nourri de l'expertise et des conseils des membres du comité scientifique et technique de l'équipement quali de DIME-SHS (Données Infrastructures et Méthodes d'Enquête en Sciences Humaines et Sociales).

« Mettre les données d'enquêtes qualitatives à disposition d'autres chercheurs peut viser trois types d'objectifs, tous évidemment légitimes : se donner les moyens de mieux préparer un nouveau terrain ; tirer plus complètement parti d'un jeu de données en le réanalysant à partir d'une nouvelle question et/ou avec des nouveaux outils ; enfin élargir la comparaison pour mettre à l'épreuve sur un plus grand nombre de cas les résultats obtenus sur un petit nombre. (…) L'archivage des enquêtes ouvre la possibilité d'aller plus loin dans cette logique de cumulativité [des sciences sociales] en donnant accès non plus seulement aux résultats publiés des enquêtes précédentes mais aussi aux données et à la compréhension dont elles ont été produites.»[77]

- Quels bénéfices sont attendus ?

  Les bénéfices du côté de l'enseignement des méthodes d'analyse seraient un premier résultat prévisible ; la possibilité de réaliser des analyses secondaires en diversifiant les stratégies et une amélioration des méthodes, un deuxième bénéfice attendu. Outre à cela, réanalyser des données peut fonctionner comme une démarche de pré-enquête et aider le chercheur à bien définir son terrain et construire son échantillon.

- Quels risques ?

  Archiver des données d'enquêtes comporte un risque à double titre : pour le chercheur et pour le(s) enquêté(s). Le risque pour le chercheur concerne la mise en cause de sa recherche ou sa réfutation. Pour les enquêtés la rupture de l'anonymat pourrait causer des problèmes juridiques et briser le pacte de confiance tacite qui scelle les contacts entre chercheurs et enquêtés. Comme le notent les auteurs et comme les entretiens le confirment, cet argument est avancé en premier lieu par les chercheurs qui hésitent à confier leurs enquêtes à une banque d'archivage[78].

  Mais il existe également deux autres risques d'ordre épistémologique : la standardisation des pratiques de recherche et l'erreur d'interprétation inhérentes à la décontextualisation des données. Quant au premier de ces risques, il est lié directement à une plus grande visibilité des méthodes : les chercheurs opérant par des méthodes plus originales, surtout les plus jeunes, risquent de renoncer à s'écarter des méthodes dominantes pour ne pas avoir à se justifier. Le deuxième risque concerne un vieux débat toujours vivant, mais très caricaturale, entre une école « positiviste », pour laquelle l'archivage des enquêtes ne pose pas de problème si accompagnées d'une documentation sérieuse et une école qui estime que les données ne peuvent pas être séparées de leur contexte de production.

[77] DUCHESNE S., GARCIA G., op. cit., p. 4.
[78] DUCHESNE S., GARCIA G., op. cit., p. 7.

- Quels modes de fonctionnement ?

  Dans son étape préparatoire, ce projet s'était tout d'abord inscrit dans le paysage des TGIR des SHS (ce qui est actuellement le cas avec son inscription dans l'EQUIPEX Dime-SHS) et avait été conçu doté d'un dispositif d'auto-archivage à interface simple, permettant au chercheur de déposer des enquêtes après se voir proposer de remplir un module d' « enquête sur l'enquête », contextualisant le mieux possible ces données dès le premier dépôt. L'équipe beQuali devrait assurer ensuite un enrichissement supplémentaire de la forme et de la description de ces données. Un prototype de site, *enQuêtes*[79], avait été réalisé pour une mise à disposition rapide de ces enquêtes dès les premières demandes de réanalyse.

5.1.2 Publier des données : évolution de la question des données vers le cercle vicieux des problèmes de publication ?

Est-ce que publier des données en revue (sous modèles fermées ou semi-fermées) et «ouvrir» les données de la recherche en tant qu'injonction institutionnel va se traduire par une tension du même type que celle existante entre la publication en open access et en accès payant pour la publication d'articles? Il est d'autant plus difficile de répondre à cette question que nombreux les facteurs à prendre en compte : 1) le chercheur qui souhaite publier article et données appuyant ce dernier, ne souhaite le faire peut être qu'à titre de justificative épistémologique et non pas en vue de proposer une réutilisation de ces données ; 2) dans ce cas ces données devront être protégées par le droit d'auteur au même titre que l'article ; 3) mais cela peut impliquer un énorme conflit avec l'environnement institutionnel de la recherche dont la proposition principale est de « décloisonner » les données » en vue de leur réutilisation

En se référant à la publication des données, certains chercheurs ont réfléchi dans des termes présentant une analogie avec les formes traditionnelles de la publication scientifique et les problématiques qui en découlent, notamment l'accès, mais aussi l'importance d'être cité correctement.

Le *Data metrics* dans les années à venir ?

A ce stade, nous pouvons prolonger cette réflexion à un autre facteur central qui en découle toujours par analogie à la logique de la publication scientifique : publier les données auraient un impact sur la réputation des chercheurs (et de la revue, mais ceci n'a pas été évoqué

[79] enQuêtes, < http://www.bequali.fr/app/>

dans les entretiens), cette fois en termes de bibliométrie ou, à présent, de « datamétrie » (data metrics)

Au Royaume-Uni et aux EUA, où les réflexions et actions de politiques de gestion des données existent depuis une décennie déjà, les effets commencent à être mesurables et les développements de systèmes alternatifs (les Altmetrics) pour constater les formes d'usage font partie actuellement des principaux programmes soutenus par les organismes de recherche et acteurs comme CDL, Plos, etc. Ci-dessous un descriptif du projet pilote américain :

"Making Data Count : Developing a Data Metrics Pilot"

Projet pilote pour le développement d'un DLM (data-level metrics) lancé en septembre 2014, à l'initiative de la California Didital Library, PLOS, DataONE.

Ce projet tiendra dans la conception et prototypage une diversité de systèmes de mesures de traçabilité et mesures de l'utilisation des données. Le détail de ce projet est disponible en ligne : <http://escholarship.org/uc/item/9kf081vf>

Objectif de ce projet pilote :

Partager les données est une activité qui demande du temps et les chercheurs ont besoin d'incitations pour entamer ce travail. Développer des DLMs permettra d'avoir des retours sur l'usage, consultations, impact des données encourageant les chercheurs à partager leurs données. Ce projet va explorer et tester des systèmes de mesures des données, nécessaires pour enregistrer les différentes activités entourant les données de la recherche.

Ce projet poursuit un autre objectif fondamental : améliorer la compréhension de l'impact du partage et utilisation des données de la recherche dans l'écosystème scientifique. Il vise à montrer également comment le facteur de traçabilité des données, rendues ainsi plus visibles par les data-metrics, pourra jouer un rôle important pour les chercheurs, permettant la découverte et la réutilisation de données fiables.

Un objectif à plus long terme consiste à augmenter l'infrastructure en ligne en termes de « matériaux pour la recherche », disponible pour les étudiants et chercheurs, aujourd'hui trop centrée sur les articles de revue.

Il sera développé sur les bases des résultats positifs du projet collectif open source lancé par PLOS en 2009, « Article-Level Metrics ». ALM fournit une vision sur l'activité autour d'un article après sa publication, à travers un large spectre de formes d'utilisation et diffusion (par exemple, consultation, partage, commentaires, citation, recommandations).

Qu'est-ce qu'un Data Journal ?

Dans nos sociétés de l'information les données occupent une place importante. La publication des données doit pouvoir remplir des objectifs de :

- Signalement ;
- accessibilité ;
- réutilisabilité.

Le contexte actuel des pratiques et initiatives dans le milieu de la recherche, montre un certain nombre de problèmes d'attribution, de reconnaissance des producteurs de données et des pratiques peu précises de citation. Les *data journals* se posent comme une solution à ces problèmes.

Dans une étude publiée en 2014[80], les auteurs analysent et comparent 116 data journals publiés par 15 éditeurs[81]. Comme le nom l'indique, un data journal est une forme de publication scientifique des données et jeux de données dont l'objectif général est de rendre les données propres au partage et à la réutilisation. L'objectif final est de fournir « information on the what, why, how and who of the data »[82]. Les datas journals sont des initiatives qui souhaitent répondre aux besoins des chercheurs de disposer de données fiables, pouvant être citées correctement et réutilisées.

Les *datas papers*, articles des data journals décrivent les jeux de données et apportent a minima les informations suivantes :

- Nature des données
- Contexte de production ou collecte
- Les raisons de leur production
- Les moyens
- Les acteurs
- Le DOI donnant accès aux données hébergées dans des plate-formes ou bases de données. Les Data Journal n'hébergent pas, en règle générale, ces données.

Pour les éditeurs scientifiques, cette alternative rend plus facile la gestion des données dont les tâches principales seront prises en charge par le centre ou plateforme externe d'hébergement de ces données. En effet, auparavant les données étaient jointes aux articles sous forme de fichiers de données. Non seulement la gestion de tous ces fichiers était assez

[80] Les remarques à suivre se fondent sur l'article de CANDELA Leonardo; CASTELLI Donatella; MANGHI Paolo; TANI Alice. « Data Journals: A Survey » Preprint of the article accepted for publication in Journal of the Association for Information and Science Technology, June 2014, DOI : 10.1002/asi.23358. En ligne : <http://www.academia.edu/9635624/Data_Journals_A_Survey>

[81] Quelques éditeurs parmi les plus connus considérés par cette étude: BioMed Central, Chemistry Central, Pensoft Publishers, SpringerOpen, Ubiquity Press, PLOSOne.

[82] Op.cit. p. 2.

compliquée et engendrait des couts supplémentaires comme, en plus, les objectifs de partage et de réutilisation se trouvaient assez compromis.

Un data paper est à la fois deux choses : une information en soi (en tant que document) et une information sur les données qu'y sont décrites. La forme du data paper ne diffère pas essentiellement de celle d'un article scientifique commun : le data paper possède un titre, des auteurs, un résumé, une table de matières, et pourra bénéficier d'une description au même titre qu'un article d'une revue traditionnelle. Les métadonnées du data paper et des données sont, toutefois, différentes et doivent être gérées séparément.
Le data paper ne doit contenir aucune mention aux résultats de la recherche qui a produit ces données, l'objectif étant de promouvoir des nouvelles exploitations des jeux de données (modalités de réutilisations potentielles) et une citation complète (méthodes, protocoles, formats). Il n'existe pas des standards de formats pour ce type de publication apparue très récemment. Mais il est convenu que les informations suivantes doivent figurer comme prioritaires dans la description des jeux de données: disponibilité, particularités (relatives au contexte de collecte ou production), périmètre, format, licence, nom de tous les contributeurs, description du projet, provenance, qualité, formes de réutilisations.

Ce nouveau phénomène laisse entrevoir quelques tendances :

- Les éditeurs se déchargent progressivement de la gestion des données, proposant souvent un lien vers des bases de données ou centre de données ; ces dernières, institutionnelles ou privées, devront en conséquence se développer dans les années à venir ;
- Les données acquièrent une valeur à part entière et non seulement une valeur qui découle de la publication des résultats de la recherche ;
- Le data metrics se développe en parallèle et peut fournir 1) des retours d'investissement aux éditeurs ; 2) notoriété au chercheur ;
- Les Data Journals ont un rôle à jouer dans la qualité de l'amélioration de la citation des jeux de données, terrain qui reste encore ouvert et qui devra gagner un contour plus précis à court terme.

Droit d'auteur / Open Science et Open data ?

Et finalement, comme nous avions déjà avancé au tout début de cette étude, les années à venir vont ramener sur la place publique le droit de la propriété intellectuelle concernant les données[83].

[83] http://libereurope.eu/news/liber-statement-on-enabling-open-science/

Le cadre juridique déterminant le périmètre de réutilisation des données est le cœur de la problématique d'ouverture, les chercheurs doivent se positionner clairement sur ce sujet.

Tout l'enjeu dans les prochaines années va se concentrer, comme pour les articles, mais de façon maintenant plus fondamentale (s'agissant des « données brutes »), dans le statut juridique des données. D'une part des programmes gouvernementaux « Open » et les impulsions institutionnelles consécutives seront sûrement le contexte avec lequel les chercheurs devront apprendre à travailler.

Nous n'avons pas pu développer dans le cadre de cette étude une approche en détail à ces questions, mais proposons de poursuivre une veille active sur ces sujets à travers un blog qui est en construction à l'adresse https://anteoculos.wordpress.com.

Méthodologie des entretiens

1) Contexte et objectifs

La mission de stage à Cairn.info

Les mois de juin et juillet 2014 ont été dédiés à temps complet à la préparation et à la réalisation des entretiens. Ces étapes ont bénéficié de l'accompagnement du responsable de stage et du soutien de l'équipe de Cairn.info.

La mission avait été définie au préalable dans les termes suivants :

« Étude prospective sur les données de la recherche en SHS : modalités de collecte et de mise à disposition de la communauté scientifique, usages, liens avec les publications, infrastructures, particularités des approches disciplinaires ».

Les étapes de recherche exploratoire et détermination d'axes de veille réalisées pendant les mois précédents ont pu constituer un terrain différencié pour passer en suite à l'élaboration de quelques questions destinées à connaître les pratiques et les réflexions portant sur les données de la recherche en SHS.

Détermination du type d'entretiens : exploratoires semi-dirigés

Les entretiens ont été conçus pour répondre à des questions assez larges, suivant un mode exploratoire, c'est-à-dire, partant de quelques problématiques et hypothèses non testées, identifiées et formulées partiellement lors de l'étude du contexte de l'actualité (étape de veille précédente).

Ces problématiques sont présentées aux chercheurs dans le but d'en apprécier leur portée concrète et l'impact réel de ces questions sur leurs pratiques et sur leurs attentes. Nous avons proposé des questions thématiques ouvertes, c'est-à-dire, laissant aux chercheurs une grande liberté de développer leur réponse dans le sens choisi par eux. A travers ma propre participation (demande de précisions, transitions, commentaire) il a été possible d'exercer une orientation souple permettant de cadrer la conversation sur les sujets centraux.

Objectifs généraux

- tester les problématiques repérées dans l'étape de veille;
- mettre en lumière d'autres aspects de ces problématiques;

- dégager des thèmes et problématiques nouvelles
- enrichir et approfondir l'approche du départ

Objectifs spécifiques

- connaître les pratiques de chercheurs visant les données de la recherche
- connaître l'état de leurs réflexions sur le sujet et l'existence de thèmes récurrents
- connaître les usages, les motivations et contraintes liées à cet usage
- identifier des tendances

Description (mots-clés) : données de la recherche, données numériques, Sciences humaines et sociales, France

Couverture géographique : France

Couverture thématique : disciplines des Sciences humaines et sociales

Unité d'analyse : chercheur

Période de collecte de données : Juin/juillet/août 2014

Méthode d'élaboration des questions

Nous avons adopté une démarche du type « entonnoir » (fig.8), c'est-à-dire, partant des sujets généraux qui gravitent autour des données de la recherche, rétrécissant leur périmètre à travers la définition d'une problématique ayant une importance suggérée par nos recherches documentaires préalables (hypothèse), pour aboutir à la formulation des questions assez larges.

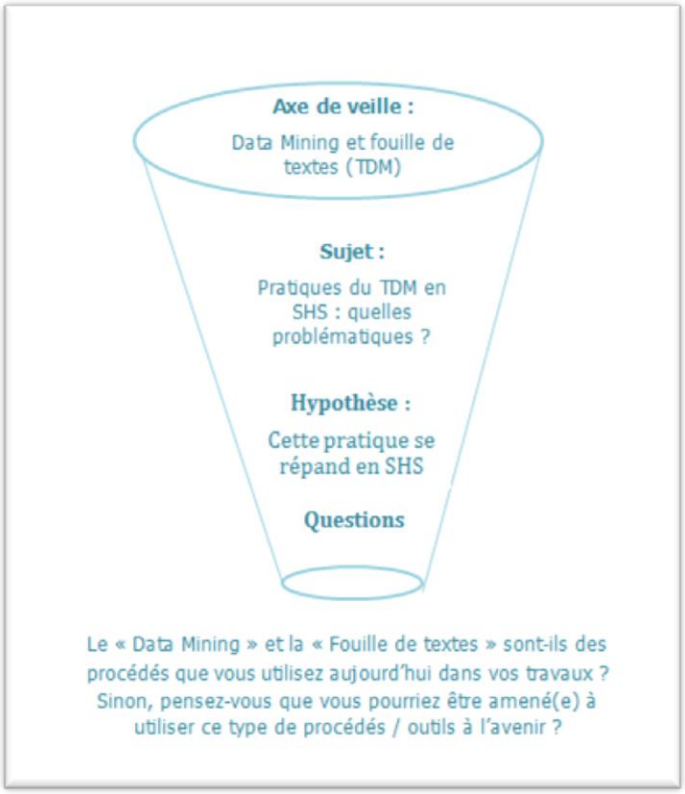


Figure 8 - Démarche « entonnoir » de la veille jusqu'à l'élaboration des questions

2) Procédure d'échantillonnage et modes de collecte

Objectif	50 entretiens dans 18 disciplines des SHS **Le nombre minimum d'entretiens a été fixé par le responsable du stage à 50 entretiens. Une représentativité des disciplines des SHS devrait pouvoir être trouvée à l'intérieur de ce chiffre de départ, ce qui obligeait à ne pas tenir compte des différentes spécialités à l'intérieur de chaque catégorie disciplinaire. Nous avons pris appui, ainsi, sur la répartition des disciplines dans le portail Cairn.info[84].**
Profil des chercheurs	L'échantillonnage des chercheurs à contacter a été réalisé à partir d'impressions obtenues dans la phase de veille de cette étude. Nous avons pu repérer différents milieux d'activité, communautés de chercheurs et statuts de ces derniers et avons essayé d'obtenir une représentativité des éléments suivants : 1) Environnement de travail : les milieux de recherche se partagent notamment entre les universités, grandes écoles, centres et laboratoires de recherche

[84] Cair.info, http://www.cairn.info/

	interdisciplinaires. 2) Statut des chercheurs : maître de conférence, chargé(e) de recherche, directeur de recherche, ingénieur. Obs.: a) Le critère d'âge n'est pas exploité ici, les chercheurs contactés n'étaient pas tenus à déclarer leur âge. b) Souhaitant augmenter les possibilités de réaliser des entretiens en face-à-face, la plupart de ceux-ci se localise en Ile-de-France, mais le critère géographique ne nous semble pas déterminant pour l'étude actuelle.
Mode de sélection	Les chercheurs ont été sélectionnés notamment par deux moyens : via l'annuaire des sites institutionnels d'appartenance et à travers du portail Cairn.info.
Mode de collecte des données	Face-à-face, téléphone, questions envoyés par mail Le premier échange lors de la demande d'entretien visait à énoncer le sujet, le contexte, la raison d'être de ces entretiens et de proposer une modalité d'échange, le face à face avancé comme première option. L'enregistrement a été écarté comme option en commun accord avec le tuteur du stage qui encadrait cette opération. L'objectif principal était une conversation libre, semi-dirigée par les questions qui ont émergé de la veille et des recherches documentaires sur les données. Pour les entretiens face-à-face, des notes ont été prises par intervalles visant à ne pas couper le flux de la conversation. Ces notes sont ensuite enrichies dès la fin de l'entretien, transcrites et jalonnées par des rubriques qui permettent une lecture rapide. Pour les entretiens téléphoniques, la prise de notes a été semblable avec la seule différence qu'elle pouvait se faire simultanément lorsque l'interlocuteur était en train de s'exprimer.
Matériel en Annexe	- questions par mail - exemple de grilles d'analyse - entretiens anonymisés

3) Lancement des entretiens et collecte des réponses

A la fin du mois de juillet 2014 nous avons pu réaliser une totalité de 53 entretiens en respectant l'objectif de la diversité disciplinaire. Le tableau ci-dessous illustre le nombre d'entretiens réalisés par discipline.

DISCIPLINE	ENTRETIENS REALISES
Anthropologie/Ethnologie	1
Anthropologie sociale et culturelle	1
Archéologie	3
Arts, architecture, danse et musique	3
Démographie, démographie historique	2
Droit	3
Economie	5
Géographie/Géopolitique	1
Histoire	7
Histoire de l'Art	1
Lettres et littérature	1
Linguistique et Sémiologie/Sciences du langage	2
Philosophie	3
Psychologie	1
Sciences administratives	2
Sciences de l'éducation	3
Sciences politiques	4
Sociologie et Societé	7
Sciences de l'information et de la communication	3
TOTAL	53

Suivant les critères auxquels nous avons essayé de répondre, le profil des répondants s'est partagé comme de la manière suivante :

PROFIL DES REPONDANTS

STATUT DES REPONDANTS
18 chercheurs/maîtres de conférence **16 directeurs de recherche** **9 chargés de recherche**

5 chercheurs/professeurs d'université

4 chercheurs à statuts divers (associé, en contrat doctoral, ATER)

1 chercheur indépendant

1 ingénieur d'études

1 professeur associé

Conclusion : Les données de la recherche, quels rôles pour les documentalistes ?

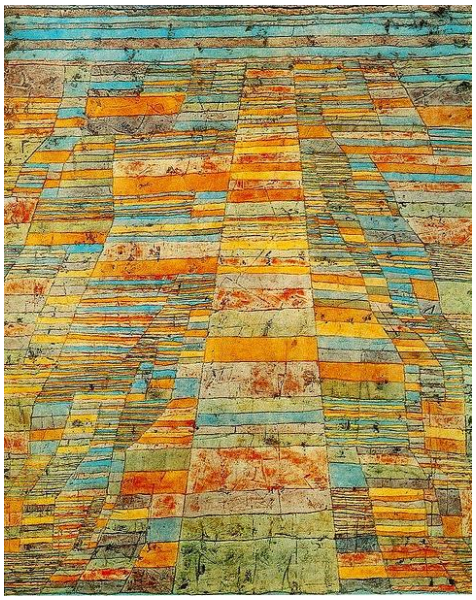


Figure 9 - Highway and Byways (1929) - Paul Klee - CC-PD

En travaillant tout au long de ce mémoire sur un sujet si riche comme les données de la recherche en SHS, nous avons pu constater que la réflexivité sur le sujet dans le champ de SHS est très variable à l'intérieur de la communauté scientifique des chercheurs en France. Les entretiens ont notamment révélé des conditions de travail très différentes au sein de cette communauté, les chercheurs travaillant en équipe étant plus réceptifs aux enjeux ouverts par les technologies du numérique en ce qui concerne la valorisation et le partage des données. Le milieu de travail interdisciplinaire des laboratoires de recherche semble un terrain plus propice à la mise en place de plans et de programmes de gestion et valorisation des données en raison des différentes pratiques et cultures qui s'y trouvent impliquées, incitant les chercheurs à se confronter à des questions qui, autrement, ne seraient pas forcement posées à l'intérieur de leur propre discipline. Ces laboratoires bénéficient aussi d'une plus grande visibilité et de soutiens institutionnels plus conséquents, ceci jouant un rôle déterminant dans la sensibilisation du chercheur à ces problématiques. Pourtant, même à l'intérieur de ce cadre, des différences subsistent : les laboratoires de recherche sont divers (par leur taille, notamment), ont des stratégies numériques très variables (formes de stockage, notamment) et présentent des grands écarts dans l'appropriation des technologies de valorisation de ces données.

Ce qui nous amène à conclure aussi que les stratégies du CNRS et de l'Inist visant à instaurer une « économie des données durable » dans le champ des SHS va se confronter aux difficultés provenant de cette disparité des conditions matérielles de travail et aux vitesses variables des transformations des pratiques à l'heure du numérique. Les stratégies de communication et de promotion dans le contexte des SHS, qui vont sans doute s'intensifier à court et moyen terme, ne produiront des effets à large échelle que si elles sont accompagnées de propositions provenant des propres communautés scientifiques, visant à combler les écarts techniques, technologiques et culturels avec un transfert de technologies et de méthodologies, instaurant ainsi une dynamique de coopération entre les différents domaines des SHS. Les groupes de travail (consortiums) d'Huma-Num[85] s'inscrivent dans cette perspective et ce type d'initiative pourra se démultiplier à court terme dans d'autres cercles scientifiques. Il serait souhaitable que ces initiatives puissent s'adresser à des équipes moins versées dans les aspects technologiques, moins favorisées par des infrastructures matérielles et devant entamer une réflexion sur les données à partir de zéro.
Il est néanmoins peu réaliste d'imaginer que les chercheurs peuvent être la seule interface pour ces transferts et se charger de toutes les étapes de travail que demandent le traitement et la valorisation des données. Pour les projets qui doivent encore être conçus, les compétences en conduite de projet des professionnels de l'information seront de première importance. Pour les projets en cours et ceux qui démarrent, ces professionnels

[85] Huma-Num. Consortiums | http://www.huma-num.fr/service/consortium

auront un rôle important pour assurer la structuration de ces données, leur description et une réflexion stratégique sur les moyens de garantir leur pérennité dans une indépendance de l'outil utilisé. Ce point nous semble particulièrement essentiel : au moment où le métier du documentaliste et celui du chercheur sont à cent pour cent réalisés dans un environnement numérique et reposent sur des technologies informatiques évolutives et contraignantes, les compétences des documentalistes s'insèrent dans l'horizon du durable, inscrites dans une réflexion sur les finalités, les usages et les concepts.

Par ailleurs, les professionnels de l'information que nous devenons aujourd'hui doivent notamment s'appliquer à développer en continu leur potentiel de veille juridique au plan national et international, de compréhension politique de l'environnement de la recherche scientifique, de connaissance concrète des pratiques de la recherche de l'unité ou laboratoire où ils souhaitent travailler. Ils doivent en outre s'ouvrir à ces contextes dans d'autres pays où les communautés scientifiques, les institutions et gouvernements répondent autrement à ces enjeux. Tous les ingrédients sont ainsi réunis pour que notre métier soit valorisé en tant que force de proposition et éclaireur au cours de débats effervescents à venir sur les données de la recherche et la construction d'une science commune.

Bibliographie

ABRIEUX Claire. « Small Data contre Big data: quand David rencontre Goliath? » in *Regards sur le numérique,* 05 mars 2012. [En ligne] http://www.rslnmag.fr/post/2012/03/05/Small-data-contre-Big-Data-quand-David-rencontre-Goliath-.aspx [consulté le 18/04/2014]

Académie des Sciences. « La Datamasse : directions et enjeux pour les données massives », conférence-débat du 18 février 2014, vidéos [En ligne] http://www.academie-sciences.fr/video/v180214.htm

BASTARD Irène ; CARDON Dominique ; FOUETILLOU Guilhem; PRIEUR Christophe ; RAUX Stéphane. « Travail et travailleurs de la données », Internet Actu, publié le 13/12/13 [En ligne] http://www.internetactu.net/2013/12/13/travail-et-travailleurs-de-la-donnee/ [Consulté le 1 janvier 2015]

BERMES Emmanuelle; avec la collaboration d'ISAAC Antoine et POUPEAU Gautier. Le Web sémantique en bibliothèque, Paris,Ed. du Cercle de la Librairie, 2013.

BERRY David. "The Computational Turn: Thinking About the Digital Humanities", Culture Machine, vol 12, July 2011. [En ligne] http://www.culturemachine.net/index.php/cm/article/view/440/470 [Consulté le 6 novembre 2014]

Big data : zoom sur les usages et les applications [Dossier] In Le monde informatique, 02 avril 2014 [En ligne] http://www.lemondeinformatique.fr/les-dossiers/sommaire-lire-big-data-zoom-sur-les-usages-et-les-applications-137.html?utm_source=mail&utm_medium=email&utm_campaign=Newsletter [consulté le 18 avril 2014]

BISHOP Libby. "Ethical Sharing and re-use of qualitative data". In *Australian Journal of Social Issues,* 44 (3), pp.: 255-72. [En ligne] http://www.data-archive.ac.uk/media/249157/ajsi44bishop.pdf [Consulté le 1 octobre 2014]

BOLLIER David. "The Promise and the Peril of Big data", The Aspen Institute *Communication and Society Program*, Washington DC: 2010. [En ligne] http://www.aspeninstitute.org/sites/default/files/content/docs/pubs/The_Promise_and_Peril_of_Big_Data.pdf [Consulté le 30 novembre 2014]

BOLTANSKI Luc ; CHIAPELLO Ève. Le nouvel esprit du capitalisme (1999), Gallimard, 2011.

BOURDELOIE Hélène. « Ce que le numérique fait aux sciences humaines et sociales », *tic&société*, Vol. 7, N° 2 | 2ème semestre 2013, mis en ligne le 31 mai 2014. [En ligne] http://ticetsociete.revues.org/1500 [Consulté le 06 novembre 2014]

BOYD Danah; CRAWFORD Kate. *Six provocations à propos des Big data* [Six Provocations for Big data, 2011], traducteur Laurence Allard, Pierre Grosdemouge et Fred Pailler, 2012 [en ligne] http://books.openedition.org/oep/273?lang=fr#ndlr[consulté le 30 novembre 2014]

BURNARD Lou. *What is the Text Encoding Initiative? How to add intelligent markup to digital resources.* New edition [online]. Marseille: OpenEdition Press, 2014 (generated 23 September 2014). Available on the Internet: <http://books.openedition.org/oep/426>. ISBN: 9782821834606.

CANDELA Leonardo; CASTELLI Donatella; MANGHI Paolo; TANI Alice. « Data Journals: A Survey » Preprint of the article accepted for publication in Journal of the Association for Information and Science Technology, June 2014, DOI : 10.1002/asi.23358 [En ligne] http://www.academia.edu/9635624/Data_Journals_A_Survey [consulté le 1 janvier 2015]

CARDON Dominique. « Utilisation du Big data dans les anlayses sociologiques ». Intervention du 22 mai 2014 à la Société Française de Statistique [Vidéo en ligne] http://www.dailymotion.com/video/x1zvszx_big-data-intervention-de-dominique-cardon-sfds-22mai-2014_news [Consulté le 2 novembre 2014]

CARDON Dominique. « Zoomer ou dézoomer? Les enjeux politiques des données ouvertes » in OWNI, 21 février 2011 [En ligne] http://owni.fr/2011/02/21/zoomer-ou-dezoomer-les-enjeux-politiques-des-donnees-ouvertes/ [Consulté le 2 novembre 2014]

CASILLI Antonio. *Comment les usages numériques transforment-ils les sciences sociales ?* In : *Read/Write Book 2 : Une introduction aux humanités numériques* [en ligne]. Marseille : OpenEdition Press, 2012 (généré le 01 janvier 2015). Disponible sur Internet : http://books.openedition.org/oep/286>. ISBN : 9782821813250.

CDSP. *Enquêtes internationales. Eurobaromètre.* Site du CDSP - Sciences PO http://cdsp.sciences-po.fr/enquetes.php?idTheme=2&idRubrique=enquetesINT&lang=FR [Consulté le 6 novembre 2014]

CEE - Sciences Po. *Dynamiques de mobilisation (Dynamob).* http://www.cee.sciences-po.fr/en/research/election-analysis/dynamiques-de-mobilisation-dynamob.html [Consulté le 6 novembre 2014]

CHARTRON Ghislaine. «Open access et SHS : Controverses », ArchiveSIC, version pré-print déposé le 24/03/2014. [En ligne] http://archivesic.ccsd.cnrs.fr/docs/00/96/52/72/PDF/Ghislaine-Chartron_RSS-fA_vrier2014-preprint.pdf [consulté le 6 janvier]

CHENU Alain; LESNARD Laurent (dir.) *La France dans les comparaisons internationales. Guide d'accès aux grandes enquêtes statistiques en sciences sociales*, Paris, Les Presses de Sciences Po, 2011.

CHIGNARD Simon. "Donnée brute ou donnée contextualisée?". In Données ouvertes [site], publié le 2 mai 2013. [En ligne] http://donneesouvertes.info/2013/05/02/donnee-brute-ou-donnee-contextualisee/[consulté le 14 novembre 2014]

CNRS Comité d'Ethique. *Promouvoir une recherche intègre et responsable: un guide.* Juillet, 2014. [En ligne] http://www.cnrs.fr/comets/IMG/pdf/guide_promouvoir_une_recherche_inte_gre_et_responsable8septembre2014.pdf [Consulté le 10 novembre 2014]

CNRS. Ouverture des données massives scientifiques. Quels risques, quels bénéfices ? Programme du Rencontre interdisciplinaire, 19 avril 2014, dans le cadre de la mission Sciences et citoyens du CNRS http://www.iscc.cnrs.fr/spip.php?article1818

CNRS. Dossier :*Big data, la déferlante des octets* in *CNRS Le Journal*, paru le 22/02/2014 [consulté le 18/04/2014]. En ligne : <https://lejournal.cnrs.fr/dossiers/big-data-la-deferlante-des-octets>

COENEN Frans. Data mining :past, present and future. The Knowledge Engineering Review, Vol. 26:1, 25–29.& Cambridge University Press, 2011 <doi:10.1017/S0269888910000378>

CORTI Louise; VAN DEN EYNDEN Veerle; BISHOP Libby; WOOLLARD Matthew. *Managing and Sharing Research Data: A Guide to Good Practice***.** SAGE: UK, 2014; p. 222.

CORTI Louise; WITZEL Andreas ; BISHOP Libby. On the Potentials and Problems of Secondary Analysis. An Introduction to the FQS Special Issue on Secondary Analysis of Qualitative Data [13 paragraphs]. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, Vol. 6, No 1 , 2005, Art. 49. [En ligne] http://nbn-resolving.de/urn:nbn:de:0114-fqs0501495 [Consulté le 6 novembre 2014]

COSTAS Rodrigo ; MEIJER Ingeborg ; ZAHEDI Zohreh ; WOUTERS Paul. The Value of Research Data - Metrics for datasets from a cultural and technical point of view. A Knowledge Exchange Report, 2013. [En ligne] www.knowledge-exchange.info/datametrics [consulté le 10 novembre 2014]

CSPLA. Missions CSPLA relative au "text and data mining" (exploration des données), juillet 2014. [En ligne] http://www.culturecommunication.gouv.fr/Politiques-ministerielles/Propriete-litteraire-et-artistique/Conseil-superieur-de-la-propriete-litteraire-et-

artistique/Travaux-du-CSPLA/Missions/Mission-du-CSPLA-relative-au-text-and-data-mining-exploration-de-donnees

DACOS Marin ; MOUNIER Pierre. *Humanités numériques – État des lieux et positionnement de la recherche française dans le contexte international*, Rapport commandé par l'Institut Français opérateur du ministère des Affaires étrangères, mars 2014. [En ligne] http://www.institutfrancais.com/sites/default/files/if_humanites-numeriques.pdf

DCC/JISC. A Comparative Study of International Approaches to Enabling the Sharing of Research Data. Version 1.6, 30 novembre 2008 [En ligne] http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/Data-Sharing-Report.pdf [consulté le 26 octobre 2014]

DUCHESNE Sophie ; GARCIA Guillaume. « beQuali : une archive qualitative au service des sciences sociales » in M. CORNU, J. FROMAGEAU et B. MÜLLER. *Archives de la recherche. Problèmes et enjeux de la construction du savoir scientifique*, l'Harmattan, pp.35-56, 2014, droit du patrimoine culturel et naturel, 978-2-343-03247-4. [En ligne] <halshs-00922690> [Consulté le 20 octobre 2014]

DUMON Olivier. « Innovation scientifique & Big data : comment gérer un tsunami de "milliards et milliards" de données » Le Huffington Post, 31 octobre 2014 [En ligne] http://www.huffingtonpost.fr/olivier-dumon/comment-gerer-big-data_b_6066246.html [consulté le 2 décembre 2014]

ETALAB. Vade-mecum sur le partage des données publiques, Septembre 2013. [En ligne] http://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/vademecum-ouverture.pdf

FRIEDMAN Jerome. H. "Data Mining and Statistics: What's the Connection?" , Proc. of the 29 th Symposium of the Interface: Computing Sciences and Statistics, May 1997, Houston, Texas. [En ligne] http://statweb.stanford.edu/~jhf/ftp/dm-stat.pdf[Consulté le 30 novembre]

GAILLARD Rémi. *De l'Open Data à l'Open Research Data: quelle (s) politique (s) pour les données de recherche*. Mémoire d'étude/janvier 2014 pour l'obtention du diplôme de Conservateur de bibliothèque, ENSSIB [En ligne] http://www.enssib.fr/bibliotheque-numerique/notices/64131-de-l-open-data-a-l-open-research-data-quelles-politiques-pour-les-donnees-de-recherche [Consulté le 14 novembre 2014]

GENET Jean-Philippe; DESROSIERES Alain. La politique des grands nombres. Histoire de la raison statistique. In: Histoire & Mesure, 1996 volume 11 - n°1-2. Varia. pp. 167-173. [En

ligne] http://www.persee.fr/web/revues/home/prescript/article/hism_0982-1783_1996_num_11_1_1724 [Consulté le 6 novembre 2014]

GESIS. *The Eurobarometer Survey Series*, mise à jour le 20 décembre 2013 http://www.gesis.org/en/eurobarometer/survey-series/ [Consulté le 6 novembre 2014]

GOËTA Samuel. "Collectionner des données ou expérimenter: une querelle des Anciens et des Modernes?" in Ar(*abes*)ques, n° 73 (jan-fev-mars 2014), pp.10-11. [En ligne] http://www.abes.fr/Arabesques/Arabesques-n-73 [Consulté le 18 novembre 2014]

GOËTA Samuel. Les coulisses de l'open data, Carnet de recherches."Raw data is an Oxymoron" : les données brutes sont-elles une fiction ? Recension [version pré print] de l'ouvrage Raw Data is An Oxymoron, Lisa GITELMAN (dir.), Cambridge, MIT Press, 2013, 182p. Billet publié le 30 avril 2013 [En ligne] http://www.coulisses-opendata.com/2013/04/30/raw-data-is-an-oxymoron-les-donnees-brutes-sont-ell(es-une-fiction/ [Consulté le 18 novembre 2014]

GUICHARD Éric. « L'internet et les épistémologies des sciences humaines et sociales », *Revue Sciences/Lettres* [En ligne], 2 | 2014, mis en ligne le 24 février 2014, consulté le 19 avril 2014. En ligne: http://rsl.revues.org/389 [Consulté le 19 octobre 2014]

GUILLAUD Hubert. "Big data: vers l'ingéniérie sociale?, InternetActu.net, article paru le 20 mai 2014. [En ligne] http://www.internetactu.net/2014/05/20/big-data-vers-lingenierie-sociale/ [consulté le 1 janvier 2015]

GUILLAUD Hubert. « De la statistique aux Big data : ce qui change dans notre compréhension du monde » [En ligne] http://www.internetactu.net/2012/12/19/de-la-statistique-aux-big-data-ce-qui-change-dans-notre-comprehension-du-monde/]

HORIZON 2020. Programme-cadre européen pour la recherche et l'innovation Horizon 2020: Lignes directrices pour le libre accès aux publications scientifiques et aux données de recherche dans Horizon 2020 [titre original : Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020], Version 1.0, 11 décembre 2013, traduction française INIST-CNRS, France, 2014.
[En ligne]
http://openaccess.inist.fr/IMG/pdf/14086_lignes_directrices_la_horizon_2020_tr_fr_version-oct2014-2.pdf

HORIZON 2020. Programme-cadre européen pour la recherche et l'innovation Horizon 2020: Lignes directrices pour la gestion des données dans Horizon 2020 [titre original : Guidelines on Data Management in Horizon 2020] , version 1.0, 11 décembre 2013, traduction

française INIST-CNRS, France, 2014. [En ligne]
http://openaccess.inist.fr/IMG/pdf/lignes_directrices_pgd_horizon_2020_tr_fr.pdf

HORIZON 2020. The EU Framework Programme for Research and Innovation
http://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020

IFDO. « IFDO Report 2014: Policies for Sharing Research Data in Social Sciences and Humanities – DASISH ». [En ligne] http://dasish.eu/news/2014/ifdoreport/ [Consulté le 26 octobre 2014]

INIST. "Science Europe : traduction française des principes pour assurer la transition vers le Libre Accès", publié le 16 mai 2013, par Thérèse HAMEAU. [En ligne] http://openaccess.inist.fr/?Science-Europe-traduction [Consulté le 18 octobre 2014].

INRA, Observatoire des Technologies. « Crossref va lancer Prospect : un service de text mining et *data mining* », mis en ligne le 28 décembre 2013 [En ligne] http://ist.blogs.inra.fr/technologies/2013/12/28/crossref-va-lancer-prospect-un-service-de-text-mining-et-data-mining/ [consulté le 17/04/2014].

JISC. The value and impact of data sharing and curation - synthesis of three recent UK studies. March, 2014 [En ligne] http://repository.jisc.ac.uk/5568/1/iDF308_-Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf

League of European Research Universities (LERU). LERU Roadmap for Research Data, Advice paper n°14 ; Paul Ayris and the Research Data Working Group, December 2013 [En ligne]
http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf
[Consulté le 26 décembre 2014]

Les enjeux éthiques du Big data Opportunités et risques. Journée du 22 mai 2014 à la Société Française de Statistique. [Interventions et programme en ligne]
http://www.sfds.asso.fr/ressource.php?fct=ddoc&i=1799

LIBER. « LIBER Responds To Elsevier's Text and Data Mining Policy », publié le 28 mars 2014 [En ligne] http://libereurope.eu/news/liber-responds-to-elseviers-text-and-data-mining-policy/[Consulté le 19 octobre 2014]

LIBER. « Ten recommendations for libraries to get started with research data management», Final report of the LIBER working group on E-Science / Research Data Management, 7 juillet 2012. [En ligne]

http://libereurope.eu/wpcontent/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

MANGHI Paolo. « Sfide tecnologiche per l'accesso aperto a tutti i prodotti della ricerca», in Bibliotime,XVII, 3, Novembre 2014. [En ligne] http://www.aib.it/aib/sezioni/emr/bibtime/num-xvii-3/manghi.htm [consulté le 1 janvier 2015]

MANOVICH Lev. « Trending: The Promises and the Challenges of Big Social Data », Uploaded 4/28/2011. [En ligne] http://manovich.net/index.php/projects/trending-the-promises-and-the-challenges-of-big-social-data [consulté le 1 janvier 2015].

MAUREL Lionel (Calimaq) « Le Royaume Uni sanctuarise les pratiques de data mining par une exception au droit d'auteur », S.I.Lex, 01 avril 2014 [En ligne] http://scinfolex.com/2014/04/01/le-royaume-uni-sanctuarise-les-pratiques-de-data-mining-par-le-biais-dune-exception-au-droit-dauteur/ [consulté le 31 décembre 2014]

MAUREL Lionel (Calimaq). "Digital Humanities, Propriété intellectuelle et Biens communs de la connaissance", S.I.Lex [blog], publié le 25 juin 2013. [En ligne] http://scinfolex.com/2013/06/25/digital-humanities-propriete-intellectuelle-et-biens-communs-de-la-connaissance/ [consulté le 14 novembre 2014]

Médialab. "Le tout est-il plus grand ou plus petit que ses parties? Regards de sociologues, économistes, physiciens, biologistes… », Colloque interdisciplinaire [vidéos en ligne] http://www.medialab.sciences-po.fr/blog/videos-du-colloque-interdisciplinaire-le-tout-est-il-plus-grand-ou-plus-petit-que-ses-parties-regards-de-sociologues-economistes-physiciens-biologistes/ [consulté le 1 janvier 2015]

MICHELAT Guy. Sur l'utilisation de l'entretien non directif en sociologie. In: *Revue française de sociologie*. 1975, 16-2. pp. 229-247. doi : 10.2307/3321036 [En ligne] http://www.persee.fr/web/revues/home/prescript/article/rfsoc_0035-2969_1975_num_16_2(_6864 [Consulté le 15 septembre 2014]

MINISTERE DE L'ENSEIGNEMENT SUPERIEUR ET DE LA RECHERCHE.. Horizon 2020: Le portail français du programme européen pour la recherche et l'innovation [Site] http://www.horizon2020.gouv.fr/cid82025/le-libre-acces-aux-publications-aux-donnees-recherche.html

Ministère de l'Enseignement supérieur et de la Recherche. Stratégie nationale - Infrastructures de recherche 2012-2020. Feuille de route, octobre 2012 [En ligne]

http://cache.media.enseignementsup-recherche.gouv.fr/file/TGIR/29/6/infras_def3_243296.pdf [Consulté le 10 novembre 2014]

MORAVCSIK Andrew. « Active Citation and Qualitative Political Science » in Qualitative & Multi-Method Research, Newsletter of the American Political Science Association, Spring 2012, Vol.10, no1, pp. 33- 37. [En ligne] https://www.princeton.edu/~amoravcs/library/active_citation.pdf

MORAVCSIK Andrew. « Transparency: The Revolution in Qualitative Research », Symposium, in American Political Science Association, January 2014. <doi:10.1017/S1049096513001789>

NAPOLETANI Domenico ; PANZA Marco ; STRUPPA Daniele. « Is Big data enough? A reflection on the changing role of mathematics in applications ». *Notices of the American Mathematical Society*, American Mathematical Society, 2014, 61 (5), pp.485-490. <halshs-00984828> [Consulté le 1 janvier 2015]

NAVARRO Jordi. « Faire de l'archiviste le maillon fort de l'open data ». In Papiers et poussières [blog], publié le 3 mai 2014. [En ligne] http://www.papierspoussieres.fr/index.php/2013/05/03/faire-de-larchiviste-le-maillon-fort-de-lopen-data/ [Consulte le 14 novembre 2014]

NOYER Jean-Max ; CARMES Maryse. « L'irrésistible montée de l'algorithmique : méthodes et concepts en SHS ». 2013. <sic_00911858>

NOYER Jean-Max ; CARMES Maryse. *Le mouvement « Open data » dans la grande transformation des intelligences collectives* In : *Les débats du numérique* [en ligne]. Paris : Presses des Mines, 2013 . [En ligne] http://books.openedition.org/pressesmines/1666 [Consulté le 6 novembre 2014]

OCDE. *OCDE Principes et lignes directrices pour l'accès aux données de la recherche financée sur fonds publics*. Organisation de coopération et de développement économiques, 2007. [En ligne] http://www.oecd.org/fr/science/sci-tech/38500823.pdf [Consulté le 12 novembre 2014]

OPEN ACCESS DIRECTORY (OAD). Data repositories http://oad.simmons.edu/oadwiki/Data_repositories

OPEN ACCESS DIRECTORY (OAD). Wiki compendium of simple factual lists about open access (OA) to science and scholarship, hosted by the Graduate School of Library and Information Science at Simmons College and supervised by an independent editorial board. [En ligne] http://oad.simmons.edu/oadwiki/Main_Page

OPEN AIRE. "Open access' a requirement for Horizon 2020", publié le 12/12/2013. En ligne: < https://www.openaire.eu/

PIERAZZO Elena. L'encodage des textes et la recherche en sciences humaines. Séminaire 2013-2014, L'Encodage du texte [vidéos][En ligne] http://msh.univ-tours.fr/content/l-encodage-du-texte [consulté le 14 novembre]

PLOS Opens. "Best Practice in Enabling Content Mining", 9 mars 2014 [En ligne] En ligne: http://blogs.plos.org/opens/2014/03/09/best-practice-enabling-content-mining/[consulté le 18 avril 2014]

POUYLLAU Stéphane. Web de données, Big data, open data, quels rôles pour les documentalistes? Manuscrit auteur, publié dans "Documentaliste - Sciences de l'Information Vol. 50 (2013) 32-33", ArchiveSIC, version pré-print déposé le 18/03/2014. En ligne [pdf] : <http://archivesic.ccsd.cnrs.fr/docs/00/96/08/53/PDF/Doc-SI_50-3-1_pouyllau-stephane_V1.pdf>

Research Information Network. *To Share or not to Share:Publication and Quality Assurance of Research Data Outputs*. Report commissioned by the Research Information Network (RIN), Juin 2008 [En ligne] http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs

RUTTER Tamsin. « Big data and open data: what's what and why does it matter? Both types of data can transform the world, but when government turns Big data into open data it's especially powerful », The guardian, 15 April 2014.[En ligne] http://www.theguardian.com/public-leaders-network/2014/mar/31/government-big-data-public-services-livechat [Consulté le 24 novembre 2014]

SARDAN Jean-Pierre Olivier. « La politique du terrain », *Enquête*, 1 | 1995, mis en ligne le 08 janvier 2007. [En ligne] http://enquete.revues.org/263 [Consulté le 15 septembre 2014]

SCHÖCH Christof. "Big? Smart? Clean? Messy? Data in the Humanities". *Journal of Digital Humanities*, 2013, 2 (3), pp.2-13. <hal-00920254>

Scholarly Communication & Research Infrastructures. "Research Data Management Case Studies". In LIBER [En ligne]http://libereurope.eu/committee/scholarly-research/research-data-management-case-studies/ [Consulté le 19 octobre 2014]

SCHUMAN Rebecca. "Will Digital Humanities #Disrupt the University?" in *Slate*, publié le 16 avril 2014. [En ligne]

http://www.slate.com/articles/technology/future_tense/2014/04/digital_humanities_and_the_future_of_technology_in_higher_ed.html [Consulté le 14 novembre]

SCIENCES PO. Horizon 2020 : le programme communautaire pour le financement de la recherche et de l'innovation. In Mission Europe de la recherche [en ligne], mis à jour le 4 avril 2014. [En ligne] http://www.sciences-po.fr/recherche/fr/mission_europe/H2020.htm [consulté le 18 octobre 2014].

SCIENCES PO. Mission Europe de la recherche [En ligne] http://sciences-po.fr/recherche/fr/mission_europe/index.htm [Consulté le 19 àctobre 2014]

Semer, essaimer. La valorisation des données de la recherche, Dossier Ar(abes)sques, n° 73 Janvier-février-mars 2014. En ligne : www.abes.fr/Arabesques/Arabesques-n-73 [consulté le 18/04/2014]

TESTART-VAILLANT Philippe. *Interview avec Michel Wieviorka : les sciences humaines et sociales à l'ère numérique* in *CNRS Le Journal*, 10/01/2014 [consulté le 18/04/2014]. En ligne : <https://lejournal.cnrs.fr/articles/interview-de-michel-wieviorka-les-sciences-humaines-et-sociales-a-lere-numerique>

THATCAMP. Manifeste des *Digital Humanities* , 26 mars 2011. [En ligne] http://books.openedition.org/oep/235 [consulté le 30 novembre 2014]

THAT CAMP COLLECTIF. *Open data en SHS : Proposé par Cynthia Pedroja, Elifsu Sabuncu, Anne-Laure Stérin*. In *THATCamp Paris 2012 : Non-actes de la non-conférence des humanités numériques*, Paris : Éditions de la Maison des sciences de l'homme, 2012 [En ligne]. http://books.openedition.org/editionsmsh/364 [consulté le 19 avril 2014]

TRABAL Patrick. « Le logiciel Prospéro à l'épreuve d'un corpus de résumés sociologiques », Bulletin de méthodologie sociologique [En ligne], 85 | 2005, mis en ligne le 28 mars 2008, consulté le 18 avril 2014. URL : http://bms.revues.org/993

University of Bristol. « What counts as research data? » In data.bris [En ligne]http://data.bris.ac.uk/research/bootcamp/data/ [Consulté le 19 octobre 2014]

University of Bristol.« Introduction to Research Data Management». In data.bris [En ligne] http://data.bris.ac.uk/research/introduction/ [Consulté le 19 octobre 2014]

URFIST Info. « "Données" de la recherche, les mal-nommées ». In URFIST, Actualités des Sciences de l'Information, billet de blog, publié le 15 novembre 2013 [En ligne] http://urfistinfo.hypotheses.org/2581 [Consulté le 19 octobre 2014]

VAN NOORDEN Richard. "Elsevier opens its papers to text-mining: Researchers welcome easier access for harvesting content, but some spurn tight controls" in *Nature,* 03/02/2014 [En ligne] http://www.nature.com/news/elsevier-opens-its-papers-to-text-mining-1.14659 [consulté le 16/04/2014]

VENTURINI Tommaso. « Médialab de Sciences Po : cartographier le web pour les sciences sociales » in INA : e-dossiers de l'audiovisuel, juin 2012. [En ligne] http://www.ina-expert.com/e-dossier-de-l-audiovisuel-sciences-humaines-et-sociales-et-patrimoine-numerique/medialab-de-sciences-po-cartographier-le-web-pour-les-sciences-sociales.html [consulté le 18 avril 2014].

Annexes

Annexe 1 – Programmes et initiatives de gestion et de valorisation des données en SHS en France et à l'international

INITIATIVES DE LA COMMUNAUTE SCIENTIFIQUE

❖ FRANCE

PANDOR - Portail Archives Numériques et Données de la Recherche
https://pandor.u-bourgogne.fr

Description : Le Portail Archives Numériques et Données de la Recherche (PANDOR) est l'outil de diffusion du Centre de Ressources Numériques Thématique (CRNT) porté par la Maison des Sciences de l'Homme de Dijon (MSH).
Cet outil en ligne est pensé comme une entrée donnant accès à l'ensemble des corpus numérisés et/ou numériques issus des programmes de recherche interdisciplinaires portés par la MSH ainsi qu'à certains catalogues informatisés. Il est le lieu de présentation du programme et de ses objectifs. Il permet de valoriser un ensemble de données numériques multi-supports et d'interroger ensemble les différents corpus constitués initialement sous forme d'entité unique.

Criminocorpus – Plateforme d'édition scientifique pour l'histoire de la justice, des crimes et des peines. Produit d'une coopération originale entre chercheurs, archivistes, documentalistes et collectionneurs, Criminocorpus met à disposition du public des outils de recherches, des sources, des articles et des expositions virtuelles.
< https://criminocorpus.org/>

Sciences Po. « Mission Europe de la recherche ». 2014. Consulté le 19 octobre2014. http://sciences-po.fr/recherche/fr/mission_europe/index.htm.

Description : Créée en 2004, la Mission Europe de la Recherche (MER) vise à renforcer l'internationalisation des recherches conduites à Sciences Po en les inscrivant dans l'Espace Européen de la Recherche. Cette inscription sert quatre objectifs : financer nos recherche, en faire reconnaître la qualité, en faciliter la diffusion et intensifier les coopérations entre chercheurs d'horizons différents.

Be quali

<http://www.bequali.fr/bequali/>

Collection des Corpus Oraux Numériques – TGIR Huma Num

<http://cocoon.huma-num.fr/exist/crdo/>

Kinsources
<http://www.agence-nationale-recherche.fr/projet-anr/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-CORP-0008>

- **INTERNATIONAL**

Organisation internationales

http://www.nsd.uib.no/nsd/english/internasjonalt.html

The Open Data Research network

http://www.opendataresearch.org/

Description : Governments, civil society organisations and companies across the world are actively engaging with open data: publishing and using datasets to promote innovation, development and democratic change.The Open Data Research network has been established to connect researchers from across the world working to explore the implementation and impact of open data initiatives. It is a joint project of IDRC and the Web Foundation, and is seeking to develop wider partnerships over the coming year. The network currently hosts the 'Exploring the Emerging Impacts of Open Data in Development Countries (ODDC)' programme.

Research Information Network – Royaume-Uni
http://www.rin.ac.uk/about

Europe

Knowledge Exchange | http://www.knowledge-exchange.info/Default.aspx?ID=1

L'initiative de l'organisation Knowledge Exchange (KE), dont le siège est au Pays-Bas, est née en 2005 visant un espace de coopération de partenaires internationaux œuvrant pour l'implémentation des infrastructures numériques et des TICs dans le milieu de l'enseignement supérieur et recherche scientifique. KE compte aujourd'hui avec cinq partenaires :

⇒ CSC - IT Center for Science - Finlande
⇒ Denmark's Electronic Research Library (DEFF) - Danemark
⇒ German Research Foundation (DFG) - Allemagne
⇒ Jisc - Royaume-Uni
⇒ SURF – Pays-Bas

Ces partenaires travaillent aujourd'hui sur les axes suivants : *Open Access*, Données de la recherche, Outils et TICs pour la recherche e Normes et standards d'interopérabilité.

Le but de cette coopération est de contribuer à la construction d'un environnement numérique européen d'information scientifique, ouvert, durable et pouvant s'étendre hors des limites européens.

Le dernier rapport de KE de 2013 *The Value of Research Data - Metrics for datasets from a cultural and technical point of view*[86] se penche sur la nécessité d'accompagner les initiatives institutionnelles et gouvernementales d'ouverture des données d'un développement des systèmes de *data metrics* (datamétrie) : création des systèmes fiables de évaluation de la qualité de données et de leurs usages.

[86] Consultable en ligne : http://www.knowledge-exchange.info/datametrics

Annexe 2 - Ressources en ligne pour la mise en place de plans de gestion des données. L'exemple du Royaume-Uni

I. Comment prendre en charge la gestion des données de la recherche ?

Actuellement, il existe une diversité de ressources en ligne disponibles pour assister les documentalistes qui accompagnent les chercheurs dans la mise en place des plans de gestion des données. Nous mentionnons ici notamment deux sources :

Le module de sensibilisation aux enjeux et bonnes pratiques de l'INIST- CNRS
http://www.inist.fr/donnees/co/Donnees_recherche_web.html basé lui-même sur des ressources diverses dont :
LERU Road map for research data
http://www.leru.org/files/publications/AP14_LERU_Roadmap_for_Research_data_final.pdf

Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics.
Le guide Data Managing and Sharing qui présente et explique toutes les questions à se poser en amont, pendant et en aval de la mise en place d'un plan de gestion des données. Cet ouvrage offre un nombre important de références accessibles en ligne.

II. Importance du concept « cycle de vie » dans la gestion des données

Dans *Managing and Sharing Research Data* [87], les auteurs soulignent l'importance grandissante du concept de « cycle de vie des données » à côté et en tant que prolongement du cycle de vie de la recherche traditionnelle[88].

[87] Op.cit.
[88] Pour voir des exemples de modèles de *Research Knowledge Creation*, consulter Humphrey, C. (2006) *e-Science and the Life Cycle of Research*, University of Alberta. [En ligne] http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

En effet, à l'ère du numérique, il importe de distinguer le projet de la recherche des données qui ont été créés par lui. Entre les années 1990 et 2000, disent les auteurs, « le cycle de vie des données de la recherche a été promu comme concept fondamental venant en appui des pratiques et de préservation et curation des données»[89].

Ci-dessous, quelques exemples de cycle de données et des ressources pour aller plus loin :

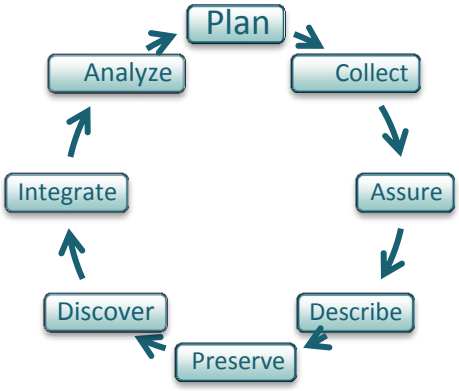


Figure 10 - The Data Lyfecycle by DataOne

< https://www.dataone.org/education-modules>

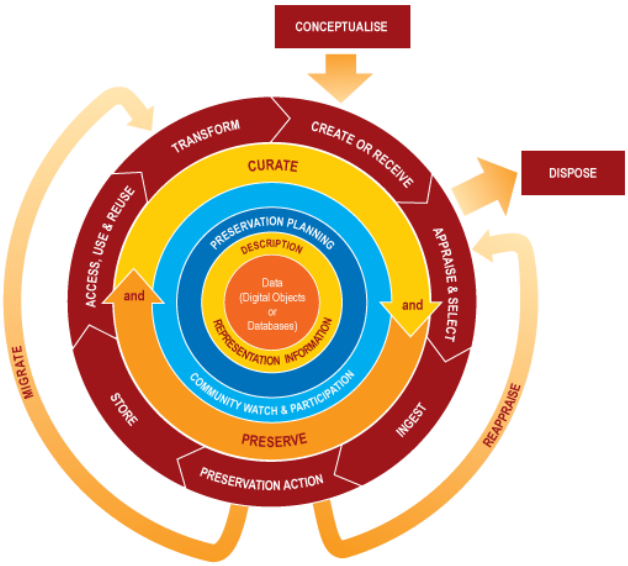


Figure 11- Curation Lifecycle model – DCC

<http://www.dcc.ac.uk/resources/curation-lifecycle-model>

[89] *Managing and Sharing Research Data,* op.cit. p. 17.

Description : Ce modèle permet une vision précise de toutes les étapes nécessaires à la gestion des données, depuis leur création jusqu'à leur cycle de vie plus complexe et interactif.

« It is important to note that the model is an ideal. In reality, users of the model may enter at any stage of the lifecycle depending on their current area of need. For instance, a digital repository manager may engage with the model for this first time when considering curation from the point of ingest. The repository manager may then work backwards to refine the support they offer during the conceptualisation and creation processes to improve data management and longer-term curation. »
<http://www.dcc.ac.uk/resources/curation-lifecycle-model#sthash.PYFgiRGM.dpuf>

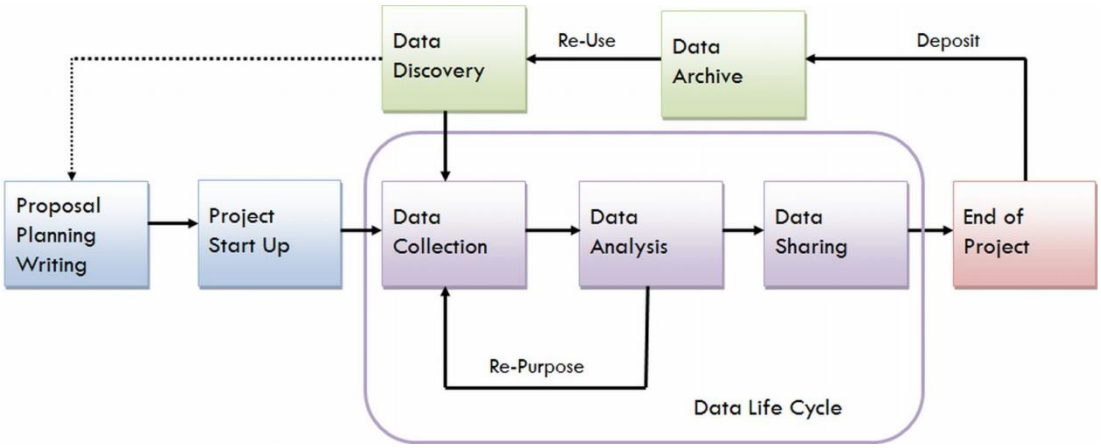


Figure 12 - Steps in Data Lifecycle - University of Virginia library - Research Lifecycle

http://data.library.virginia.edu/data-management/lifecycle/

Autres ressources :

California Digital Library, What is the research life cycle ?
<**http://fr.slideshare.net/joanstarr/the-research-data-life-cycle**>
Présentation de plusieurs modèles et finalités de cycles de vie

JISC – Research Lifecycle diagram
<http://webarchive.nationalarchives.gov.uk/20140702233839/http://www.jisc.ac.uk/whatwedo/campaigns/res3/jischelp.aspx>
Cycle de vie des données et autres sources

Data management during Lifecycle
http://guides.library.oregonstate.edu/lifecycle

Annexe 3 - Tableaux et grilles d'analyse

TABLEAU 1 - DEFINITION DE DONNEES DE LA RECHERCHE PAR DISCIPLINE

Le tableau ci-dessous rassemble les réponses des chercheurs concernant le type de données produite ou utilisée dans la recherche. Entre parenthèses ( ) : numéro de l'entretien

Discipline (*)/ **Domaine de recherche**	Réponse à la question : ***« Dans le cadre de votre pratique, qu'est-ce qu'une « donnée de la recherche » ?***
Anthropologie-Ethnologie **(*4*)*** **Langue guarani et paléographie**	manuscrits guaranis et ses transcriptions (paléographie)
Anthropologie sociale et de l'environnement **(*9*)** **Eco-anthropologie et ethnobiologie**	données ethno géographiques (enregistrement de témoignages) et carnets d'observation de terrain
Géoarchéologie **(*10*)** **Paléoenvironnements**	mesures physico-chimiques sur les archives sédimentaires et interprétation de ces données par comparaison ou à la lumière d'autres résultats
Archéologie (*12*) **La vie matérielle médiévale en Europe Occidentale**	base de données d'images, carnets de fouille, plans, dessins
Arts (*21*) **Esthétique du cinéma et de l'art des nouveaux média**	documentation bibliographique et archives d'images animées
Arts (*39*) **Esthétique du cinéma et de l'audiovisuel** **La figure du témoin : formes, usages et enjeux**	notes, observations, données statistiques du cinéma, analyse de films et documentajires, forum internet
Arts (*44*)	données ds et corpus de photographies

Histoire et esthétique de la photographie	
Démographie (*6*) **Histoire socio-politique des élites et des pouvoirs au XVIe et XVIIe siècle**	archives de contrats de mariage parisiens
Démographie (*28*) **Sida, migrants et santé en France**	entretiens qualitatifs et analyse d'enquêtes quantitatives
Droit (*37*) **Droit des médias**	données classiques du droit, base de données juridiques
Droit (*45*) **Droit du travail, relations employeur/salariés dans l'emploi public**	données de bases de données juridiques, données d'entretiens qualitatifs
Droit (*53*) **Normes, droit, monde arabe, Égypte, constitution, famille, justice, démocratisation, charia**	réseaux sociaux (Twitter), vidéos en ligne (You Tube), décisions des Tribunaux, émissions de télévision, actualités en ligne
Economie (*47*) **Analyse économique des élections, participation électorale, modes de scrutin, opinion à l'égard des migrants**	Données quantitatives (sondages)/questionnaires sur table
Economie (*32*) **Economie de la santé : addictions, vaccination, offre de soins**	informations quantitatives présentées en tableur, enquêtes
Economie (*35*) **Economie mathématique, protocole Bitcoin d'un point de vue micro-économique**	articles, code, programmes informatiques
Economie (*42*)	statistiques publiques, entretien, enquêtes,

Économie du développement, emploi informel, genre, emploi du temps.	méthodologie
Economie (*47*) **Analyse économique des élections, participation électorale, modes de scrutin, opinion à l'égard des migrants**	Données du Ministère, INSEE, Quételet, Instituts statistiques, micro-données des bureaux de vote
Géographie (*23*) **Migrations internationales, mobilités et changement urbain au Sénégal**	données enquêtes quantitatives, données INSEE
Histoire (*11*) **Histoire des sciences de l'homme et histoire de la justice**	sources, données brutes enrichies de métadonnées, données secondaires produites par d'autres chercheurs
Histoire (*24*) **Histoire du travail, des conflits sociaux et des techniques**	prise de notes sur d'ouvrages, consultation/dépouillement d'archives, recueil et constitution de données statistiques
Histoire (*49*) **L'Amérique latine et la Première Guerre mondiale**	données d'archives pour constitution des corpus
Histoire (29) **Histoire du droit du travail**	prises de note et 8000 photos des B.O. du Travail dans les colonies
Histoire (*1*) **Histoire orale**	Archives orales numérisées
Histoire (*2*) **Histoire de l'action au XXe siècle, histoire des organisations et des gouvernements.**	Références bibliographiques, archives annotées
Histoire (*36*) **Les discours sur l'égalité/inégalité des femmes et des hommes**	manuscrits du XVII, notes, paléographie, codes
Histoire de l'Art (*5*)	*Muqarnas* (élément d'architecture de l'art islamique)

Histoire de l'art almoravide, systèmes d'information géographiques (SIGs) appliqués aux humanités	
Linguistique (*43*) **Bilinguisme, langues en contact, langues et migration**	enregistrement des observations participantes, entretiens enregistrés, transcription des entretiens, analyse des interactions
Linguistique (*50*) **Les corrélats acoustiques et articulatoires de la parole expressive (expressions d'émotions et d'attitudes)**	données acoustiques, données physiologiques, données linguistiques, méthodologie
Littérature (*18*) **Théorie littéraire appliquée à la littérature française contemporaine**	bibliographies (base de données et corpus de références), analyse des discours littéraires, sources littéraires
Philosophie (*7*) **Philosophie allemande contemporaine**	corpus de textes philosophiques
Philosophie (*13*) **La condition animale, approches phénoménologiques de l'animalité, définitions occidentales de l'animalité**	corpus de textes philosophiques, articles de revue, ouvrages
Philosophie (*16*) **Philosophie allemande post-hégélienne, philosophie politique, philosophie contemporaine**	base de données, Big data, corpus de textes philosophiques
Psychologie (*52*) **Psychologie de la perception et psychologie cognitive expérimentale**	données de "temps de regard" de l'enfant
Sciences Administratives (*30*) **Luxe, rapport aux objets**	entretiens, observations participante et non participante

Sciences Administratives *(41)* **Plans sociaux des entreprises, syndicalisme, négociation en entreprise**	bibliographies et statistiques publiques
Sciences de l'Education *(40*)	notes d'observation, entretiens qualitatifs, enquêtes larges, méthodologie
Sciences de l'Education *(46)*	données d'enquêtes
Sciences de l'Education *(48)*	carnets de bord des étudiants, notes d'observation, entretiens enregistrés, transcriptions
Sciences de l'information et communication *(15)*	Ras
Sciences de l'information et communication *(26)*	notes de séminaires, webmaires, lectures
Sciences de l'information et communication *(51)*	données d'entretiens, données d'observation, traces médiatiques (données du web)
Sociologie *(3)*	archives/sources
Sociologie *(14)*	données quantitatives (forums de presse en ligne)
Sociologie *(17)*	entretiens, notes d'analyse des affaires judiciaires, observations non participante, méthodologie
Sociologie *(22)*	données d'archives/données nativement numériques
Sociologie *(34)*	enquêtes qualitatives/rapports de synthèse ; méthodologie
Sociologie (*33*)	questionnaires, entretiens, compte-rendu d'observations
Sociologie *(38)*	INSEE, INED, Eurostat/analyse de ces données (calcul de tableaux)
Sciences politiques *(8)*	analyse de données d'enquêtes
Sciences politiques *(19)*	données numériques /cartographies
Sciences politiques *(20)*	observations de terrain, entretiens qualitatifs et documentation primaire (archives, documents divers)
Sciences politiques *(25)*	données statistiques, données quantitatives

GRILLE D'ANALYSE THEMATIQUE

Retours/Points de vue	Partage/ est d'accord	Ne partage/n'est pas d'accord	discipline
Problématiques principales – Données SHS			
Les questions éthiques (confidentialité et anonymat) et juridiques sont la problématique principale à une diffusion des données de la recherche.	1, 2, 3 5,9 Archives orales, données de terrain 14 (analyse au cas par cas difficile), 15 (le cadre juridique est obsolète face au numérique), 17, 20,22, 25,33,41,43,48	6	Histoire orale, histoire, anthropologie, sociologie, eco-anthropologie, philosophie des sciences, sciences admin, linguistique
Les **données qualitatives** posent de problèmes particuliers à leur diffusion, notamment éthiques/liées à l'anonymisation	20,33		
Un problème essentiel à l'ouverture des données est le travail nécessaire en amont (traitement, normalisation, documentation, contextualisation, limites, encodage, enrichissement de métadonnées) pour qu'elles puissent **être réutilisables**	5, 9,14, 19 (codes), 22, 26, 27,28, 33, 40, 43,46		Histoire, eco-anthropologie, sociologie des TICs, démographie, éducation, linguistique
Avec les données, il faut aussi justifier de la méthodologie et des algorithmes d'exploration des données/partager aussi les programmes	14, 25		Sciences politiques, Sociologie des TIcs

Le chercheur doit s'entourer des personnes ayant une culture des données (statisticiens) ou des compétences en gestion de l'information	14, 28		Sociologie des Tics, Histoire
L'accès aux sources/ressources est l'enjeu principal	2, 3, 4, 7, (numériser des revues anciennes) 21, 37		Histoire, sociologie, anthropologie, philosophie, cinéma, droit
Problème de **sous-exploitation des données** produites et gérées par des instituts ou réseaux institutionnels/ grandes enquêtes. En particulier, l'accès aux micro-données (données avant traitement statistique,	8, 23, 41,42		Sciences politiques, géographie, sciences administratives, Economie
Un problème de premier ordre est : quelles formes de diffusion adopter et pour quels buts, quel public ?	14		Sociologie des Tics
Problématiques d'archivistique : Pérennité, stockage, structuration et mise à disposition des données	22,26		Sociologie des tics
Comment valoriser les données produites par les chercheurs avant le numérique ?	22		Sociologie des Tics
Lourdeur administratif de Quetelet / INSEE : il faudrait faciliter l'accès aux données	23, 31		Géographie, sciences admin
Davantage d'infrastructures et des politiques d'archivage sont nécessaires	26		SIC

Publier les données est-il vraiment utile face à l'abondance de publications déjà existantes et au manque de temps de chercheurs de tout lire et trier ?	24		histoire
TICs/TDM en SHS			
Il faut être ouvert mais rester critique envers le TDM et nouvelles technologies	2, 4,6,32		Histoire, anthropologie, économie
Le numérique est une révolution/produit des transformations des pratiques en SHS	3, 4, 12,16 (accès aux sources), 27		Sociologie, anthropologie, archéologie, philosophie
Face aux nouvelles technologies, la démarche théorique et intellectuelle ne doit pas passer en second plan mais reste primordiale/les SHS restent des discipline du sens et de l'interprétation	4, 6, 8, 9, 11, 34		Anthropologie, Démographie, Sciences politiques, Histoire
Pratiques de la recherche (culturelles, techniques, politiques)			
Le chercheur doit conserver sa liberté de décision d'ouverture ou non de ses données/ et d'effacer ses traces	2,3, 13,26		Histoire, sociologie, philosophie
La réutilisation des données n'a du sens que dans les disciplines travaillant avec des données quantitatives	3		Sociologie
La réutilisation des données produites par d'autres chercheurs se fait de manière courante	10 (dans le cadre des collaborations),11 (données du	12, 3, 36 (le chercheur doit produire ses données),49	Archéologie, Histoire, Economie, droit, Histoire

	type « outils »), 27, 32,33	(les données sont déjà des interprétations)	
Il existe une culture du partage / partage naturel entre chercheurs	4, 6,27,32, 34 (de la méthodologie)		Anthropologie, Démographie, histoire médiévale, sociologie
Il existe une culture de la propriété chez les chercheurs	7 (crainte plagiat)		philosophie
Les chercheurs sont aujourd'hui partagés entre une culture de la propriété et du partage	8 , 17, 34		Sciences politiques, sociologie
Les chercheurs en SHS ne sont pas à l'aise dans le partage de leurs méthodologies de recherche	22,36,40(à propos des historiens)	34 (cela constitue une obligation)	Sociologie, histoire
Une recherche de qualité se construit sur ses propres données/le chercheur doit produire/ traiter ses données **La production des données est inhérente au processus créatif/intellectuel**	3, 26, 36		Sociologie, Sic, Histoire
Ouvrir/Diffuser/ Partager des données : quand, quoi et pourquoi ?			
Partager les données est essentiel à la constitution de réseaux scientifiques/faire circuler les savoirs/construire la science à plusieurs/ouvre des nouveaux horizons	4,5, 12,26, 27		Anth , Hist ; archéologie médievale ; SIC
L'ouverture des données ne doit pas être une fin en soi mais avoir des objectifs et intérêts précis et être analysées au cas par cas selon un certain nombre de problématiques particulières	1, 26, 20		Histoire, Sciences politiques, Sic

Les initiatives d'ouverture des données en SHS vont de pair avec le développement de l'interdisciplinarité/ Humanités numériques	5, 19 (Venturini) ,29 (ambauche de deux ingénieurs)		Histoire de l'Art, disciplines textuelles, histoire du droit
Donner une plus grande visibilité aux données dites « mineures » est important (littérature grise, données « erreurs »)	10,31,33		Archéologie, économie
Les données du type « outils de la recherche », élaborées au cours de recherche, offrent un plus grand intérêt à être diffusées	11		histoire
Les donnes quantitatives « brutes » ont une plus grande vocation à être ouvertes/partagées à des fins de vérification, réutilisation, replicabilité/reproductibilité de la recherche	6 (réutilisation, vérification), 11, 38,42		Démographie, histoire,sociologie,economie
Mutualiser les travaux techniques de préparation des données aurait un grand intérêt	9		Eco-anthropologie
Partager des données de terrain ou autres avec l'équipe	27, 20,43,52	9 (partage rare),	Oui :Histoire médiévale, sciences politqiues, linguistique Non : eco-anthropologie
Pas de partage collectif, mais peuvent partager si sollicités	3,20, 34, 35, 48		
Partage au sein de l'équipe	32,41,52		Sciences admin
Partager uniquement les résultats	13		
Valoriser et partager la méthodologie de collecte et traitement/analyse des données est essentiel et aussi	17, 28		Sociologie, démographie

important que diffuser les données / Il n'y a pas d'intérêt à diffuser des « données brutes » en absence d'une contextualisation ou d'une diffusion de la méthodologie			
Diffusion des étapes dans des séminaires et communications scientifiques	34, 39,43,48		Sociologie, cinéma, linguistique
Dans le processus de recherche, il n'y a pas d'étapes intermédiaires à diffuser.	13		Philosophie
Seulement les résultats doivent être/ sont publiés	24, 30, 37		Histoire
Ne souhaite pas partager	30,51		Sciences administratives, Sic
Diffuser/publier : comment ?			
Diffusion rapide via des réseaux sociaux / blogs/ sites	4, 26, 44,47,53		Anthropologie, Sic, Esthétique de la photographie, droit
Création de plateformes de dépôt par et pour les chercheurs	31, 40		économie
Publication des résultats + accès aux sources/ données parallèle	11, 25, 38		histoire
Publication des résultats + accès aux sources/ données en aval	52,53		
Publier les résultats en revue **spécialisée** continue d'être la meilleure manière de diffuser les travaux scientifiques	16,37,39, 50,		Philosophie, Droit, cinéma
Publier les données et la méthodologie	17, 36		Sociologie
Offrir à la publication des données un espace propre : *data publication*	46		Education

Revues « augmentées » à l'exemple des revues en Economie	22	
A la Robert Darnton ?	11, 22	Sociologie, histoire
Ouvrir les données avant la publication	5 (est important pour faire connaître son travail)	32 (peu probable, à cause de la concurrence)
En amont, en aval, en parallèle selon le type de recherche et les objectifs	27, 32	Histoire médiévale, économie
Une publication en aval est plus probable, les chercheurs sont compétitifs	50	linguistique
Motivations des chercheurs à préparer, ouvrir, partager les données		
➢ Les chercheurs sont plus motivés si reconnaissance institutionnelle et cadre de travail collectif (17, 32, 23, 29,39 ➢ Les chercheurs doivent pouvoir avoir un accès facile aux données qu'ils ont produit. La solution de dépôt des données actuelle, Quetelet, n'est pas satisfaisante, car l'accès aux données se traduit très souvent par des démarches administratives lourdes (23 ➢ Les chercheurs doivent se sentir concernés par la mutualisation et l'échange des données, ce qui permet d'avancer plus vite et de créer un esprit de travail collectif parfois rare chez les chercheurs en SHS (23, 27,40,46,47 ➢ Créer des outils adaptés pour la recherche (23,40,46 ➢ Ouvrir les données est aussi important que citer ses sources : exigence épistémologique (46, ➢ La diffusion des données aura des effets positifs de transparence et reconnaissance des « coulisses » de la recherche (42 ➢ De plus en plus de revues, à l'instar des revues en Economie, vont exiger les données (40,42,47 ➢ Contribution à la réputation du chercheur et du laboratoire qui a produit les données (52		
Freins au partage/ouverture		
➢ Temps et travail nécessaire pour le traitement des données (22,23,28,40,43		
➢ problèmes juridiques à traiter en amont (22		
➢ problèmes de déontologie (22		
➢ culture de la propriété des données (22		

➢ préparer les données et justifier des méthodologies peut être perçu comme des contraintes par les chercheurs (23
➢ habitude au travail de recherche individualiste (23, 34
➢ l'aspect technologique tient à distance quelques chercheurs
➢ initiatives individuelles difficiles faute de financement et motivations (32
➢ problèmes de confidentialité : une anonymisation excessive rend les données inutilisables (41,
➢ la concurrence entre les chercheurs et les institutions peut freiner des initiatives d'ouverture (42

EXEMPLE DE GRILLE D'ANALYSE UTILISEE

discipline	N° entretien	données	type de donnée	type et périmètre de diffusion ou partage appliqués	Type de traitement appliqué	type et périmètre de diffusion souhaités ou préconisés	motivation principale	problème/frein principal pour la diffusion	réutilisation des données produites par d'autres chercheurs

Annexe 4 – Les entretiens

I. Questionnaire d'orientation des entretiens

1) Dans le cadre de votre propre recherche, comment définiriez-vous la notion de « données de la recherche » ? Quel(s) type(s) de « données » produisez-vous ? Que faites-vous de ces « données » produites ? (traitement, stockage, diffusion, etc) Rencontrez-vous de problématiques particulières liées à la nature de votre recherche ou aux infrastructures de votre institution pour le traitement de ces données ?

2) Quelles pourraient être selon vous les motivations pouvant amener des chercheurs à diffuser, partager réutiliser des « données » ? Seriez-vous vous-même disposé(e) à le faire ou le faites-vous déjà ? Avez-vous déjà (ré)utilisé des données mises à disposition par d'autres chercheurs ?

3) Par rapport aux publications des chercheurs, quel type de diffusion pour ces données vous semblerait-il potentiellement intéressant : en amont de la publication d'articles ou d'ouvrages, parallèlement à celle-ci, ou bien en aval ?

4) Le « data mining » et la « Fouille de textes » sont-ils des procédés que vous utilisez aujourd'hui dans vos travaux ? Sinon, pensez-vous que vous pourriez être amené(e) à utiliser ce type de procédés / outils à l'avenir ?

5) De votre point de vue, les technologies du numérique transforment-elles les pratiques en SHS ? Ont-elles transformé votre propre pratique ? Si oui comment ?

II. Index par discipline

Discipline	N° entretien
Anthropologie	**4, 9**
Archéologie	**10, 12**
Arts	**21, 39, 44**
Démographie	**6, 28**
Droit	**37, 45, 53**
Economie	**31*, 32, 35, 42, 47**
Géographie	**23**
Histoire	**1, 2, 5, 11, 24, 27, 29, 36, 49**
Linguistique	**43, 50**

Littérature	**18**
Philosophie	**7, 13, 16**
Psychologie	**52**
Sciences Administratives	**30, 41**
Sciences de l'éducation	**40, 46, 48**
Sciences de l'Information et Communication	**15, 26, 21**
Sociologie	**3, 14, 17, 22, 33, 34, 38**
Sciences Politiques	**8, 9, 20, 25, 31***

III. Liste des institutions d'appartenance des chercheurs interviewés

Analyse et Traitement informatique de la Langue Française (ATILF) – Université de Nancy, Nancy

Centre Alexandre Koyré (CAK-EHESS), Paris

Centre d'études et de recherches internationales (CERI), Sciences Po, Paris

Centre d'études européennes (CEE), Sciences Po, Paris

Centre de données Socio-politiques, Sciences Po, Paris

Centre de recherche en économie et statistique (GENES-CREST), Malakoff

Centre de Recherches Historiques (CRH), EHESS, Paris

Centre Population et Développement (CEPED) - Université Paris Descartes, Paris

CERLIS (Centre de recherche sur les liens sociaux) - Université Paris Descartes, Paris

Cités, Territoires, Environnement et Sociétés (CITERES)- Université François-Rabelais, Tours

Ecole des Hautes Etudes en Sciences Sociales (EHESS), Paris

Ecole Normale Supérieure (ENS), Cachan

Ecole Normale Supérieure (ENS), Paris

Ecole Pratique des Hautes Etudes (EPHE), Paris

Groupe d'Analyse et de Théorie économique (GATE), Lyon-Saint-Etienne

Groupe d'archéologie médiéval (GAM-CRH EHESS), Paris

Groupe de Recherche sur les Enjeux de la Communication (GRESEC) – Université de Grenoble 3, Grenoble

Institut d'administration des entreprises (IAE) - Paris Sorbonne

Institut de recherche et développement (IRD)- Paris Descartes, Paris

Institut de recherche sur le cinéma et l'audiovisuel (IRCAV), Université Sorbonne Nouvelle, Paris

Institut des Hautes Etudes de l'Amérique Latine (IHEAL)-Université Sorbonne Nouvelle, Paris

Institut des Sciences Sociales du Travail - ISST Paris Sorbonne, Bourg-la-Reine
Institut Francilien Recherche Innovation Société (IFRIS), Université Paris-Est Marne-la-Vallée
Institut National de la Recherches Agronomique (INRA), Paris
Institut National des Langues et Civilisations (INALCO)- Paris
Laboratoire de démographie et histoire sociale (LaDéHis), EHESS, Paris
Laboratoire de Phonétique et Phonologie (LPP), Université Sorbonne Nouvelle, Paris
Laboratoire de Psychologie de la Perception- Université Paris Descartes, Paris
Laboratoire de recherche en gestion et économie (LaRGE), Strasbourg
Laboratoire des techniques, territoires et société (LATTS), Université Paris-Est Marne-la-Vallée
Laboratoire des usages de France Telecom R&D
Laboratoire Droit et changement social (DCS), Nantes
Laboratoire Interdisciplinaire – Solidarité Sociétés Territoires (LISST-CERS – CNRS)
Laboratoire Interdisciplinaire des Droits des Médias et des Mutations Sociales (LID2MS), Aix-Marseille
Lille Economie et Management (LEM), Lille
Maison René-Ginouvès - Université de Paris 10, Nanterre
Médiation Information Communication Art (MICA) – Université de Bordeaux, Bordeaux
Migrations internationales, Espaces et Sociétés (MIGRINTER), Université de Poitiers, Poitiers
Musée National d'Histoire Naturelle, Paris
Observatoire sociologique du changement – Sciences Po, Paris
Université Paris Ouest- Nanterre - La défense
Universidad de Granada, Espagne
Université de Bourgogne
Université de Nice Sophia Antipolis
Université de Paris Sorbonne-Paris IV, Paris
Université de Strasbourg
Université Paris Descartes - Paris
Université Paris IV-Sorbonne

IV. Sommaire des entretiens

ENTRETIEN N°1 (09/06/2014) – HISTOIRE

Modalité d'entretien

Téléphone

Statut

Maître de conférences

Domaines de recherche

Archives orales, histoire orale, patrimoine immatériel, histoire des organisations.

Retours / Positions

FAIRE DU CAS PAR CAS

Cette chercheuse développe actuellement un projet de numérisation d'archives orales, dont les résultats ne sont pas encore publics. Elle et le reste de cette équipe se sentent fortement concernés par la question des données de la recherche et de leur ouverture, même si cette question est loin d'être simple à appréhender.

Constituer et diffuser des fonds comme celui que cette équipe a pu constituer pose des questions d'ordre éthique, mais aussi d'ordre juridique, notamment vis-à-vis des règles fixées par la CNIL. S'ajoute à cela le problème financier, la numérisation d'archives orales étant une activité en général assez couteuse et les financements publics étant souvent difficiles à obtenir pour ce type d'initiative.

De manière plus générale, le partage de données entre les chercheurs pose souvent des problèmes du même ordre, mais bute aussi sur l'absence d'une culture de mutualisation des résultats bruts produits dans le cadre des recherches, autrement que sous la forme de publications et de conférences.

S'il est évident que la numérisation de fonds, de corpus et de sources est en voie de développement au sein du monde la recherche, ces travaux doivent être conduits avec prudence et une approche globale et systématique n'aurait pas de sens sur cette question. Les problèmes posés doivent être vus au cas par cas.

PAS DE FIEVRE TECHNOPHILE EN HISTOIRE

En France, la fièvre technologique n'a pas à proprement parler envahi des disciplines comme l'histoire, seuls quelques jeunes chercheurs s'étant fortement investis dans des projets qui s'inscrivent dans la veine des Humanités Numériques. Il s'agit peut-être d'une question de génération, mais l'hypothèse d'une montée en puissance de la technologie dans des disciplines comme l'histoire lui semble peu probable : les recherches en histoire ne peuvent se faire sans des sources solides et des références étayées.

ENTRETIEN N° 2 (10/06/2014) – HISTOIRE

Modalité d'entretien

Face-à-face

Statut

Directeur de recherche

Domaines de recherche

Histoire du XXe siècle, histoire des organisations et des gouvernements

Retours / Positions

L'ENJEU DES DONNEES TIENT A L'ACCESSIBILITE DES SOURCES POUR LES CHERCHEURS

A son sens, la question des données de la recherche doit avant tout se comprendre comme celle des sources accessibles aux chercheurs. L'enjeu à ce niveau est très important : le nombre, la qualité et la facilité d'accès aux sources déterminent grandement la qualité du travail des chercheurs dans un domaine comme l'histoire des organisations.

En ce sens, le fait qu'elles soient de plus en plus massivement rendues publiques, de façon numérique ou physique (pour les historiens, le contact matériel avec les sources est encore très important), est une bonne chose. Les chercheurs œuvrent d'ailleurs en ce sens quand ils le peuvent : il a, par exemple, fait don récemment d'un fonds d'entretiens à une bibliothèque spécialisée. Mais cela ne va pas sans poser un certain nombre de problèmes éthiques, juridiques et politiques, notamment dans un cas comme le sien où il s'agit d'archives de l'ex-URSS.

LAISSER AU CHERCHEUR LA LIBERTE DE DIFFUSER OU NON SES DONNEES

Le travail de recherche produit certes des éléments que l'on pourrait globalement qualifier de « données » (tableaux Excel, synthèse d'entretiens, notes intermédiaires, etc.), mais dans un domaine de recherche où l'approche n'est pas quantitative, comme l'histoire,

l'utilisation de cette matière ne fait pas grand sens. Cette question peut certainement avoir du sens dans les sciences dures, ou plus généralement dans les disciplines utilisant des méthodes à proprement parler quantitatives, mais pour son domaine, l'enjeu n'est pas là.

Le chercheur doit pouvoir garder sa méthode de travail pour soi et ne pas forcément dévoiler les chemins parcourus pour aboutir à une publication ou une communication. Généraliser la diffusion de versions intermédiaires d'un travail de recherche pourrait même représenter un danger, si cela conduisait à restreindre la liberté du chercheur à faire évoluer son approche, son raisonnement, voire ses conclusions. Le chercheur doit avoir la liberté de choisir s'il souhaite ou non partager ses données, toute contrainte à ce niveau, dans la veine de la société de « surveillance » qui apparaît peu à peu, serait tout à fait regrettable.

RESTER OUVERT MAIS AUSSI CRITIQUE VIS-A-VIS DES NOUVELLES METHODES DE PRODUCTION DES SAVOIRS

Pour les pratiques de fouille automatisée de corpus de textes ou de données, il est nécessaire de réfléchir plus en profondeur aux conditions dans lesquels sont réalisés ces projets. Les chercheurs connaissent encore mal les outils et leurs applications potentielles mais sont a priori ouverts à toute méthode étayée qui produirait des résultats intéressants.

ENTRETIEN N°3 (13/06/2014) – SOCIOLOGIE

Modalité d'entretien

Téléphone

Statut

Directeur de recherche

Domaines de recherche

Sociologie du conflit

Retours / Positions

LES PRATIQUES DE LA RECHERCHE EN SHS SE TRANSFORMENT A L'HEURE DU NUMERIQUE

Ayant des responsabilités académiques et administratives au sein de la communauté scientifique, ce chercheur en sociologie estime être témoin d'un changement en cours dans les pratiques de recherche de toutes disciplines y compris en Histoire où, à son sens, les recherches s'effectuent généralement d'une manière plus traditionnelle. Les changements des pratiques ne se bornent pas à la génération de jeunes de chercheurs, mais commencent peu à peu à gagner du terrain parmi les chercheurs des générations antérieures : il est désormais possible d'avoir un accès simple et rapide à une multiplicité de sources et d'archives, notamment au format numérique, et cela modifie le travail de très nombreux chercheurs, quel que soit leur champs de recherche spécifique.

LE CHERCHEUR A BESOIN DE PRODUIRE SES PROPRES DONNEES

Si l'intérêt d'un accès facilité à un nombre croissant de sources pour les chercheurs parait bien réel, la réutilisation des données produites par les autres chercheurs est une question plus compliquée. Des archives rassemblées par lui, fruits de son travail de recherche portant sur le terrorisme, ont par exemple été mises en ligne il y a maintenant une dizaine d'années. Durant cette période, seules deux personnes appartenant déjà à son cercle de connaissances lui ont fait des demandes pour la réutilisation des données mises à disposition. Dans le domaine de la sociologie, un chercheur a besoin de produire ses propres données dans le cadre de ses travaux, c'est un processus inhérent à son activité de recherche. La réutilisation des données produites par d'autres est une problématique qui concerne plutôt des disciplines qui travaillent avec des informations quantifiables.

L'AVENEMENT PROGRESSIF D'UNE CULTURE ORIGINALE

L'enthousiasme de certains chercheurs en SHS pour les nouvelles technologies n'est pourtant pas limité à un groupe restreint d'acteurs militant pour les « Humanités Numériques ». La création de grandes infrastructures de la recherche en SHS répond à une demande large de la communauté scientifique dans ces domaines. De même, le débat actuel sur l'ouverture et le partage des données, dans le sillon de celui sur le libre accès aux

publications scientifiques, confirme le développement progressif d'une culture originale sur ces questions en SHS, guidée notamment par le souhait d'atteindre de plus larges publics et de démocratiser le processus de recherche. Cette culture a toutes les chances de s'installer très largement, à condition que la liberté d'action du chercheur soit respectée et que les problématiques juridiques soient prises à bras le corps.

ENTRETIEN N° 4 (16/06/2014) ANTHROPOLOGIE/SOCIOLOGIE

Modalité d'entretien

Téléphone

Statut

Chercheure/ Maître de conférences

Domaines de recherche

Langue guarani, paléographie

Retours / Positions

UNE REVOLUTION DES PRATIQUES A TRAVERS LA NUMERISATION DES CORPUS

L'avènement du numérique dans la recherche est une véritable révolution. Des manuscrits introuvables depuis plus de quarante ans sont désormais numérisés et signalés dans des bases de données du monde entier, en étant parfois accessibles directement. Se déplacer pour consulter des sources n'est plus nécessaire et, lorsqu'un manuscrit n'est pas encore numérisé, il est souvent possible d'en faire la demande dans les bibliothèques et archives de l'Amérique Latine.

UNE CULTURE DU PARTAGE EN ANTHROPOLOGIE

En anthropologie, le partage des données collectées est très courant et fait quasi systématiquement partie du processus de recherche des chercheurs. L'échange de données est indispensable à la qualité des travaux, permet de créer des réseaux internationaux et de

construire un savoir commun que chacun peut utiliser / réutiliser à sa guise. Ce type de pratiques ne se retrouve cependant pas forcément dans toutes les disciplines.

Dans son cas, l'ouverture des données produites au cours des transcriptions de manuscrits guaranis est mise en ligne au fil de l'eau sur une plateforme. Une grande partie de son travail est ainsi accessible avant toute publication dans des revues ou sous d'autres formes.

NE PAS OUBLIER L'IMPORTANCE DE LA DEMARCHE THEORIQUE ET INTELLECTUELLE DU CHERCHEUR

Les nouvelles technologies apportent indéniablement de nouvelles opportunités pour les SHS, à commencer par la numérisation des corpus et la possibilité d'augmenter les liens entre eux, de tisser des relations entre les contenus.

Il est cependant important de ne pas oublier que le rôle du chercheur est d'élaborer des hypothèses de recherche, qu'il doit valider ou invalider au regard des faits. L'enthousiasme pour les outils qui permettent de réaliser ce travail ne doit donc pas empêcher de s'interroger sur la démarche et les but poursuivis. Créer des bases de données et les relier entre elles, par les technologies du web sémantique notamment, c'est important mais pour autant que l'objet de la recherche le justifie. Le développement des outils de text mining est aussi intéressant pour l'étude du guarani, à sa connaissance, n'existent pas encore et envisage de travailler dans l'avenir sur ce projet.

ENTRETIEN N° 5 (16/06/2014) – HISTOIRE DE L'ART

Modalité d'entretien

Face-à-face

Statut

Chercheur post-doctorant

Domaine de recherche

Histoire de l'art almoravide, systèmes d'information géographiques (SIGs) appliqués aux humanités

Retours / Positions

LES SIG AU SERVICE D'UNE MEILLEURE EXPLOITATION DES DONNEES EN SHS

Cette chercheure a décidé d'utiliser des SIGs dans sa recherche sur l'art almoravide après avoir pris connaissance d'un projet développé à l'EHESS, qui utilisait le géo-référencement des marques de tailleurs de pierre dans la cathédrale de Tolède. L'objectif était de rendre exploitable et intelligible des données issues d'observations d'éléments d'architecture islamique repérés dans différents sites autour du bassin méditerranéen.

L'OUVERTURE DES DONNEES ACCOMPAGNE LE DEVELOPPEMENT DE L'INTERDISCIPLINARITE

Il existe actuellement en SHS une forte incitation à l'interdisciplinarité et à l'appropriation des technologies numériques par les chercheurs. Cette impulsion accompagne une tendance à l'ouverture des données, même si celle-ci est encore limitée à certaines disciplines.

Elle-même souhaiterait publier et partager les données qu'elle a pu rassembler sur une plateforme en ligne et en libre accès. Deux contraintes retardent cependant cette mise en ligne : la quantité de travail à consentir pour structurer les données collectées et l'étude des contraintes juridiques liées à la publication notamment de données provenant d'observations faites en Syrie.

LE LIBRE ACCES AUX DONNEES PERMETTRA UNE MEILLEURE CIRCULATION DES SAVOIRS

Pour elle, la démarche d'ouverture des données en SHS est extrêmement importante car c'est une manière de faire connaître son travail et de connaître le travail de ses collègues avant leur publication. En SHS, les étapes précédant les publications des résultats peuvent prendre beaucoup de temps, parfois des années, et la circulation facilitée des données permettrait de remédier en partie à ce problème.

ENTRETIEN N°6 (18/06/2014) – DEMOGRAPHIE

Modalité d'entretien

Face-à-face

Statut

Directeur d'études / chercheur CNRS

Domaines de recherche

Histoire sociale des pouvoirs politiques, XVIe et XVIIe siècles, histoire des élites.

Retours / Positions

LA DEMARCHE STATISTIQUE CONDUIT NATURELLEMENT A L'OUVERTURE DES DONNEES

En ce moment, il travaille, conjointement à d'autres membres de son laboratoire, à la collecte de données issues des contrats de mariage parisiens auprès des archives notariales de Paris. Ce travail est quasiment manuel, la plupart de ces archives n'ayant pas été numérisée et se trouvant dans des microfiches. Les données sont ensuite intégrées dans une base de données simple, réalisée sur Excel, et serviront à la réalisation de graphiques et à l'élaboration d'analyses statistiques. Ces tableaux sont partagés par les membres du laboratoire mais une diffusion à plus large échelle est envisagée. Son équipe travaille actuellement aux modalités de cette mise à disposition avec un ingénieur de recherche, spécialiste des problématiques de la « mise en données » des sources et à la constitution de bases de données pour la recherche. L'ouverture de ces données n'est pas particulièrement problématique dans ce cas, car il s'agit de données concernant des personnes disparues depuis longtemps. Méthodologiquement parlant, la mise à disposition de ce type de données est une étape obligatoire de la démarche, pour attester de la pertinence des interprétations que pourrons réaliser les chercheurs « producteurs » et permettre des retours critiques et constructifs.

LES METHODES QUANTITATIVES EN SHS

Actuellement, il y a parmi les historiens et chercheurs en sciences sociales une tension assez forte entre la tendance déconstructiviste de la microhistoire - opposée aux méthodes quantitatives – et le courant de l'histoire qui travaille avec des méthodes statistiques. Quoiqu'il en soit, dans l'un comme dans l'autre courant, des théories orientent la démarche épistémologique.

LA MONTEE EN PUISSANCE DU DATA MINING SERAIT LE REFLET D'UNE CRISE EPISTEMOLOGIQUE DES SHS

Comme l'affirmait Alain Desrosières, la démarche statistique ne peut pas faire l'économie de la construction d'une vraie problématique comme point de départ. La production de données

doit être critique, réfléchie et permettre la prise en compte de facteurs d'erreur. Cette école de pensée a permis des travaux aussi importants que ceux de Bourdieu, Boltanski, Chiapello, Furet, etc. Le recours au data mining dans le contexte de massification des données disponibles prend, en revanche, le contrepied de cette époque d'utilisation construite et réfléchie des données en sciences sociales. Dans ce sens, la montée en puissance serait le reflet d'une véritable crise épistémologique dans les SHS, mettant en doute la capacité des chercheurs de construire des véritables objets de recherche. Rien ne prouve aujourd'hui qu'en sciences humaines et sociales, l'analyse de grands volumes de données, quelles que soient les performances des outils utilisés, ne débouche sur un niveau de compréhension supérieur des problématiques abordées.

ENTRETIEN N°7 (18/06/2014) – PHILOSOPHIE

Modalité d'entretien

Face-à-face

Statut

Directeur de recherche et directeur de revue

Domaines de recherche

Philosophie allemande contemporaine, métaphysique, phénoménologie et logique.

Retours / Positions

EN PHILOSOPHIE, LES CHERCHEURS ONT ESSENTIELLEMENT EN TETE LA PUBLICATION EN LIVRE OU EN REVUE

Pour ce chercheur en philosophie, les « données » sont avant tout des corpus de textes ou des éditions de référence, comme par exemple les Husserliana dans l'édition réalisé par Herman Van Breda.

L'ouverture des données de la recherche et leur diffusion ne sont pas, à son sens, une problématique susceptible de concerner les travaux comme ceux qu'il mène. Peut-être est-ce là un enjeu important pour d'autres disciplines, mais pour la philosophie, les chercheurs

visent essentiellement la publication de leurs travaux sous forme de livre ou d'articles de revue, après un travail éditorial soigné.

Il est vrai qu'en philosophie, les chercheurs travaillent et publient à une vitesse moins contraignante que dans d'autres disciplines, même si cela a beaucoup changé depuis qu'on pousse les chercheurs à accumuler les citations pour être évalués comme actifs et intéressants.

LA NUMERISATION D'ANCIENS NUMEROS DES REVUES EST PAR CONTRE UN ENJEU ESSENTIEL

A son avis, à l'heure qu'il est, le grand enjeu pour la philosophie est la numérisation d'anciens numéros de revues. Des articles parfois très importants sont devenus rares et difficilement accessibles. Etant lui-même directeur de revue, il se sent particulièrement concerné par cette question.

L'UTILISATION DU « TEXT MINING » POURRAIT ETRE UTILISE POUR L'ETUDE DE CERTAINS AUTEURS

Le text mining pourrait éventuellement avoir un intérêt en philosophie pour l'étude de certains auteurs, comme Nietzsche par exemple, dont on a récemment découvert des manuscrits avec des versions différentes de certains textes aujourd'hui étudiés.

ENTRETIEN N° 8 (18/06/2014)- SCIENCES POLITIQUES

Modalité d'entretien

Skype

Statut

Chargé de recherches

Domaines de recherche

Comportements électoraux et politiques en France

Retours/Positions

PROBLEMATIQUE GENERALE DES DONNEES QUANTITATIVES EN FRANCE

Il ne produit pas lui-même des données d'enquêtes et des grandes enquêtes, ces dernières réalisées la plupart du temps par des instituts de sondages et parfois par des chercheurs. Son travail principal consiste à dépouiller ces données et à en réaliser l'analyse.

Sa façon de comprendre la question des données de la recherche est ainsi très large et pourrait difficilement être transposée en seulement quelques lignes. Néanmoins, dans ce qui concerne uniquement les données des grandes enquêtes, il entrevoit les problématiques suivantes :

- Le coût élevé du financement des prestataires et instituts de sondages pour la réalisation des grandes enquêtes alors que…

- …très souvent, les différentes formes de réutilisation des données d'enquête souhaitées par les chercheurs ne sont pas prévues en amont à la préparation de ces enquêtes. La nature hétérogène de ces données rend difficile une exploitation optimale et génère des problèmes de sous-exploitation de ces données.

- Problématique d'archivage et stockage de ces données, questions centrales aujourd'hui comme par le passé. Il suffit de rappeler la volumétrie considérable des données des années 60 et 70 qui se sont perdues.

- La France est en retard, en comparaison aux EUA, dans l'établissement des bonnes pratiques concernant la politique des données électorales

QUESTION CULTURELLE MAIS AUSSI POLITIQUE POUR LA GESTION DE CES DONNEES EN FRANCE

Par ailleurs ces problématiques, il identifie deux sources de problèmes ayant des origines, d'une part, dans les pratiques des chercheurs concernant les données de la recherche et d'autre, dans le pouvoir centralisateur des institutions qui détiennent et contrôlent ces données.

Concernant les pratiques des chercheurs, le paysage en France est partagé : alors qu'on assiste à une ouverture progressive de ces données par des initiatives individuelles ou collectives, il est à la fois toujours facile de constater l'existence d'une culture de la propriété des données dans le milieu scientifique et des modes de travail plus individualistes. Hormis quelques disciplines pour lesquelles l'ouverture des données est une obligation

méthodologique, les chercheurs ne sont pas tous à l'aise avec l'idée de permettre des vérifications de leurs résultats ou avec le fait de dévoiler leur méthodologie - ce qui découlerait naturellement d'une ouverture des données.

Il serait souhaitable, à son avis, d'évoluer vers des nouvelles formes de recherche en SHS, et de saisir les bénéfices des nouveaux usages émergents du numérique.

LES POLITIQUES INSTITUTIONNELLES DE GESTION DES DONNEES LIMITENT LEUR USAGE DANS LA RECHERCHE

En France il est très important que les institutions qui gèrent et détiennent des données d'enquêtes ou des statistiques trouvent un compromis entre l'objectif d'ouverture de ces données aux chercheurs et leurs politiques de contrôle sur l'accès et sur les usages que le chercheur en fera. Aujourd'hui, le déséquilibre est prédominant au détriment de la liberté des chercheurs. Le réseau Quetelet, par exemple, avec son lourd système bureaucratique est un frein à l'initiative dans la recherche.

Un autre frein qui doit être évoqué concerne la durée de l'embargo pour les données. Il est personnellement favorable à leur mise à disposition immédiate.

Les chercheurs peuvent, parfois, contourner ces dispositifs et faire le « one to one », c'est-à-dire se transmettre des données entre eux. A sa connaissance, ces pratiques restent assez limitées à des projets menés collectivement.

UN USAGE LIMITE DU DATA MINING EN SHS

Il n'utilise pas les procédés du data mining et estime que leur usage est, pour l'instant, assez limité en SHS. Le modèle épistémologique prédominant des recherches en SHS, c'est-à-dire, théorique, a bien des années de survie devant lui et n'est pas, à son avis, sérieusement mis en question face à l'apparition de ces technologies.

ENTRETIEN N°9 (20/06/2014) – ANTHROPOLOGIE

Modalité d'entretien

Téléphonique

Statut

Chargée de recherche au CNRS

Domaines de recherche

Eco-anthropologie et ethnobiologie

Retours / Positions

LE PARTAGE DES DONNEES DE TERRAIN POSE DE NOMBREUX PROBLEMES

Cette chercheure développe des recherches dans un domaine assez peu connu : l'anthropologie de l'environnement. Il s'agit d'un champ de recherche en plein développement, porté notamment par les travaux de Bruno Latour et de Philippe Descola, opérant par analyse de controverses sur les questions environnementales.

Elle travaille donc avec des données ethnogéographiques : des enregistrements (témoignages) et des carnets d'observations de terrain. Le partage des données se fait uniquement au sein de son équipe et une diffusion plus large semble difficilement envisageable. Tout d'abord parce que les données nécessitent d'un travail de contextualisation important pour avoir une valeur scientifique à part entière. Ensuite parce que ces données touchent de près la vie des personnes concernées et ne peuvent / doivent pas être diffusées sans leurs accords. A ce jour, de simples accords oraux de confidentialités sont passés avec les personnes qui ont témoigné, autorisant l'utilisation de ces enregistrements dans le cadre restreint de l'équipe. Mais la déontologie empêche d'aller plus loin. En règle générale, les anthropologues ne partagent pas forcément leurs données de terrain et quand cela arrive, c'est dans le cadre d'un projet collectif.

MUTUALISER CERTAINS TRAVAUX TECHNIQUES POUR LA COMMUNAUTE SCIENTIFIQUE

Elle a déjà réalisé des analyses de données textuelles de type « text mining » lors d'un projet développé avec une collègue sociologue. Elles se sont servies pour ce faire du logiciel open source Prospéro et ont analysé un corpus de 1.400 pages web.

La plus grande difficulté a été de formater des documents mis en ligne au format PDF dans un format interprétable par le logiciel, ce qui a été très long et très pénible. Dans ce cas, elle pense qu'il serait souhaitable de mutualiser les tâches préparatoires à ce type d'analyse de

déposer les textes « traités » dans une base de données commune. Il serait par-là possible de créer une communauté scientifique tout en instaurant une éthique de travail entre chacun.

A son sens, les nouvelles technologies ne révolutionnent pas fondamentalement les pratiques de la recherche dans son domaine. Elles rendent cependant possible de nouvelles modalités d'exploitation des données et apportent des solutions aux problèmes de conservation et de stockage des informations. Les logiciels de traitement de corpus textuels constituent, par exemple, une véritable opportunité pour les SHS, mais sans l'observation de terrain et la modélisation théorique qui constituent des étapes essentielles du processus de recherche, ces outils n'auraient pas de sens.

Lorsqu'elle a participé, avec d'autres chercheurs, à une analyse de controverses environnementales, celle-ci été orientée préalablement par des hypothèses. Le logiciel utilisé a ensuite été paramétré en fonction des catégories préétablies avant l'étape d'automatisation de l'analyse du corpus. Ils ont ainsi élaboré des dictionnaires (lexiques) mais qui ne peuvent être utilisés en dehors du contexte de cette recherche particulière.

ENTRETIEN N°10 (20/06/2014) – ARCHEOLOGIE

Modalité d'entretien

Téléphone

Statut

Chargée de recherches CNRS

Domaines de recherche

Archéologie environnementale, géoarchéologie, paléoenvironnements

Retours / Positions

ACCROITRE LA VISIBILITE DES DONNEES CONSIDEREES COMME « MINEURES »

En matière de données, elle produit des mesures physico-chimiques sur les archives sédimentaires. Les données sont brutes et ne sont qu'une partie du résultat qui implique, lui, l'interprétation des données. Les résultats obtenus sont ensuite publiés en revue.

Les données sont parfois réexaminées à la lumière de nouveaux résultats, pour les comparer et les préciser, ou pour tirer des conclusions sur d'autres échelles d'analyse (plus petites ou plus grandes).

En matière de réutilisation des données, elle utilise parfois des données anciennes et des données produites par d'autres collègues. Cependant, cela se passe généralement dans le cadre de collaborations avec ses collègues, car les chercheurs aiment être systématiquement associés aux réflexions issues de leur travail.

A son avis, la publication reste le meilleur support de diffusion. Il faut cependant offrir des supports pour diffuser les données qui paraissent mineures, car les revues ne font de la place que pour les données qui apparaissent a priori comme principales.

ENTRETIEN N°11 (20/06/2014) – HISTOIRE

Modalité d'entretien

E-mail

Statut

Directeur de recherche

Domaines de recherche

Histoire des sciences de l'homme et histoire de la justice

Retours / Positions

LES DONNEES DE LA RECHERCHE DOIVENT ETRE DISTINGUEES SELON LEUR DEGRE D'ELABORATION

La notion de « données de la recherche » peut s'entendre de deux manières au moins, soit qu'il s'agisse de nommer les données sur lesquelles s'appuie la recherche ou les données produites par la recherche. Dans le cadre du tournant numérique des SHS, les données produites par la recherche bénéficient d'une nouvelle exposition et de possibilité d'accessibilité, d'enrichissement et d'interopérabilité accrue.

Sur la plateforme hypermédia conçue par lui et des coopérateurs, cette distinction a été réalisée pour classer les données selon leur degré d'élaboration. Les « données brutes », annotées ou enrichies de métadonnées, sont classées dans les « sources ». Les données secondaires sont le produit d'une élaboration qui suppose – dans la plupart des cas - de déstructurer le document initial (agrégation, traitement, extraction sélective) et de s'affranchir de la représentation initiale du support pour élaboration d'une nouvelle visualisation des données. Ces données produites par la recherche sont plus proches d'« outils » ou d'instruments de recherche (typiquement, les « bases de données »).

DIFFUSER ET PARTAGER LES OUTILS ET INSTRUMENTS DE RECHERCHE SE FAIT DE PLUS EN PLUS FREQUEMMENT

Suivant la distinction proposée ci-dessus, les chercheurs ont plus naturellement compétence et intérêt à diffuser des données de type instruments de recherche, souvent élaborées au cours d'une recherche, même si les recoupements avec la mise à disposition ou l'aide à l'accès aux données sources sont fréquents (ex : bibliographie, catalogue manuscrit, catalogue de ressources etc.). La plateforme collective créée propose ce type de partage, et il utilise pour ses propres recherches des outils et instruments de recherche parfois conçus par d'autres chercheurs.

Les modalités de diffusion ou de publication de ces données sont multiples et dépendent du type de données, des objectifs de la publication et du public visé. Un modèle général est difficilement imaginable, mais celui proposé par Robert Darnton peut être considéré comme théoriquement idéal : le chercheur présente un résultat en donnant accès à ses sources, qui constituent alors des éléments de preuves et de vérification.

LES NOUVELLES TECHNOLOGIES SE HEURTENT CEPENDANT A L'IMPORTANCE DE L'INTERPRETATION EN SHS

Le « data mining » et la « Fouille de textes » sont des procédés qu'il n'utilise pas pour l'instant, même s'il n'écarte pas la possibilité de les utiliser un jour. Cependant, quelques doutes sur l'analyse automatique des données subsistent car les SHS restent des disciplines du sens et de l'interprétation. L'élaboration de la procédure de fouille doit donc être très bien problématisée dès le départ.

ENTRETIEN N°12 (23/06/2014) – ARCHEOLOGIE

Modalité d'entretien

E-mail

Statut

Directrice de recherche/chercheure

Domaines de recherche

Étude de la vie matérielle médiévale en Europe occidentale à partir des sources iconographiques, écrites et archéologiques

Retours / Positions

LES « DONNEES DE LA RECHERCHE » : UNE POLYSEMIE

Dans son domaine, les données de la recherche impliquent au moins un double sens. Ce sont d'une part des outils : dans son cas, une base de données d'environ 40.000 images, des carnets de fouilles, des plans, des dessins, des croquis issus de ses propres recherches ou de celles d'autres chercheurs laboratoire, numérisées et indexées.

Il s'agit, d'autre part, des ressources, d'articles, des livres, des synthèses, des notes de séminaires et communications des colloques produits aussi par les chercheurs ayant pour objectif la transmission des résultats des recherches à un public plus ou moins spécialisé. Cependant, une définition unique de « données de la recherche » en SHS lui semble assez difficile.

DES PRATIQUES DE PARTAGE POUR LA CONSTRUCTION D'UN EDIFICE COMMUN DE LA RECHERCHE

Après trente ans d'activité scientifique, elle continue à penser que la recherche est un édifice construit à plusieurs et que le partage des données est un but et une nécessité. De plus, cela accroît la visibilité des chercheurs et donne lieu à de nouveaux échanges. Très fréquemment sa base de données d'images est mise à disposition d'étudiants ou de chercheurs. Pour sa propre recherche, elle utilise souvent des données issues des travaux d'autres chercheurs, que ce soit des thèses mises en ligne, différents sites scientifiques, des revues en ligne, des bases de données d'images issues de différentes bibliothèques.

Le type de diffusion pour ces données est prioritairement électronique bien que, à son avis, le support papier continue d'être essentiel et souhaitable. Mais c'est certainement très positif

que les chercheurs déposent sur Hal leurs productions et que les revues soient numérisées par différents sites d'hébergement (Persée, Revue.org, Cairn).

DES NOUVEAUX MOYENS POUR LA RECHERCHE GRACE AU NUMERIQUE

Chercheuse du « temps des livres » et peu au courant des procédés offerts par les nouvelles technologies, le data mining et la fouille de textes restent à l'écart de sa pratique. Le numérique a sans doute transformé les pratiques des SHS dans le sens d'une accessibilité simple et rapide à des données, riches en nombre et en diversité, qu'il aurait été impossible d'avoir auparavant. Les chercheurs aujourd'hui peuvent réaliser leurs recherches plus rapidement grâce à des points d'accès simplifiés aux sources et à son coût financier relativement bas. Il y a à peine vingt-cinq ans, par exemple, il était impossible de donner à faire à un étudiant un mémoire de recherche en iconographie médiévale. L'achat des clichés était beaucoup trop onéreux et cela impliquait un temps long de dépouillage de manuscrits par manuscrit des fonds des bibliothèques. Aujourd'hui, l'accès à un ensemble de collections et à des manuscrits entièrement numérisés est possible sans effort. En revanche, une grande prudence est nécessaire dans le choix de données, parfois à valeur très inégale, et leur rapprochement, au risque de réaliser des inventaires à la Prévert.

ENTRETIEN N°13 (23/06/2014) – PHILOSOPHIE

Modalité d'entretien

Téléphonique

Statut

Philosophe, directrice de recherche et directrice de revue

Domaines de recherche

La condition animale, phénoménologie

Retours / Positions

LA DIFFUSION SCIENTIFIQUE REPOSE ESSENTIELLEMENT SUR LA PUBLICATION EN REVUE

A son sens, les données utilisables dans son domaine sont avant tout des corpus de textes philosophiques. Elle travaille aujourd'hui essentiellement en se déplaçant dans des bibliothèques, et en se servant de plateformes d'édition en ligne comme Jstor et Cairn, mais ne trouve pas confortable la lecture sur support numérique.

Concernant son processus de recherche, elle considère qu'il n'a pas d'étapes intermédiaires susceptibles d'être diffusés avant une publication. Par ailleurs, elle ne diffuse pas non plus ses contributions à l'occasion de colloques, car elle trouve déjà très difficile de supporter les enregistrements de ces évènements où les réponses sont souvent faites à chaud et les positions prises sans temps de réflexion. Elle estime qu'il y actuellement une quête incessante à tout enregistrer, et ne partage pas l'engouement de certains à avancer dans ce sens. Le droit à l'oubli est, à son avis, une composante de la liberté humaine.

Afin de diffuser son travail de chercheuse, elle rédige des articles dont la plupart sont sollicitée et commandée par les revues. Récemment elle a pris conscience de l'importance d'accroître sa visibilité sur Internet et a créé un site rassemblant une grande partie de sa production scientifique qui se trouve en accès libre.

Elle dirige une revue dont tous les articles sont consultables et téléchargeables en accès libre.

ENTRETIEN N°14 (23/06/2014) – SOCIOLOGIE

Modalité d'entretien

téléphone

Statut

Chercheur/ Maître de conférences

Domaines de recherche

Sociologie des TICs, innovation dans le monde des médias et de la communication

Retours / Positions

DOCUMENTER ET CONTEXTUALISER LES DONNEES « BRUTES » POUR LES RENDRE REUTILISABLES

Il comprend tout d'abord « données de la recherche » comme celles produites par les disciplines des SHS adoptant une méthodologie quantitative. Il faut cependant distinguer les

données créées par les chercheurs dans des cadres institutionnels et gouvernementaux précis, comme celles, certifiées, de l'INSEE, et les données produites par les chercheurs au cours de leurs activités scientifiques.

Un de ses travaux de recherche a impliqué la collecte d'une grande quantité de données dans les forums de presse en ligne. Ces données ont été ensuite intégrées à une base de données et enrichies d'autres données contextuelles (données sur les personnes, contexte de la collecte, données de l'INSEE, etc.).

Les cartographies et l'article issus du travail sur ces données ont été partagés, mais pas la base de données elle-même. En effet, le partage de ces données « brutes » n'aurait pas été de grand intérêt en tant que source d'information. D'autre part, il y aurait certainement un grand intérêt à contextualiser, documenter et décrire ces données pour les rendre réutilisables dans de bonnes conditions.

TROIS DIFFICULTES A L'ADOPTION D'UNE « CULTURE DES DONNEES » EN SHS

En ce sens, il pense que les chercheurs doivent commencer à documenter davantage leurs données en vue de leur ouverture, mais qu'un certain nombre de contraintes freinent actuellement ce mouvement.

En premier lieu, les difficultés d'ordre juridique relevant parfois du type de données mais aussi de la difficulté à analyser au cas par cas les implications juridiques d'une ouverture.

Deuxièmement, les chercheurs devraient interagir davantage avec des équipes ou des personnes ayant une « culture des données » - des statisticiens notamment - habitués à traiter de façon qualitative les données.

Enfin, une autre difficulté : les algorithmes qui sont derrière l'analyse des données doivent eux-mêmes être l'objet de justification et de documentation pour qu'une réutilisation soit possible.

OUVRIR, CERTES, MAIS COMMENT ?

La réflexion sur l'ouverture des données de la recherche dans le contexte de l'Open Data doit d'abord répondre à la question suivante : quelles plateformes d'exploitation et de diffusion créer ? Le faire de façon centralisé comme le fait l'INSEE, ou de façon plus éparse ? Comment valoriser les données pour qu'elles puissent toucher leurs publics potentiels ?

A son avis, la question de la publication éditoriale et de la publication ou diffusion des données sont deux choses très différentes, qui peuvent sûrement avoir des liens (enrichissements) mais qu'il est important de dissocier. Il ne faut pas oublier que le travail du chercheur n'est pas seulement de produire des données, mais aussi de réaliser des comptes rendus qui synthétisent et interprètent ces données.

LE DATA MINING POUR DES PROJETS TRES PRECIS

Le data mining est un procédé qui se justifie pour des projets ponctuels et innovants, par exemple, des projets de cartographie du Web de Medialab ; « La fabrique de la loi » (Medialab et CEE-SciencesPo) et le projet de l' OTMedia (Observatoire TransMedia). « La fabrique de la loi » combine trois sources :le catalogue Open Data du Sénat, les APIs de Nosdéputés.fr et de NosSénateurs.fr et les sites officiels du Sénat et de l'Assemblée Nationale.

ENTRETIEN N°15 (23/06/14) – PHILOSOPHIE ET ANTHROPOLOGIE DES SCIENCES ET DES TECHNIQUES

Modalité d'entretien

e-mail

Statut

Directeur de recherche

Domaines de recherche

Enjeux stratégiques et géopolitiques des technologies

Retours / Positions

EN FRANCE, LA REFLEXIVITE DES SCIENTIFIQUES SUR LEURS PROPRES PRATIQUES EST ENCORE FAIBLE

Il estime que la question des données de la recherche est aujourd'hui abordée de manière large par un nombre restreint d'équipes de recherche en France. Ce sont en général des équipes transdisciplinaires qui tentent d'éviter un certain nombre de postures idéologiques et des préjugés habillés dans le "sens commun" des doxas héritées de longues années

d'enfermement dans un modèle scolastique, dominant et réducteur, des pratiques scientifiques.

La question des données doit être aujourd'hui examinée dans le contexte des débats autour de différentes économies des politiques scientifiques et écologies cognitives qui sont à l'horizon des collectifs de pensée, en France comme à l'étranger. A son sens, dans un contexte si large il est néanmoins possible de définir quelques axes orientant l'analyse :

- systèmes de critériologies et d'évaluations scientifiques ;

- les dimensions plus ou moins processuelles des pratiques éditoriales

- le Content-Mining et les nouvelles pratiques de cartographies etc.

- le développement de nouvelles fonctionnalités éditoriales surtout dans le secteur des sciences dures et moins, paradoxalement, dans le secteur des SHS, apparemment englué dans une approche faible et parfois infantile des «Digital Humanities».

A son sens, ce qui rendrait aujourd'hui difficile une analyse de ces questions du point de vue des pratiques des acteurs intéressés, c'est-à-dire les chercheurs, c'est que la réflexivité des scientifiques en sciences dures et en SHS sur leurs propres pratiques d'écriture et d'édition n'est pas forcément forte. De ce fait, certains travaux de sociologie récents mais aussi moins récents continuent d'être indispensables à l'analyse de la situation.

Il est aussi remarquable qu'en France les questions juridiques relatives au "content mining », sur les travaux/publications et sur les données, soient insuffisamment traitées et mieux abordées à l'étranger. Ces questions peinent à se frayer un chemin allant vers des transformations pratiques ou des remises en cause d'un cadre juridique devenu obsolète face au numérique. Ceci est dû très probablement au poids des conservatismes institutionnels encore très présents en France et en décalage par rapport aux transformations des sociétés de l'information. Cela mériterait certainement une analyse réflexive plus fine de la part des chercheurs ; or, aujourd'hui ces questions essentielles qui concernent la circulation et la transmission des savoirs n'occupent pas une place d'importance dans les recherches.

Pourtant, à son avis, les données de la recherche en SHS sont aujourd'hui un terrain fécond à la production de nouveaux modèles éditoriaux et l'occasion de bâtir une nouvelle culture scientifique, avec des conséquences épistémologiques et pratiques, dont les contours restent difficiles à définir à l'heure qu'il est, mais qui pourrait s'annoncer plus démocratique. Il est probable que les initiatives le plus intéressantes ne viendront pas du côté des institutions,

trop attachées encore à un modèle centralisateur et uniformisateur, mais plutôt des éditeurs et de quelques collectifs de chercheurs.

ENTRETIEN N°16 (24/06/2014) – PHILOSOPHIE

Modalité d'entretien

Face-à-face

Statut

Chercheur CNRS, enseignant

Domaines de recherche

Philosophie allemande post-hégélienne, philosophie politique, philosophie contemporain

Retours / Positions

LES REVUES DONNENT UNE IDENTITE A LA RECHERCHE EN PHILOSOPHIE

Les « données de la recherche » peuvent se comprendre de diverses façons : les données des bases de données, les données du Web (Big Data) et, en philosophie, les ressources elles-mêmes, c'est-à-dire les corpus de textes philosophiques.

Il se rend compte des changements liés à l'utilisation des nouvelles technologies en SHS, mais en philosophie cela reste discret. Néanmoins, l'accès rapide à des documents numérisés dans des bases de données rend le travail de recherche beaucoup plus dynamique.

En philosophie, la publication en revue permet d'identifier clairement les domaines de recherche, les réseaux politiques et philosophiques d'un chercheur, ce qui est très important pour comprendre la matière proposée.

Il a des collègues qui s'intéressent de près à l'analyse exploratoire des données textuelles, appliquée notamment à des auteurs dont les œuvres ont été publiées et traduites plusieurs fois en France, comme Heidegger, par exemple.

Il investit actuellement les réseaux sociaux et anime un blog, il produit ainsi un certain nombre de données d'« autopromotion ».

ENTRETIEN N°17 (24/06/2014) – SOCIOLOGIE

Modalité d'entretien

Face-à-face

Statut

chercheur/ATER

Domaines de recherche

Sociologie de l'expertise psychiatrique

Retours / Positions

Dans son dernier projet de recherche, il a travaillé avec des corpus composés de trois types de données qualitatives :

- soixante-et-onze entretiens enregistrés auprès des experts psychiatriques,
- des notes d'analyse d'archives des affaires judiciaires
- des observations non participantes de vingt-deux procès.

Il a partagé ces données uniquement avec des collègues et s'est posé récemment la question d'une diffusion plus large, mais le travail de mise en forme et de documentation de ces données est complexe et exigerait peut-être d'y consacrer beaucoup de temps. Par exemple, la gestion des questions d'éthique et de confidentialité pour chacune de ces données serait particulièrement longue à faire, ainsi que la mise en forme de l'historique de chaque procès.

Il serait volontaire à le faire mais dans le cadre d'un programme institutionnel incluant la communauté scientifique dans son ensemble. A son sens, la reconnaissance de ce type de travail de préparation des données doit être objet d'une reconnaissance institutionnelle à part entière, au même titre que les résultats de la recherche. Cette valorisation est fondamentale pour motiver les chercheurs qui, autrement, n'y verraient pas d'intérêt.

MUTUALISER LES DONNEES MAIS AUSSI LES METHODES DE TRAITEMENT DES DONNEES

Dans ces conditions, il estime qu'il serait très intéressant que les chercheurs de différents domaines de spécialisation, puissent mutualiser, en plus de données, les différentes méthodologies de traitement de ces données. En ce qui concerne sa recherche, connaître les méthodologies de traitement des données qualitatives d'autres projets de recherche serait très utile.

A son voir, cette orientation constituerait un vrai projet collaboratif de construction d'outils pour la recherche, alors que le simple partage des données « brutes » lui semble dépourvu de sens. En absence de la méthodologie de collecte et de traitement de ces données, leur valeur épistémologique est contestable.

Ce raisonnement peut être appliqué aux éventuelles pratiques éditoriales qui souhaiteraient diffuser les données en parallèle aux articles. Il est certainement intéressant d'offrir l'accès à des corpus d'entretiens, mais uniquement si le chercheur a pris les mesures adéquates de traitement de ces données en amont et explicité sa méthode.

CULTURES DE PROPRIETE ET DE PARTAGE : LES CHERCHEURS ENCORE DIVISES

En France, les chercheurs en SHS ont pris un peu de retard, à son avis, dans la réflexion sur les conditions souhaitables pour un partage de données en SHS. Ils risquent d'être surpris par des décisions prises par les instances supérieures et ressentir des politiques en faveur de l'ouverture comme contraignantes. La question urgente à se poser et à répondre est, à son avis : « A qui appartiennent les données ? ». Les chercheurs sont encore divisés entre une culture de la propriété intellectuelle et une culture du partage.

Il ne connaît pas le data mining et ne travaille pas avec de données quantitatives.

ENTRETIEN N°18 (25/06/2014)- LITTERATURE

Modalité d'entretien

téléphone

Statut

Chargé de recherches CNRS

Domaines de recherche

Théorie littéraire appliquée à la littérature française contemporaine

Retours / Positions

Il y a actuellement, à son avis, un bouleversement de la « vie littéraire » dû à l'avènement du numérique. Les conditions de la création littéraire ont changé, des nouveaux supports existent et en conséquence des nouveaux genres apparaissent. En parallèle, des nouvelles pratiques de lecture surgissent. Il est donc de s'attendre que tous ces changements favorisent des nouvelles formes de diffusion de l'information scientifique.

Il estime que parler des « données de la recherche en SHS » doit forcément amener à un effort de définir des typologies de données pour un paysage qui reste actuellement très flou et indéfini. La datafication – « tout est donnée » - concernant tout type d'information rend ce travail indispensable.

LES DONNEES D' « AUTO-DOCUMENTATION » DU CHERCHEUR

De manière générale, il aperçoit dans son domaine trois types des données :

- Bibliographiques, avec création de bases de données bibliographiques ou corpus de références ;

- Contenus : analyses de discours littéraires ;

- Données d'éditions textuelles ou, autrement dit, les sources littéraires elles-mêmes.

Dans le contexte compétitif de l'évaluation scientifique que connaissent les sciences aujourd'hui, un autre type de donnée envahit l'espace web, qui mérite une attention particulière: les données d'auto-documentation du chercheur.

Produites par le chercheur lui-même, ces données sont à la fois une forme de promotion de sa recherche et une manière d'augmenter sa visibilité sur le Web.

Ces données peuvent se présenter très différemment, allant de la simple forme d'un cv en ligne, passant par les réseaux scientifiques, jusqu'à des formes plus élaborées comme les

Carnets de la recherche, offrant souvent des possibilités de télécharger des articles. Directe ou indirectement liées au processus de la recherche lui-même, c'est un fait que la production de ce type de données connaît une augmentation d'importance aux yeux des chercheurs et qu'elle consomme un temps encore sous-évalué de l'activité scientifique. A son sens, il faudrait s'attarder sur ce type de données pour examiner, d'une part, si une nouvelle forme de communication scientifique, parallèle à la publication, se développe et si, d'autre part, des initiatives individuelles de partage d'articles ou des données sont en cours à travers les réseaux sociaux.

ENTRETIEN N°19 (26/06/2014) – SCIENCES POLITIQUES

Modalité d'entretien

Skype

Statut

professeur associé

Domaines de recherche

Méthodes et outils numériques, cartographie, innovation sociale

Retours / Positions

DEVELOPPER DES METHODES ET OUTILS NUMERIQUES POUR INNOVER DANS LA RECHERCHE EN SHS

Les données dans le cadre de son activité sont des données numériques. Cela peut correspondre à des données nativement numériques ou numérisées à posteriori. Il réalise des cartographies des données du web et plus spécifiquement, des cartographies de controverses, dans le cadre d'un projet Ce projet, se développe sur deux volets, technologique et pédagogique, visant innover dans la pédagogie de l'enseignement supérieur à travers le développement de méthodes et d'outils permettant aux étudiants d'appréhender des phénomènes complexes de société.

Il coordonne, par ailleurs, plusieurs projets pilotés par des équipes de chercheurs émanant de différentes disciplines des SHS, souhaitant inscrire les outils numériques au cœur de leur pratique de recherche. La demande la plus fréquente vient des équipes ou laboratoires travaillant avec des données textuelles et des disciplines littéraires.

L'objectif du travail de soutien technique et méthodologique vise à rendre exploitable des données hétérogènes, issues d'approches diverses, ayant des formats différents et véhiculant des informations de nature variée. A travers la prise en charge technique il est possible de comparer et de croiser ces données. Cela se traduit, ensuite, par une représentation visuelle, cartographie ou autre, de ces données et de ses interactions, appuyant ainsi le processus interprétatif d'analyse.

La tendance à réaliser ce genre de démarche exploratoire et expérimentale, commence à se répandre au sein des Humanités numériques qui souhaitent innover le processus de recherche en SHS à travers les possibilités offertes par les outils numériques.

Les données produites par ces analyses sont destinées à la publication sur des plateformes.

DEMULTIPLIER LES PLATEFORMES DE CES DONNEES POUR TESTER LEUR PERTINENCE EN PETITE ECHELLE

Pour lui, il est important, à l'heure qu'il est, de démultiplier ces projets expérimentaux de plateformes tout en conservant leurs dimensions réduites pour une meilleure prise en charge des imperfections et des défauts. Le traitement des données reste aujourd'hui très instable pour envisager une centralisation de ces plateformes.

De manière générale, l'ouverture des données de la recherche permettrait une plus grande transparence du processus des recherches et servirait à vérifier leur solidité. C'est un objectif louable mais qui doit aller de pair avec un travail sérieux de mise en forme de ces données visant leur interopérabilité c'est-à-dire, un travail sur les codes qui doivent également être mis à disposition.

ENTRETIEN N° 20 (27/06/2014) – SCIENCES POLITIQUES

Modalité d'entretien

e-mail

Statut

directeur de recherche

Domaines de recherche

Relations internationales et sociologie politique de la globalisation, Turquie, Iran, Afrique sub-saharienne, théorie et méthodologie de l'Etat

Retours / Positions

UN PARTAGE DES DONNEES RESTREINT DANS L'ETUDE DES RELATIONS INTERNATIONALES

Dans son travail de recherche, il est amené à travailler avec une variété importante de données. Les sources secondaires constituent une partie conséquente de son matériau de recherche, mais il travaille essentiellement sur des observations de terrain, des entretiens qualitatifs et de la documentation primaire (archives, documents divers). Ces données sont intégrées dans ses écrits scientifiques (livres, articles). Mais leur caractère, qualitatif et même souvent confidentiel, exclut un partage plus formel et plus systématique de ces sources - ce qui n'empêche pas des partages plus ponctuels et informels avec des collègues, sur la base d'une relation de confiance quant à leur usage.

UN TIRAILLEMENT ENTRE L'ETHOS SCIENTIFIQUE ET LA COMPLEXITE DES SITUATIONS

A priori, le partage des données relève de l'ethos scientifique, mais celui-ci ne peut être défini que de manière pragmatique, selon les circonstances. Par exemple, tous ses dossiers relatifs à différents pays d'Afrique sont ouverts à ses collègues qui en font la demande, pourvu qu'il soit certain qu'ils en feront un usage déontologiquement conforme à la manière dont ils ont été constitués (confidentialité des sources, anonymat des enquêtés, non citation des sources classifiées, etc.). Par ailleurs, il utilise des données mises à disposition par des chercheurs, par exemple des fonds déposés dans des archives. Il est certain qu'Internet offre des possibilités intéressantes de publication et de diffusion, mais cela ne peut être systématique et doit se conformer aux règles susmentionnées. Compte tenu du caractère très qualitatif de sa démarche, il n'envisage pas d'utiliser les procédés de collecte et analyse automatisée de données.

Une politique d'ouverture des données de la recherche doit être conçue en prenant en considération toute la gamme des problématiques qui peuvent apparaitre, et surtout éviter un surcroît de bureaucratisation hautement nuisible à la productivité des chercheurs. Il faut bien comprendre que le travail de ceux-ci est de plus en plus pollué par la lourdeur des procédures du New Public Management.

ENTRETIEN N°21 (30/06/2014) – ARTS/ CINEMA

Modalité d'entretien

téléphone

Statut

Chercheur/professeur d'université

Domaines de recherche

Esthétique du cinéma et de l'art des nouveaux média

Retours / Positions

DES DONNEES « SOURCES » POUR LA RECHERCHE EN ESTHETIQUE DU CINEMA

Elle travaille dans le domaine de l'esthétique du cinéma et de l'art des nouveaux médias.

Dans sa recherche, elle se sert de sources diverses : documentations bibliographiques et images animées.

De son point de vue, les technologies du numérique facilitent beaucoup la recherche même si la France a, à son avis, beaucoup de retard par rapport à l'Allemagne qui propose plus de ressources internationales et moins de barrières pour y accéder. Elle réalise ses recherches bibliographiques sur le web, consulte les articles surtout sur Jstore et documents et archives d'imagés animées à travers le catalogue de la FIAF (International Federation of Films Archive).

Dans sa recherche, les données qu'elle produit sont vouées au travail de publication éditoriale, les étapes intermédiaires ne faisant objet d'aucun type de partage, sauf lors des conférences et colloques où il lui arrive de dévoiler un certain nombre de réflexions qui sont, pour ainsi dire, des étapes non formalisées de sa recherche. Mais cela reste purement occasionnel et se fait dans un cercle de chercheurs travaillant sur des thèmes proches. D'une manière générale, elle n'envisage pas de diffuser ses notes en amont à la publication car elles correspondent à des étapes d'une réflexion personnelle originelle.

ENTRETIEN N°22 (01/07/2014) – SOCIOLOGIE

Modalité d'entretien

Face-à-face

Statut

chercheur

Domaines de recherche

Les médias alternatifs, mouvement altermondialiste et les organisations militantes internationales, nouveaux médias, Web 2.0

Retours/ Positions

LES DONNEES DE LA RECHERCHE DE « L'AVANT-NUMERIQUE» : COMMENT LES VALORISER ?

La questions des données de la recherche en SHS se pose, à son avis, selon des problématiques relatives à deux types de données :

- données d'archives existant dans d'autres supports physiques et en d'autres formats (non numériques) ;

- données nativement numériques.

Alors que les programmes massifs - hautement couteux- de numérisation d'archives sont la conséquence des dispositifs visant l'exploitation du premier type de données, pour le deuxième on retrouve des problèmes liés à la pérennité, stockage, structuration et mises à disposition de ces données. Logiquement, une fois numérisées, les données du premier type posent des problèmes similaires à celles du deuxième type.

Une question qui mérite d'être posée est celle de savoir comment valoriser les données «produites » par les chercheurs avant la généralisation de l'environnement numérique actuel de la recherche. Un exemple : des archives papier qu'il a conservées pendant des années, fruit de ses recherches (et qu'il a fini par jeter), auraient-elles pu connaître une forme d'exploitation quelconque ? Autres exemples sont possibles, il suffit de penser aux différents supports d'enregistrement qui ont défilé des années 1980 à aujourd'hui et qui doivent se trouver à l'abandon dans les laboratoires. Le débat sur les données de la recherche devrait

avoir également un volet portant sur la récupération des données des recherches stockées dans d'autres supports, avant le Web.

LES CHERCHEURS EN SHS DANS LE SILLON DES SCIENCES EXACTES

Il est certain que les SHS se voient aujourd'hui confrontées à des réflexions déjà entamées depuis quelques années dans les sciences exactes sur l'ouverture des données de la recherche. Comme il est souvent rappelé, les sciences mathématiques ont adopté depuis très longtemps le partage des données. Et la culture scientifique des sciences dures s'est construite sur le principe de la vérification ou de la preuve. En SHS, en revanche, les chercheurs ne sont pas encore à l'aise avec l'idée de devoir rendre visible, en parallèle aux données, la méthodologie de leur collecte et/ou production. Aux Etats-Unis, où cela a été imposé par loi, la contrainte finira tout de même par produire des effets bénéfiques sur la transparence des pratiques des chercheurs. Bien qu'il ne soit pas pour la dimension contraignante, il estime que cela pourrait avoir en France un effet positif.

UN POTENTIEL EDITORIAL CERTAIN POUR LES DONNEES SHS MAIS DES DIFFICULTES D'ORDRE DIVERSE A GERER

Actuellement en France une évolution intéressante serait en cours vers des nouvelles formes éditoriales de la recherche scientifique. Il pense notamment à ce qui se passe en Economie, discipline où des revues mettent en place des articles « augmentés », c'est-à-dire, enrichis de données. A ses yeux, ce modèle devra se répandre vers d'autres disciplines.

La particularité des données en SHS risque néanmoins de se heurter à nombre de difficultés comme par exemple :

1) problème de motivation des chercheurs : les exigences épistémologiques comme la gestion des « métadonnées », le travail de contextualisation, etc., étant très chronophages, les chercheurs risquent de se détourner des bénéfices qui en résulteraient pour la recherche;

2) problèmes juridiques : les techniques courantes d'anonymisation sont assez faibles et pas fiables à un cent pour cent. Mais l'anonymisation excessive implique la perte d'information et rend ces données inutilisables.

3) problèmes de déontologie : les chercheurs travaillent le plus souvent sous des « contrats de confiance » passés entre chercheurs et personnes enquêtées. Une mise en ligne des

données en masse nécessiterait des dispositifs juridiques qui mettraient cet équilibre en risque.

4) Les chercheurs en SHS en France se considèrent très souvent «propriétaires des données». Cela provient de la dimension « auteur/littérateur » que possède la recherche en SHS.

Cela n'empêche nombre d'initiatives, comme les blogs « Carnets de la recherche » par exemple, à mi-chemin entre le partage des données et la réflexion sur la production des données. Des projets comme celui de Robert Darnton c'est-à-dire, ce nouvel objet « livre » dont les parties communiquent avec le tout, sont ambitieux mais trouveraient-ils des lecteurs ?

ENTRETIEN N°23 (02/07/2014) – GEOGRAPHIE

Modalité d'entretien

Téléphone

Statut

Chargé de recherche CNRS

Domaines de recherche

Migrations internationales, mobilités et changement urbain au Sénégal ; famille et logement des migrants internationaux en France

Retours/ Positions

Ses travaux de recherche s'appuient principalement sur des données d'enquête quantitatives produites et exploitées dans le cadre d'un partenariat avec l'Institut National d'Etudes Démographiques

Au sein de son laboratoire ils sont plusieurs chercheurs à s'intéresser aux modalités d'exploitation des données produites par les chercheurs du laboratoire.

Depuis 2005, une base de données recense tous les travaux des chercheurs du laboratoire et des quelques chercheurs à l'international travaillant sur la circulation migratoire, avec l'objectif de donner un accès intégral à un maximum de ressources.

La réflexion se transpose actuellement aux « données de la recherche ». L'activité scientifique du laboratoire se développe actuellement dans quatre axes de recherche, réunissant différentes compétences interdisciplinaires. La création d'un cinquième axe de recherche intitulée «axe méthodologique » est en cours de discussion au sein du laboratoire et verra certainement le jour. L'objectif de sa création est d'intégrer à la réflexion méthodologique des chercheurs un volet sur la mutualisation et le partage des données au sein du laboratoire. Des groupes travaillent actuellement pour définir les besoins, les contraintes et les objectifs de ce projet visant, à terme, la mise en place d'un dispositif qui puisse concentrer et fédérer des données hétérogènes : quantitatives, qualitatives, observations de terrain.

La possibilité de développer ce projet est tributaire des nouveaux recrutements d'ingénieurs d'études dans le laboratoire, qui viennent apporter leur assistance aux chercheurs en matière d'architecture de l'information et traitement des données. Actuellement, les groupes travaillent sur les questions suivantes :

- Méthodologie du traitement des données visant le partage : travail sur les métadonnées et le code, charte des bonnes pratiques à destination des chercheurs;

- Gestion des données exigeant des mesures de secret statistique et confidentialité ;

- Stockage et accessibilité de ces données.

Dans un premier temps, la diffusion de ces données se fera uniquement à l'échelle du laboratoire. Cela permettra de maîtriser les éventuels problèmes avant l'élargissement du périmètre de la diffusion de ces données. Le but est, tout de même, d'arriver à une diffusion le plus large possible.

LES CHERCHEURS ONT INTERET DE S'INVESTIR DANS DES PROJETS QUI PERMETTENT DE GERER ET EXPLOITER LEURS DONNEES

Il y a un réel intérêt de développer cette sorte de projet pour, au moins, les raisons suivantes :

- Les chercheurs doivent se sentir concernés par ce projet car mutualiser et échanger des données permet d'avancer plus vite et de créer un esprit de travail collectif parfois rare chez les chercheurs en SHS,

- Les enquêtes et d'autres matériaux produits sont souvent sous-exploités,

- Les chercheurs doivent pouvoir avoir un accès facile aux données qu'ils ont produit. La solution de dépôt des données actuelle, Quetelet, n'est pas satisfaisante, car l'accès aux données se traduit très souvent par des démarches administratives lourdes,

- Il s'agit de créer un outil à l'image de la recherche et, dans le cas de son laboratoire, spécialement conçu pour permettre une réappropriation des données par les chercheurs spécialistes. Déléguer les étapes de réflexion et de travail en amont à des techniciens pourrait avoir comme résultat la création d'un outil inapproprié.

LES NOUVELLES FORMES DE TRAVAIL POURRAIENT TOUTEFOIS ETRE PERÇUES COMME DES CONTRAINTES PAR LES CHERCHEURS

Certains chercheurs peuvent se montrer réticents à réaliser un travail minimum de formalisation de leur méthodologie de collecte des données et du traitement structurant ces données. Rendre compte de ce travail pourra obliger, parfois, à modifier le processus de recherche en amont, intégrant la réflexion des données du départ.

Il faudra également consacrer du temps à ce travail et être plus ouvert à la discussion collective et à l'interaction des équipes dans le processus de recherche, ce qui peut poser des problèmes à certains chercheurs habitués à des formes de travail plus individualistes.

ENTRETIEN N°24 (02/07/2014) – HISTOIRE

Modalité d'entretien

e-mail

Statut

Maître de conférences

Domaines de recherche

Histoire du travail, des conflits sociaux et des techniques

Retours/ Position

Il s'étonne de cette notion - « donnée de la recherche » - qui n'existe ni dans la littérature scientifique ni dans l'épistémologie. En outre, le partage des informations scientifiques existe avant l'avènement d'Internet.

Tout travail de recherche implique évidemment la collecte de données multiples. Dans son cas il s'agit principalement de la prise de notes sur des ouvrages, la consultation/dépouillement d'archives manuscrites ou imprimées, le recueil ou la constitution de données statistiques, variables en fonction de chaque objet de recherche. C'est d'ailleurs l'essence du travail de recherche de transformer ces données éclatées, dispersées, en récit lisible et de donner du sens à ces informations dispersées.

Ce type de « données » reste sa propriété, il peut les échanger avec des collègues de travail mais il ne voit pas l'intérêt de les mettre en ligne. Chaque chercheur fonctionne en relation avec d'autres et en réseaux avec des pairs, des collègues, qui partagent ses préoccupations et ses objets de recherche, et cela aussi depuis longtemps, avant Internet.

Les données brutes peuvent être échangées dans ces micro-communautés pour aider les autres dans leur propre recherche, pour croiser des informations mais, à son avis, elles ne sont ni destinées à la publication ni à devenir publiques.

PUBLIER LES DONNEES NE PRESENTE PAS UN REEL INTERET

A vrai dire, à son sens, les questions d'une diffusion des données semblent se poser davantage pour les sciences dures où les enjeux de vérification des procédures soulèvent des questions plus sensibles qu'en sciences humaines.

Aujourd'hui, en général, le problème central est le manque de temps pour lire toutes les publications disponibles et produites dans chaque champ spécialisé. La démultiplication des revues et plateformes des revues offrent aux chercheurs une abondance de matériaux qu'il est souvent difficile à trier.

En outre, il considère que l'objectif des publications de qualité est précisément de permettre aux chercheurs/lecteurs un accès à des formes abouties de la recherche, résultats d'un processus mené à terme par les chercheurs. Publier des données ou des étapes intermédiaires de la recherche aurait-il donc un sens? Il y a un nombre réduit de lecteurs qui seraient disposés à consulter ces données et remonter le processus de recherche, alors que l'article consiste justement dans l'interprétation et l'explication de ces données. Au moment où les chercheurs manquent de temps pour l'essentiel des lectures dans leur propre domaine, il est peu réaliste d'imaginer qu'ils se précipiteront dans l'exploration des données produites par d'autres chercheurs.

Il ne voit pas non plus un réel intérêt à la mise à disposition ouverte de toutes ces données sur internet, d'autant qu'il existe déjà une grande quantité de données, d'archives et d'imprimés en ligne.

DE LA FOUILLE DES DONNEES POUR EXPLORER DES CORPUS TEXTUELS

Pour lui, les expressions de « data mining » et « fouille de textes » renvoient à des méthodes très diverses. Il y associe néanmoins le traitement classique statistique et quantitatif de grands agrégats d'informations. Comme beaucoup de chercheurs dans son domaine, il lui arrive d'y recourir au besoin. Par exemple, avec des outils comme Frantexte pour explorer la fréquence de tel ou tel mot dans un corpus, mais aussi avec des outils de recherche fournis par Googlebooks pour explorer par mots clés dans la masse des corpus numérisés.

LE NUMERIQUE : UN OUTIL PARMI D'AUTRES

Il est difficile de ne pas tomber dans des lieux communs lorsqu'on doit commenter les changements apportés par le nouvelles TICs. De manière simple, il estime que désormais il n'est plus possible de travailler sans le numérique, mais contrairement aux prophètes et entrepreneurs du numérique qui y voient une « révolution » et une source de discontinuité radicale dans les pratiques de recherche, il y voit plutôt un outil parmi d'autres qui poursuit des trajectoires antérieures. Par ailleurs, ayant commencé la recherche il y a une dizaine d'années, c'est-à-dire au moment de la montée des technologies numériques, les chercheurs de sa génération ne sont sans doute pas les mieux placés pour évoquer comment cela a modifié les pratiques personnelles. Enfin, ce qu'on appelle « technologies numériques» renvoie à une grande diversité d'outils et de dispositifs, il faudrait voir au cas par cas pour chacun d'entre eux.

ENTRETIEN N°25 (02/07/2014) – SCIENCES POLITIQUES

Modalité d'entretien

Face-à-face

Statut

Chargé de recherche CNRS

Domaines de recherche

Le temps dans les sociétés contemporaines

OUVERTURE DES DONNEES ET OPEN DATA

Il considère qu'en matière de politique d'Open Data il y a encore de progrès considérables à faire sur le volet technique. Dans l'état actuel, ces données sont quasi inexploitables et leur utilité pour la recherche est moindre. Le principe d'une ouverture des données publiques est louable, mais il estime que mettre à disposition des données sans en expliciter la méthodologie, sans documenter celle-ci par le contexte précis de leur collecte, abouti au résultat contraire de l'esprit démocratique qui anime l'Open data au départ. En tout cas, aujourd'hui, les chercheurs ne peuvent pas se servir sérieusement de ces données.

D'autre part, l'ouverture des données de la recherche a comme problème principal la question de la confidentialité et de l'anonymat des données. L'anonymisation limite l'usage de ces données dans la recherche et la gestion des autorisations à obtenir auprès des personnes enquêtées est un système assez lourd à gérer.

Au CDSP (Centre de Données Socio-politiques) les versions anonymisées sont mises à disposition des chercheurs ainsi que les versions non anonymisées. Mais pour ces dernières l'accès est restreint, soumis à des conditions d'utilisation, donnant parfois lieu à des frais de paiement.

DIFFERENTES DIMENSIONS POUR LA QUESTION DES DONNEES DE LA RECHERCHE EN SHS

Il identifie trois dimensions distinctes où la question des données évolue différemment et qu'il faudrait prendre en compte pour une compréhension des enjeux actuels pour les SHS:

- Les chercheurs, les institutions de recherche et les programmes de valorisation de la production scientifique ;

- Les bibliothèques et les questions liées au moissonnage et ouverture des métadonnées ;

- Les éditeurs scientifiques qui essaient de se positionner stratégiquement face aux exigences épistémologiques de la « réplicabilité » de l'expérience et qui commencent à exiger (en économie par exemple) les données accompagnant les résultats de la recherche.

A son avis, une ouverture des données de la recherche n'a de sens que si elle vise également le partage des programmes et algorithmes utilisées pour analyser ses données de

manière à permettre à d'autres chercheurs d'en reproduire l'analyse ou de tester ces algorithmes sur leurs propres données.

La compétition étant tellement exacerbée entre les chercheurs dans son domaine, il imagine mal ce partage pendant des recherches en cours, mais pense que les chercheurs seront poussés de plus en plus à travailler leurs données et à les publier en parallèle aux publications, comme l'a fait Piketty, par exemple.

ENTRETIEN N°26 (02/07/2014) – SIC

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Culture de l'information

Retours/ Positions

EN SIC L'OUVERTURE DES DONNEES EST INDISPENSABLE

Cette chercheuse ne travaille pas avec des données quantitatives. Les données produites proviennent des notes de séminaires et de ses lectures.

Elle partage le plus possible et parfois instantanément : dépôt dans les archives ouvertes, partage sur son blog, partage dans le réseau Twitter. Plus récemment, elle essaye l'expérience de mettre en place un pad de prises de notes collectives et des webmaires (séminaires sur le web).

La démarche d'ouverture d'un grand nombre de données lui semble indispensable dans le domaine des SIC. Elle a eu l'occasion d'en voir quelques effets, rapides et persuasifs : effets réseaux qui vont découler en des collaborations scientifiques, échanges de connaissances et circulation des savoirs. Ces bénéfices sont peut-être moins évidents pour les disciplines « classiques » des SHS, bien que visibles en sociologie et en sciences politiques.

L'OUVERTURE DES DONNEES NE DOIT PAS ETRE UNE FIN EN SOI

Pourtant un partage massif des données et une ouverture à tous azimuts n'est peut-être pas souhaitable. Il faut, à son avis, éviter le trompe-l'œil d'un idéal d'ouverture totale et, au contraire, analyser au cas par cas l'intérêt et les bénéfices du partage et de la diffusion des données. Bien qu'elle soit elle-même favorable à la publication d'un maximum des données, il ne faut pas oublier qu'une recherche de qualité se construit sur ses propres datas et sur un rapport particulier du chercheur à ces données dans le processus créatif de la recherche qui peut ressembler, bien des fois, au rapport de l'artiste à son œuvre. Il est donc naturel que le chercheur puisse ne pas souhaiter partager un certain type de données avant de les avoir exploité dans sa recherche.

Par ailleurs, rendre accessible des données implique un travail important de normalisation et contextualisation qui doit être mené collectivement.

Actuellement elle réfléchit beaucoup sur les problèmes d'archivage des données. A son avis, il n'y a pas assez des politiques d'archivage. Il faudrait développer davantage de structures, à l'exemple d'Huma-num et de CINES. Dans son laboratoire, elle mène, auprès des collègues, une campagne pour l'adoption des politiques d'archivage en vue de travailler conjointement les formats des données en garantissant leur pérennité. Mais la question est loin d'être simple et il y a, à son sens, énormément de développements de politiques d'archivage et de sensibilisation à être réalisés en France.

ENTRETIEN N°27 (03/07/2014) – HISTOIRE

Modalité d'entretien

e-mail

Statut

Ingénieur d'études

Domaines de recherche

Histoire médiéval, espaces urbains et culture (XIIIe-XVe siècle), histoire des activités économiques médiévales

Retours/ Positions

« DONNEES BRUTES » A DISTINGUER DES « DONNEES DE LA RECHERCHE »

Dans le cadre de ses recherches en histoire, les données de la recherche sont avant tout des données recueillies en archives.

Il importe toutefois de distinguer la « donnée brute », le texte d'archive lui-même - qu'il qualifierait plutôt de « données de terrain » -, des données de la recherche qui correspondent plus à un travail de pré-analyse.

Il s'agit de constituer un corpus portant sur le thème qu'il souhaite traiter mis sous forme de notes synthétiques ou de bases de données selon la nature de la source envisagée. Les données ainsi produites sont ensuite utilisées dans des publications ou des interventions lors de colloques ou de séminaires. Elles font aussi l'objet d'un partage avec des collègues qui travaillent sur la même problématique.

LE PARTAGE DES DONNEES OUVRE DES NOUVEAUX CHAMPS A LA RECHERCHE

Il partage naturellement ses données avec des collègues. De son point de vue, le travail de recherches est avant tout un travail collectif car il n'est pas possible d'appréhender par le seul travail individuel un ensemble des sources disponibles, surtout quand la problématique ne se focalise pas sur une échelle locale.

Le partage permet également d'envisager des types de sources qu'on ne connaît pas et par là rend possible une réelle ouverture vers d'autres horizons géographiques ou vers d'autres méthodes de travail.

Il y a toutefois un bémol pratique au partage des données : la compétence spécifique de celui qui les a recueillies. C'est, à ses yeux, plus difficile d'utiliser les données d'autre chercheur si celui-ci ne rend pas disponible ou ne communique pas la façon dont elles ont été recueillies et quelles en sont les limites.

DIFFUSION DES DONNEES EN AMONT, EN AVAL OU EN PARALLELE

Par rapport aux publications des chercheurs, le type de diffusion de ces données va dépendre de la configuration de la recherche. Si celle-ci s'intègre à un projet collectif, le partage peut se faire en amont entre les membres du collectif. Par contre, l'ouverture totale

doit, selon lui, se faire en aval de la publication afin de protéger le travail des recherches en cours. Dans certains cas, il peut se faire en même temps que la publication, sous la forme d'un volume de pièces justificatives, par exemple

Le « data mining » et la « fouille de textes » sont des procédés assez peu utilisés en histoire médiévale, bien que le data mining soit employé par les médiévistes anglais pour traiter de l'histoire économique. Il faut dire que les sources anglaises le permettent mais pas les françaises !

Le data mining ne correspond de toute façon pas à sa façon de travailler, plus proche de l'échelle locale et de la micro-histoire. D'autre part, il est nécessaire de préciser que le progrès de la recherche en histoire médiévale est tributaire de l'exploration d'archives inédites, pour laquelle des outils de mining offrent des possibilités exceptionnelles.

Par contre le text mining fait bien partie de ses pratiques, notamment à travers les sites de publications en ligne ou des bibliothèques du type Gallica.

Les technologies du numérique ont transformé les pratiques en SHS à plusieurs égards. Il pense tout d'abord à la mise en ligne d'archives ou de bibliothèques. L'accès à des sources manuscrites inédites conservées en province est désormais largement facilité.

Enfin, un site comme Gallica permet des recherches en plein texte qui modifie complètement la façon de lire ou de rechercher l'information. Il faudrait ajouter l'accès en ligne aux revues, avec là aussi des recherches en plein texte qui permettent d'identifier rapidement ce dont on a besoin.

ENTRETIEN N°28 (03/07/2014) – DEMOGRAPHIE

Modalité d'entretien

Téléphone

Statut

Directrice de recherche

Domaines de recherche

Sida, santé reproductive et rapports de genre, migrants et santé en France.

PARTAGER DES DONNEES HETEROGENES

Actuellement elle analyse des enquêtes qui ont été menées depuis deux ans par des prestataires (institut de sondage Ipsos) auprès des immigrants de l'Afrique Sub-saharienne en Ile de France.

Les données collectées par ces enquêtes quantitatives en face-à-face seront accessibles à la seule équipe pendant 3 ou 4 ans et ensuite déposées dans le réseau Quetelet, comme cela se passe toujours dans son domaine.

Plusieurs équipes appartenant à des organismes divers (CEPED, INSERME, INPES) ont travaillé sur la préparation de ces enquêtes pendant plusieurs mois. Les travaux préparatoires qu'elle a réalisés se basent sur des entretiens qualitatifs selon une approche biographique, le but étant, dans la suite, la constitution d'une enquête pilote.

Elle estime que ces travaux préparatoires sont effectivement très riches et présentent un intérêt à être diffusés, notamment du point de vue méthodologique. Une difficulté se présente, néanmoins, originaire de la particularité de cette étude d'être à cheval sur deux disciplines, la démographie et l'épidémiologie, et des types différents de données utilisées, qualitatives et quantitatives. Il faudrait réfléchir à un moyen de présenter ces données «brutes » en les contextualisant, mais cela implique beaucoup de temps disponible pour leur préparation et les chercheurs ne sont souvent pas en mesure d'accomplir seuls ce travail.

PARTAGER LES METHODOLOGIE S D'ANALYSE DE DONNEES

Ce qui est envisageable, par contre, c'est de réaliser de plateformes de partage des travaux intermédiaires et des méthodologies d'analyse de données appliquées à de cas concrets. Cela serait, à son sens, d'une grande utilité pour d'autres chercheurs et pourrait créer une dynamique de collaboration scientifique.

Pour l'instant, les équipes ont diffusé seulement un « working paper », téléchargeable en libre accès.

ENTRETIEN N°29 (03/07/2014) – HISTOIRE

Modalité d'entretien

Téléphone

Statut

Directeur de recherche

Domaines de recherche

Histoire du droit du travail

Retours/ Positions

UN VASTE PROJET DE GESTION DES DONNEES DE LA RECHERCHE

En ce moment il développe avec l'équipe de son laboratoire (19 personnes), et d'autres centres et laboratoires partenaires, un grand projet sur l'histoire social du droit des colonies.

A l'origine de ce projet, se trouvent des prises de note et 8000 photos des Bulletins Officiels du Travail dans les colonies, fruits de quatre mois de travail aux archives d'Aix-en-Provence.

Une fois de retour à son environnement de travail, l'intérêt d'exploiter ces documents et de les partager avec le laboratoire lui a semblé évident.

Par la même occasion, deux ingénieurs de recherche spécialisés en humanités numériques sont recrutés et mettront en place un outil collaboratif de dépôt et partage des données.

Autour de cette nouvelle possibilité de partage, le projet va se développer et inclure d'autres axes de recherche sur l'histoire du droit des colonies en élargissant son périmètre géographique et en incluant des chercheurs d'autres laboratoires. Des équipes sont constituées et des aires géographiques sont attribuées à chacune. Pendant plusieurs mois ces équipes vont dépouiller et photographier des archives, les documenter (par des notes et métadonnées) et les mutualiser dans la plateforme collaborative.

Les ingénieurs de recherche se chargeront de garantir la qualité des métadonnées descriptives, l'accessibilité des documents et leur diffusion au sein des laboratoires, mais chaque chercheur est libre d'y déposer ses données.

Quelques difficultés apparaissent du côté de certains usagers qui ne sont pas à l'aise avec l'aspect technologique de la plateforme. Un des ingénieurs a réalisé un tutoriel à leur intention mais cela n'a pas été suffisant dans certains cas. Il faudrait prévoir d'autres solutions pour que la technologie ne vienne pas s'interposer entre le chercheur et le but poursuivi, le partage et la documentation des données.

A sa connaissance, plusieurs projets semblables sont en cours à la Maison des Sciences de l'Homme. Leur originalité réside non seulement dans les ressources numérisées qu'y sont partagées mais aussi dans le fait de mettre à disposition les notes et observations des chercheurs.

Il y a donc mutualisation des données «sources» et des données «produites» par les chercheurs.

Toutefois, à ce stade, aucun partage plus large n'est envisagé car il s'agit des données des recherches en cours.

ENTRETIEN N°30 (03/07/2014) – SCIENCES ADMINISTRATIVES

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Luxe, le rapport aux objets

Retours / Positions

UN TRAVAIL D'ANALYSE DES DONNEES COLLECTEES DESTINE A LA PUBLICATION

Elle réalise actuellement une grande collecte de données sur plusieurs aspects de l'industrie du luxe à travers trois procédés :

- Entretiens menées auprès des acteurs de l'industrie du luxe : vendeurs, manageurs, hôtels, restaurants de luxe ;

- Observation non participante dans les lieux du luxe ;

- Observation participante (en tant que cliente) ;

Il s'agit uniquement de données qualitatives, elle ne travaille jamais avec des données quantitatives et ne fait pas du data mining.

Elle n'envisage pas de mettre ses données à disposition car cela représente beaucoup de temps de travail. Elle communique néanmoins ces données dans les colloques et conférences auxquels elle participe. Autrement son travail d'analyse des données est destiné à la publication d'articles dans des revues de gestion ou de sociologie.

ENTRETIEN N°31 (07/07/2014) – ECONOMIE

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Economie financière, microéconomie bancaire, politiques publiques, comportements électoraux et politiques.

Retours / Positions

LES REVUES SCIENTIFIQUES NE S'INTERESSENT QU'AUX RECHERCHES REUSSIES

Ce chercheur développe ses recherches au croisement de deux disciplines : sciences économiques et sciences politiques. Ses recherches actuelles portent sur les systèmes de paiement et il se sert notamment des données produites par l'INSEE, Eurostat et des données des entreprises fabricant des cartes bancaires.

Dans sa recherche, il produit de données quantitatives à partir de sondages et des « questionnaires sur table » (visant une population spécifique). L'étape de production et exploitation des données est toujours orientée par une théorie préalable pour laquelle ces

données constituent les preuves ou la confirmation. Lorsque, au contraire, les données contredisent cette théorie, ou ne fonctionnent pas comme preuve, la recherche en cours est le plus souvent abandonnée ou retravaillée, car elle ne trouvera jamais une forme quelconque de publication. Pourtant les erreurs ou inexactitudes des théories ont souvent une valeur épistémologique très intéressante et il est fréquent que les chercheurs en discutent de façon informelle. Il n'est pourtant pas sûr, et c'est regrettable, que ce type de donnée trouve un jour sa place dans d'autres types de publications.

LES ARTICLES « AUGMENTES » DES DONNEES : UN INTERET CERTAIN POUR LES SHS

Dans le domaine des sciences économiques, la question du partage et de la diffusion des données ne se pose pas aussi fortement qu'en sciences politiques, car dans les premier cas les chercheurs ne sont que très rarement des producteurs de données, alors que dans le second, les chercheurs produisent très souvent de données quantitatives. L'intérêt de trouver des espaces de publication de ces données est certain, comme l'atteste la revue French Politics qui a créé la rubrique « Data, Measure and Methods ».

La diffusion de ses travaux se fait par publication dans des revues « peer-review » et à travers le dépôt des versions pré-print dans SSRN. Pourtant, à son sens, à l'heure d'Internet les revues n'ont plus vraiment de sens et les chercheurs ont intérêt à s'affranchir du système d'évaluation par publication dans les revues « prestigieuses ». La création de plateformes spécialisées donnant un accès libre aux travaux de chercheurs suffirait largement à promouvoir la diffusion des connaissances scientifiques.

De façon analogue, il estime que l'infrastructure centralisant aujourd'hui la gestion et le traitement des données quantitatives, le réseau Quetelet, constitue un frein pour l'innovation en matière de politique des données scientifiques en SHS, du fait du lourd système administratif mis en place pour l'accès à ces données. Il faudrait développer des systèmes alternatifs pour leur diffusion et à son avis cela prendra la forme de l'auto-dépôt des données par les chercheurs, dans des sites et plateformes créés par les communautés scientifiques.

ENTRETIEN N°32 (07/07/2014) – ECONOMIE

Modalité d'entretien

Téléphone

Statut

Maître de conférences en économie

Domaines de recherche

Economie de la santé

Retours / Positions

LES « DONNEES DE LA RECHERCHE » A DISTINGUER DES INFORMATIONS « BRUTES »

Par « données » elle comprend tout d'abord des informations quantitatives, généralement présentées dans un tableur (individus en ligne et variables observées en colonne).

Mais le terme données « de la recherche » sous-entend des données « produites », ou utilisées à titre secondaire, dans un but de production de savoir scientifique. Ces dernières impliquent dès lors des étapes d'analyse et d'interprétation.

UN CYCLE DE VIE « CLASSIQUE » POUR CES DONNEES AU SEIN DE SON LABORATOIRE

Dans le cadre de sa recherche il lui arrive plus fréquemment d'utiliser des données déjà produites que de les produire elle-même.

A propos de l'aspect de production de données quantitatives, elle a contribué à la réalisation de quelques enquêtes dans le champ des comportements de santé. Elle a ainsi participé à la constitution des échantillons et rédaction des questionnaires, constitués de questions fermées et soumis à quelques centaines ou milliers d'individus.

Son laboratoire gère ces données suivant une pratique assez « classique » et qui ne pose pas de problème particulier. De manière générale, suite à la collecte, les données sont stockées par son laboratoire de rattachement et presque immédiatement mises à disposition des chercheurs du laboratoire. La plupart du temps elles serviront à des objectifs de traitement statistique dont les résultats seront publiés ultérieurement dans les revues.

Au bout de quelques années, lorsque les chercheurs ont fini d'exploiter ces données, elles sont mises à la disposition de l'ensemble de la communauté de chercheurs à travers la mise en ligne sur un site spécialisé.

Les motivations qui pourraient amener les chercheurs à diffuser des données sont, à son sens, étroitement dépendantes des conditions de collecte et production de ces données. Tout d'abord, un chercheur travaillant en solo ne trouvera pas autant de soutien financier et logistique pour bien préparer ces données à une publication. Ceci constitue naturellement un frein à des initiatives individuelles. A ses yeux, un programme de diffusion des données a une plus grande probabilité de réussite lorsqu'elle émane d'initiatives collectives, d'abord parce qu'elle bénéficie d'une reconnaissance au sein d'une communauté de chercheurs et peu ainsi obtenir des soutiens institutionnels importants, mais aussi parce que le travail de préparation et gestion de ces données ne peut être réalisé par un seul chercheur.

UNE DIFFUSION DE CES DONNEES EN PARALLELE ET EN AVAL PLUTOT QU'EN AMONT DES PUBLICATIONS

Quoiqu'il en soit de ces conditions, la motivation principale au partage de données serait la mention explicite de la source des données dans les publications, ce qui permet de contribuer à la réputation du chercheur/du laboratoire qui les a produites.

En conséquence au risque de concurrence entre chercheurs et laboratoires, une diffusion en amont des publications semblerait peu probable. Une diffusion parallèle peut être intéressante pour vérifier certains résultats de publications. Une diffusion en aval peut permettre de traiter de problématiques non abordées par les producteurs des données (soit parce qu'ils n'y ont pas pensé, soit parce qu'ils ne souhaitent pas traiter ces questions).

Dans son cas, il ne lui est jamais arrivé de partager ou de diffuser des données à titre individuel mais uniquement en tant que membre du laboratoire. Pareillement, elle a souvent réutilisé des données mises à disposition par des institutions et très rarement par des chercheurs à titre individuel.

Elle a déjà utilisé le data mining mais de façon assez marginale et envisage de recourir à la fouille de texte prochainement. Il est certain que ces outils possèdent leur intérêt mais il faudrait rester assez critique dans leur utilisation. En effet, les capacités de calcul croissantes offertes par l'informatique facilitent les approches empiriques quantitatives au détriment des approches théoriques et qualitatives.

ENTRETIEN N°33 (07/07/2014) – SOCIOLOGIE

Modalité d'entretien

e-mail

Statut

directrice de recherche

Domaines de recherche

La médecine et la santé en sciences humaines et sociales

Retours / Positions

ANONYMISER LES DONNEES : ENJEU CENTRAL POUR FACILITER LEUR DIFFUSION

Dans son processus de recherche elle produit trois types de données :

- des données quantitatives dans un travail par questionnaire,

- des données qualitatives sous forme d'entretiens

- des données qualitatives sous forme de compte-rendu d'observation.

Ces données sont traitées avec beaucoup de précaution car elles ne peuvent circuler sans une complète anonymisation, en particulier les données qualitatives. Elle les exploite seule ou, plus souvent, collectivement, avec une équipe.

DIFFUSER LES DONNEES QUALITATIVES : DES POINTS DE VUE PARTAGES

Une réflexion est menée depuis plusieurs années au sein de son laboratoire sur l'intérêt et les conditions du partage ou d'une plus ample diffusion des données et, particulièrement, des données qualitatives. Pour ces dernières, les avis divisent la communauté des chercheurs, en sociologie en particulier, mais aussi en anthropologie. Certains chercheurs seraient prêts à mettre à disposition ces données sous certaines conditions, d'autres sans doute beaucoup moins.

A son sens, le principal problème se situe en amont, dans la nécessité de travailler la préparation de cette éventuelle diffusion, dès la production des données. Il faudrait, par exemple, être extrêmement scrupuleux concernant les problèmes de confidentialité. Comment rendre publique ou diffuser en ligne des entretiens réalisés avec des personnes et basés uniquement sur des contrats de confiance établis sur le moment ? Il s'agit d'une véritable question. Prendre le choix de ne pas le faire implique sûrement une « perte » de

données. D'autre part, il est impossible de savoir si les recherches sont ou non meilleures grâces à la mise à disposition de ces données.

DIFFUSER LES DONNEES BRUTES MAIS AUSSI LA LITTERATURE « GRISE »

Elle travaille occasionnellement avec des données mises à disposition par d'autres chercheurs mais celles-ci concernent notamment d'enquêtes diffusées sur des sites officiels. Autrement, cela peut arriver dans le cadre de recherches collectives, où les chercheurs mettent en commun toutes les données.

Lorsqu'un chercheur a "terminé" vraiment une recherche et qu'a priori il n'y reviendra plus, il pourrait mettre à disposition ses données pour d'autres chercheurs. Cela est en particulier intéressant pour des travaux de terrain auprès de personnes disparues, par exemple, ou pour travailler de façon comparative sur plusieurs populations.

Un autre point, souvent moins évoqué, concerne les exigences de la littérature scientifique qui font disparaître la littérature « grise ». Celle-ci représente pourtant les « coulisses » de la recherche, où il est possible de trouver des matériaux très riches dans un format différent de celui des revues scientifiques et qu'il serait très intéressant d'exploiter. Toutefois, dans ce cas aussi il faut protéger, non seulement le chercheur d'utilisations détournées de ses matériaux, mais aussi les personnes qui les confient.

Les technologies ont sûrement changé les pratiques en SHS même si elle ne se sent pas à l'aise avec tous les aspects de ces changements. Un des aspects le plus frappant est celui des transformations opérées sur les ressources bibliographiques, du point de vue de la recherche et de l'accès. Devant la richesse d'informations et multiplication des outils, il y a un risque de débordement et d'une très haute formalisation voire standardisation des productions.

ENTRETIEN N°34 (08/07/2014) – SOCIOLOGIE

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

La professionnalisation à l'université, l'accueil des étudiants étrangers à Paris, réformes scolaires

Retours / Positions

Elle participe actuellement à la formalisation des résultats de l'enquête réalisée cette année sur la condition de vie des étudiants étrangers en France. Il s'agit d'une grosse enquête quantitative réalisée à l'initiative d'un organisme public d'étude et de recherche sur la population étudiante. Elle ne participe pas au dépouillement de ces données - ces étapes sont prises en charge par des techniciens statisticiens-, mais uniquement à l'analyse interprétative de ces données et conséquente restitution dans des rapports et synthèses. Ce travail est réalisé collectivement. S'agissant le plus souvent d'enquêtes commandées par le Ministère de l'Education Nationale, les données collectées seront versées ensuite dans une base de données intégrant le Réseau Quetelet et deviendront accessibles selon les conditions de ce dernier.

LES METHODOLOGIES SONT PARTAGEES ENTRE SOCIOLOGUES, MAIS PAS FORCEMENT D'AUTRES TYPES DE DONNEES

Mais la plupart du temps elle travaille sur des enquêtes qualitatives et communique très souvent sa démarche méthodologique dans les colloques, séminaires et publications.

Dans son domaine, toutes les données ayant valeur de justification méthodologique (constitution des échantillons, réponses aux questionnaires, etc.) sont naturellement ouvertes mais le plus souvent en parallèle ou après la formalisation des résultats. Cela constitue, depuis longtemps, le mode de travail des sociologues et est une exigence épistémologique de la sociologie.

Il est rare, toutefois, qu'elle ou ses collègues partagent d'autres types de données issues d'étapes plus réflexives de la recherche ou des données collectées par le chercheur à titre individuel sauf, bien entendu, dans un contexte de projet collectif.

Mais le partage des données selon d'autres formes, moins formelles, existe : de chercheur à chercheur ou de chercheur à étudiant, à l'occasion d'échanges dans les locaux de l'université. Si un chercheur est sollicité par d'autres chercheurs à partager certaines données, il le fera sans trop de difficultés, mais il le fera difficilement de sa propre initiative et sans un objectif très précis.

En effet, de ce point de vue, elle observe que les nouvelles technologies n'ont pas changé beaucoup les modes de travail des chercheurs, assez individualistes encore, ainsi que leur rapport aux données produites par leurs recherches, qui restent très majoritairement leur propriété.

LES TECHNOLOGIES NE REMPLACENT PAS LE TRAVAIL De TERRAIN DU SOCIOLOGUE

Le text mining est un procédé assez courant parmi ses étudiants mais malgré son existence ancienne et les développements d'outils assez performants dans l'actualité, ce procédé n'est pas probant d'un point de vue épistémologique. Ses résultats restent le plus souvent largement contestables. A son sens, aucune performance technologique ne vaut l'observation et l'analyse empirique réalisé dans une approche qualitative.

ENTRETIEN N°35 (09/07/2014) – ECONOMIE

Modalité d'entretien

Téléphone

Statut

chargé de recherche

Domaines de recherche

Economie mathématique

Retours / Positions

Chercheur en économie, ses recherches portent sur les aspects théoriques des décisions (sociales ou individuelles) mais également sur l'économie de la santé et l'épidémiologie. Plus récemment il s'est intéressé au protocole Bitcoin d'un point de vue micro-économique.

Il n'utilise aucune donnée du type « chiffre », mais seulement des articles, du code (via Github) et des programmes informatiques.

Dans sa recherche il produit du langage mathématique et informatique, de textes et d'analyses visant la publication.

Il ne partage que les codes et des bouts de code, comme le font d'ailleurs tous qui s'en servent de Github.

Parmi ses collègues, l'habitude du partage n'est pas non plus monnaie courante. Dans la pratique, les chercheurs ne proposent pas d'eux-mêmes leurs données, mais ils peuvent les mettre à disposition suite à la demande d'un collègue. En tout cas il n'existe pas de partage collectif.

LE DESEQUILIBRE DES MODELES DE AUTO-ARCHIVAGE ET DE PUBLICATION EN REVUES SPECIALISEES

L'enjeu plus important dans son domaine réside, à son avis, dans la diffusion des résultats des recherches et dans un déséquilibre de deux modèles : celui de diffusion des working paper et de la publication dans les revues à renom.

D'une part, dans son laboratoire les chercheurs déposent leurs working papers dans des plateformes comme SSRN et IDEAS. C'est une initiative positive de faire circuler les savoirs mais qui comporte un certain nombre d'inconvénients, le plus grand étant l'absence de processus de peer-review de ces articles. Il lui est déjà arrivé plus d'une fois de consulter des articles de qualité discutable dont les développements étaient insuffisamment fondés. Un autre problème est la difficulté de savoir quelle version a été consulté, si elle correspond à la dernière déposée ou à la même consulté il y a un an, par exemple.

D'autre côte, les revues d'économie qui garantissent la qualité de la ressource et le renom du chercheur, en majorité anglophones, publient sous un modèle économique qui ne va pas sans être abusif. Par exemple, la simple soumission d'un article accompagne presque toujours le paiement d'une « submission fee ». Le lecteur/chercheur, de son côté, devra aussi payer pour en avoir accès.

A son avis, dans ce dernier cas, il s'agit d'autre chose que de la pure diffusion de l'information scientifique. Il s'agit de payer un signalement et un certificat de qualité.

Dans son domaine il est extrêmement difficile d'échapper à ces deux extrêmes car leur travail est mesuré et évalué par les publications.

Devant cette situation les chercheurs sont résignés. Il y a tout de même des initiatives intéressantes comme « Economics Bulletin », revue Open Access à comité scientifique qui publie de notes de recherche, commentaires et premiers résultats des recherches en cours.

LE TDM COURANT EN ECONOMETRIE

Il s'est déjà servi du data mining pour collecter des données publiques, mais cela reste très occasionnel. Ces pratiques sont, toutefois, très courantes dans le domaine de l'économétrie.

ENTRETIEN N°36 (09/07/2014) – HISTOIRE

Modalité d'entretien

Téléphone

Statut

docteur/chercheur indépendant

Domaines de recherche

Les discours sur l'égalité/inégalité des femmes et des hommes

Retours / Positions

QUELQUES HISTORIENS A LA TETE DES INITIATIVES DU TEI (TEXT ENCODING INITIATIVE) DANS LES ANNEES 90

Elle se considère un cas particulier dans son domaine. Au début des années 90, avant que les Humanités numériques fassent leur apparition en France, on parlait de « Histoire et informatique ». Il s'agissait d'un nombre réduit de chercheurs qui, opposés à la séparation chercheurs/techniciens, ont ressenti le besoin de collecter, transcrire (réaliser la paléographie) et analyser leur propres données, prenant ainsi en main, à ce moment-là, les possibilités offertes par les outils informatiques. Ce besoin se traduisit par des réflexions collectives et groupes de travail cherchant des solutions d'exploitation des données à court et long terme.

Dans le cas de sa propre recherche, la problématique était de trouver une solution de travail et exploitation des données hétérogènes collectés à travers la paléographie de manuscrits datant de la période comprise entre la Renaissance et le XVIIe siècle. L'écriture de 1580 et celle de 1640 étaient complétement différentes et techniquement c'était très compliqué de travailler avec ses sources. Dans le but de trouver une solution technique à l'exploitation de ces données hétérogènes, elle s'est rapprochée à des chercheurs travaillant sur la TEI (Text Encoding Initiative) à l'INRIA et s'est auto-formée en encodage en XML.

L'HISTORIEN DOIT TRAITER LES DONNEES SOI-MEME POUR S'APPROPRIER SON OBJET DE RECHERCHE

Elle se considère un cas à part dans sa discipline, comme l'est aussi celui de Robert Descimon avec qui elle partage plusieurs points de vue épistémologiques. Par exemple, pour tous les deux, le traitement de données est partie intégrante du processus de recherche. Collecter, transcrire et encoder ses propres données permet de s'imprégner et de bâtir un objet de recherche tout autrement que faire traiter ses données par quelqu'un d'autre. Pour cette raison, elle ne peut pas, en tant que chercheuse, travailler avec les données produites par d'autres personnes.

A son sens, aujourd'hui la séparation chercheur/technicien n'a pas été dépassée, contrairement à ce qu'on aurait pu attendre compte tenu des progrès technologiques et des logiciels plus faciles à prendre en main par des non informaticiens.

Les historiens, par exemple, estiment qu'ils ne doivent pas être des paléographes et s'occupent quasi exclusivement du travail interprétatif de l'analyse des données. En faisant cela ils nourrissent cette représentation désuète qui sépare théorie et technique dans les SHS et qui « déclasse » cette dernière au rang d'une activité sans valeur scientifique à part entière.

Les chercheurs impliqués dans les Humanités Numériques sont intéressés par toutes ces questions mais l'abordent, là encore, d'une façon purement théorique. Les trois quart ne font pas et ne souhaitent peut-être pas réaliser eux-mêmes le travail sur les données. Cet état des choses est indépendant du fait de travailler ou non avec des données quantitatives, même s'il est vrai que la « quantitativité » des années 70 a laissé des souvenirs négatifs en histoire et que beaucoup d'historiens réagissent contre une quantification de l'histoire (l'approche de la micro-histoire, par exemple).

La tendance aujourd'hui est donc de s'entourer d'ingénieurs ayant des compétences techniques et technologiques et qui réalisent le travail sur les données de la recherche.

Pour elle, en revanche, l'outil informatique et, maintenant, les technologies numériques, sont, très concrètement, des instruments du chercheur et l'environnement de sa recherche. Mais ces instruments ne sont pas neutres et leur action ne s'exerce pas sur une matière passive. Les instruments sont des extensions du travail intellectuel du chercheur. Comment alors séparer ces deux moments, préparation des données et interprétation de celles-ci ?

LES DONNEES AINSI QUE LA METHODOLOGIE DE COLLECTE ET D'ANALYSE DOIVENT ETRE DIFFUSEES

En histoire, la plupart des chercheurs ne souhaitent pas dévoiler leur méthodologie. À son avis, ceci est un contresens et totalement contradictoire à la démarche scientifique. Dans ces sens, elle pense que les données doivent pouvoir être mises à disposition ainsi que la méthode de collecte et d'analyse de ces données. Pour évoquer un exemple très récent, il suffit de se rappeler les attaques subies par Piketty lors de la parution de son dernier ouvrage. Comment aurait-il pu les contredire, s'il n'avait pas, auparavant, partagé les données au public ?

Elle a pris il y a bien longtemps le parti de diffuser et les données et la méthode qui a servi à collecter et analyser ces données. Cette diffusion a lieu sur son site personnel et après la publication des résultats.

LE WEB SEMANTIQUE NE VA PAS REVOLUTIONNER LES SHS

Lorsqu'on évoque les principaux impacts des technologies du web sur les SHS on évoque souvent le Web sémantique et son soi-disant caractère révolutionnaire. Elle pense que le web sémantique, malgré l'intérêt qu'il comporte, est de l'ordre d'un idéal qui va bientôt déchanter ces adeptes. Ces technologies ont des limites ainsi comme, aujourd'hui on le sait, les technologies d'OCR dont on évoquait il y a quelques années le caractère révolutionnaire.

ENTRETIEN N°37 (03/07/2014) – DROIT

Modalité d'entretien

Téléphone

Statut

maître de conférence

Domaines de recherche

Droit des médias

Retours / Positions

DES DONNEES CLASSIQUES EN DROIT ET UNE DIFFUSION DES RESULTATS DE LA RECHERCHE CHEZ DES EDITEURS HISTORIQUEMENT PLACES

Il développe des recherches dans le domaine du droit des médias. En tant que chercheur il se sert des données assez classiques, textes de droit, réglementations, lois. L'organisation et la classification de ces données est un fait historique de longue data et dont la réflexion appartient au cœur même du système juridique. Les bases de données juridiques se sont ainsi très tôt développées en France et les textes sont aujourd'hui facilement accessibles. Actuellement, avec l'Open Data d'autres sources d'informations utiles aux juristes sont également accessibles.

Dans son domaine la diffusion des résultats de recherche se fait presque exclusivement par la publication d'articles et ouvrages chez des éditeurs historiquement placés dans cette discipline. Il y a très peu de publications en revues open access, d'une part parce que les auteurs/chercheurs en droit sont rémunérées par ces éditeurs spécialisés, d'autre part par ce que ces revues sont peer-reviewed et offrent une garantie de qualité dans les processus d'évaluation scientifique.

EN SCIENCES DURES, LA QUETE D'UN CADRE JURIDIQUE TRANSPARENT

Il a récemment participé à des réunions avec des chercheurs en sciences exactes pour apporter son expertise sur un nombre de questions et difficultés touchant les données de la recherche. Le problème principal se place du côté du droit de la propriété intellectuelle.

Les chercheurs souhaitent pouvoir trancher sur des questions juridiques touchant la réutilisation des données, en tant que producteurs et en tant qu'utilisateurs. Pour faire la part de ces questions, une distinction importante doit être réalisée : les données brutes et les données ayant subi un traitement (éditorial dans une publication, intégration en bases de données). Les premières sont des données libres de droit et pourraient être utilisées par les chercheurs sans demande d'autorisation. La difficulté réside dans la définition des « données brutes », car dès que ses données subissent un traitement quelconque elles ne seraient plus tout à fait brutes.

DEVELOPPER DES PRATIQUES BALISEES PAR DES CONTRATS OU DES LICENCES AUTOUR DU DROIT D'AUTEUR EN SHS

En SHS des problèmes proches se posent, notamment dans les négociations de Couperin et les grands éditeurs. Il estime que des enjeux importants pour la recherche gravitent autour du droit d'auteur qui protège les œuvres mais aussi les données en SHS. Il ne faudrait pas espérer, dans un premier moment, de modifier la loi en France, ce qui peut s'avérer extrêmement long, mais plutôt développer des pratiques institutionnelles, balisées par des contrats ou des licences, dans la quête d'un compromis permettant aux chercheurs de réutiliser des données protégées et de partager, s'ils le souhaitent, leurs données dans des conditions choisies par eux.

ENTRETIEN N°38 (09/07/2014) – SCIENCES POLITIQUES

Modalité d'entretien

e-mail

Statut

Chercheur en contrat doctoral

Domaines de recherche

Homogamie : approches temporelles, longitudinales et comparatives d'un sujet classique.

Retours / Positions

LES DONNEES INSTITUTIONNELLES COMME MATERIAU DE LA RECHERCHE

D'un point de vue très précis, il ne produit pas de données mais utilise des bases de données fournies par l'Insee, l'Ined ou Eurostat, soit en accès libre, soit sur demande (par exemple via le réseau Quételet).

Au sens large, il produit des données en analysant ces bases brutes, par exemple, en calculant des tableaux. Ces résultats sont publiés dans des articles scientifiques mais il

compte aussi de mettre à disposition sur sa page personnelle des tableaux bruts qui permettent à d'autres personnes de reproduire ses analyses.

METTRE A L'EPREUVE DES RESULTATS ET DE PROLONGER DES ANALYSES

À son sens la diffusion est utile pour renforcer des analyses publiées en permettant à d'autres personnes de mettre à l'épreuve ses résultats.

La publication devrait aussi permettre à un autre chercheur de prolonger une analyse déjà publiée sans avoir à repartir de zéro. Mais les bases de données brutes ne peuvent généralement pas être rendues publiques puisqu'elles ne lui appartiennent pas, seulement des traitements plus limités (comme des tableaux).

Il n'a jamais eu l'occasion de réutiliser des données publiées par un autre chercheur, personne dans son domaine n'ayant mis à disposition un tel matériau. En revanche, il a souvent comparé ses résultats avec ceux publiés dans des articles à partir des mêmes bases de données.

DIFFUSER DES DONNEES EN PARALLELE AUX PUBLICATIONS EST PLUS INTERESSANT

Une diffusion potentiellement intéressante pour ces données, à son avis, se ferait en parallèle ou rapidement après la publication d'articles et ouvrages car une fois que le chercheur change de projet, il est difficile de s'y remettre et de préparer des données correctement.

Avant, les données risquent d'être encore travaillées par les chercheur et, en outre, rendre public un travail sans l'avoir publié pourrait permettre à quelqu'un de doubler la personne qui a fait le travail avant qu'elle en retire le crédit qui lui est dû.

De même, pour les utilisateurs, il semble plus intéressant d'avoir accès aux données en même temps qu'à la publication associée.

UNE FOUILLE DE TEXTES « CLASSIQUE » ET UN ENVIRONNEMENT DE TRAVAIL DOMINE PAR LE NUMERIQUE

Dans ses recherches, il peut occasionnellement utiliser la fouille de textes, mais plus au sens de l'analyse textuelle «classique », sur des corpus de taille raisonnable (de quelques dizaines

à quelques milliers d'articles de presse) et d'origine connue (journaux, articles de presse), que sur des corpus gigantesques récupérés sur le Web.

Arrivé trop récemment dans la recherche, il estime qu'il est difficile de parler d'une d'évolution ou d'une transformation des pratiques de recherche en SHS, mais l'omniprésence du numérique dans sa propre pratique de recherche est un constat. Il recherche et lit des articles scientifiques presque exclusivement sur ordinateur, gère sa bibliographie avec Zotero, fais des traitements statistiques sous R...

ENTRETIEN N°39 (10/07/2014) – ARTS/CINEMA

Modalité d'entretien

Face-à-face

Statut

Chercheur/maître de conférence

Domaines de recherche

Esthétique du cinéma et de l'audiovisuel, théorie des genres documentaires

Retours / Positions

DES SOURCES POUR LA RECHERCHE EN ESTHETIQUE DU CINEMA ET UNE DIFFUSION DANS LA PRATIQUE DE L'ENSEIGNEMENT

Actuellement il travaille et développe des études sur le genre documentaire de l'audiovisuel.

Il se sert notamment des données statistiques du cinéma, des données bibliographiques, de la littérature académique, des critiques et forums sur internet.

Naturellement, il se sert aussi des films et d'audiovisuels qui sont relativement accessibles, quelques-uns enregistrés à la tv, autres récupérés sur Internet dans des sites peer-to-peer ou disponibles dans la Cinémathèque et à l'INA.

Du point de vue de sa production, il produit des données bibliographiques, des articles, ouvrages, contributions diverses et des séries d'entretiens filmés. Ces derniers sont en général produits dans le cadre des travaux de recherche collectifs avec des sociologues.

La diffusion des étapes des recherches en cours se fait couramment dans les séminaires de master. C'est une occasion pour lui d'approfondir ses réflexions et d'avoir des retours de ses étudiants et collègues.

DES MODES DE PARTAGE TRADITIONNELS POUR DES RECHERCHES A CARACTERE TRES LITTERAIRE

Autrement, dans son domaine, les chercheurs partagent des réflexions sur leurs travaux récents ou en cours dans les colloques et journées d'études. Il considère que ces activités sont bien plus que des simples formalismes académiques mais constituent une véritable occasion de rencontrer et échanger des informations et, très souvent, de démarrer des projets collectifs.

Les chercheurs en esthétique ne partagent pas leurs données autrement que dans les formes déjà mentionnées, probablement parce que leur manière de procéder est très littéraire et comporte une bonne partie de travail intellectuel individuel. Mais aussi par ce qu'une « vie de laboratoire » authentique fait défaut, il n'y a pas des locaux pour ainsi dire, institutionnalisés, qui favorisent un échange plus dynamique entre les chercheurs. Malgré cela, il pense assister à une évolution progressive vers des formes plus collectives de la recherche.

Du côté des publications, les actes de ces colloques jouent un rôle fondamental dans son domaine, constituant une partie importante des ressources utilisées par les chercheurs.

Il publie dans des revues spécifiques au domaine du cinéma et parfois dans des revues de critique de cinéma.

UNE COLLECTE MANUELLE DE DONNEES DU WEB

A sa connaissance, certains chercheurs de l'IRCAV utilisent des outils de collecte automatisée de données sur Internet. Il s'agit d'une étude croisée de la réception des films à travers collecte des données des avis des internautes et de la critique sur Internet. Dans son cas, il n'utilise pas ce genre de technologies, sa collecte dans les forums sur Internet est souvent très ciblé sur certains films ou documentaires et pratiquée manuellement.

ENTRETIEN N°40 (10/07/2014) – SCIENCES DE L'EDUCATION

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Informatique pour tous, usages de l'informatique dans les écoles, didactique de l'informatique et des disciplines informatisées

Retours / Positions

DU QUALITATIF ET DU QUANTITATIF POUR DES ANALYSES QUALITATIVES SUR L'USAGE ET LES OPPORTUNITES DES TICS EN MILIEU SCOLAIRE

Elle a une première formation en Informatique et s'est ensuite orientée vers les Sciences de l'Education. Actuellement ses recherches portent sur les usages des nouvelles technologies dans le milieu scolaire (collège, lycée) et les MOOCs dans le milieu universitaire.

En ce sens, ses recherches ont une double approche : observation des pratiques en milieu scolaire et universitaire et un travail d'expertise à travers l'interrogation des enseignants sur les besoins et les retours d'expérience de l'utilisation de ces TICs.

Sa méthode de collecte est exploratoire, c'est-à-dire, réalisation d'entretiens qualitatifs sans hypothèse de départ, analyse de ces entretiens, puis dégagement de questions dans l'objectif de réaliser une enquête plus large.

Pourtant, la démarche reste toujours qualitative, le but étant de tester les questions qui ont pointé au départ et les réadapter si nécessaire. L'analyse ultime vise à identifier les usages et les opportunités des TICs dans l'enseignement.

Ces données sont des enregistrements qui seront, dans un premier moment, transcrits intégralement sans rien omettre ou changer (pas d'anonymisation). Cette version est conservée et accessible à deux ou trois personnes impliquées dans le projet. Une deuxième version est anonymisée et diffusée dans un réseau court de chercheurs.

PRENDRE DES DECISIONS EN CONNAISSANCE DE CAUSE GRACE AU PARTAGE DES METHODOLOGIES ET DES DONNEES DOCUMENTEES

En 2013 elle a participé à la conception et conduite du concours Castor visant à la fois à faire découvrir aux jeunes collégiens et lycéens l'informatique et le numérique et à analyser les pratiques de ces jeunes. Les données collectées concernent un certain nombre de données personnelles (âge, établis, genre, classe) mais aussi les traces informatiques laissées par ces élèves (nombre de clics, clics retour etc.). Plus de 170.000 candidats ont participé, répondant 18 questions et produisant une grande volumétrie des données « traces ». Du point de vue technique le traitement de ces données est de l'ordre du Big data. Ces données mises en base de données doivent être converties en tables analysables par des logiciels de traitements statistiques. La question qui se pose est : quelle table réaliser ? Quelles variables choisir ? C'est à ce moment que l'intérêt de travailler en collaboration avec d'autres chercheurs devient patent, et qu'avoir documenté ces choix et partagé les méthodologies permet de prendre de décisions en concertation. Il est donc extrêmement important de mettre en place non seulement un dispositif de partage mais des programmes permettant aux chercheurs de communiquer entre eux et de travailler leurs données de manière homogène.

SENSIBILISER LES CHERCHEURS A L'IMPORTANCE DE CULTIVER LEURS DONNEES

D'ailleurs elle n'est pas la seule dans son laboratoire à avoir cet avis. Actuellement, au sein du laboratoire, un projet de sensibilisation des collègues à la question du traitement des données est développé : étapes et techniques d'anonymisation, contextualisation des données, enjeux liés au partage et à la méthodologie. Ce projet a un caractère de «recommandation » non coercitive.

La réception des chercheurs à ce projet de mutualisation et culture des données est hésitante : d'un côté ils comprennent que cela représente une charge supplémentaire de travail et se demandent si cela en vaut la peine. D'autre côté, ils commencent à se rendre

compte que de plus en plus de revues demandent les données et la méthodologie, ne se contentant plus du seul article, en particulier pour les résultats de recherches basées sur des données du web. Alors s'ils veulent continuer à publier, ils devront progressivement préparer leurs données en amont.

EN HISTOIRE, LA CRAINTE D'UNE « REIFICATION » DES DONNEES…

La question de dévoiler la méthodologie pose aussi d'autres types de problèmes. A ce propos, elle pense particulièrement à un collègue historien, très imperméable à l'idée de diffuser sa méthodologie par crainte, disait-il, d'un « réification » des données de l'histoire ». Cette position souligne, à son sens, l'opposition entre les disciplines à pratiques plus objectivistes, pour lesquelles l'idée de «reproductibilité» de l'expérience n'est pas contradictoire avec leur démarche, et les disciplines à une méthode interprétative, comme l'histoire. »

…mais de multiples intérêts et bénéfices pour la recherche en général

En définitive, pour elle, l'intérêt des chercheurs à traiter leurs données et leur méthodologie en vue du partage est multiple :

- le partage des données rend possible une plus grande collaboration entre chercheurs ;

- les méthodes peuvent être confrontées et réutilisées, ainsi que les données ;

- une autre manière de faire de la science, plus transparente et collaborative, peut voir le jour ;

- les chercheurs pourront partager aussi les outils numériques utilisés pour traiter ces données.

DES EXPERIENCES-PILOTE A TAILLE REDUITE AU DEPART

Il est tôt, estime-t-elle, pour répondre d'une ouverture du type Open data, qui permettrait aux citoyens de remonter des données produites par les chercheurs vers les politiques et réformes politiques. Mais l'enjeu, à son avis, est de cette taille.

Pour l'instant la priorité est d'œuvrer vers un changement des pratiques de chercheurs en matière de préparation de ces données et de les sensibiliser à l'intérêt de le faire. Cela n'est

pas possible sans la création d'expériences-pilotes qui seront forcément d'une taille réduite au départ et d'une ouverture limitée aux chercheurs du laboratoire.

ENTRETIEN N°41 (10/07/2014) – SCIENCES ADMINISTRATIVES

Modalité d'entretien

Face-à-face

Statut

Chercheur/maître de conférence

Domaines de recherche

Plans sociaux des entreprises, syndicalisme, négociation en entreprise

Retours / Positions

LES DONNEES DES STATISTIQUES PUBLIQUES: DES DONNEES MACROECONOMIQUES

Il utilise des données bibliographiques, des données de statistiques publiques et des données d'entreprise. Ces dernières, sont toujours produites par lui et par des collègues.

Les statistiques publiques, essentiellement des données du type « macroéconomiques », ne lui sont pas vraiment d'une grande aide, car il a besoin de données d'entreprises à granularité assez fine pour réaliser des recherches sur les pratiques en entreprise. Les données mises à disposition par l'INSEE sont donc pour lui des « données secondaires ». INSEE détient d'autres données d'entreprises qui lui seraient fort utiles mais auxquelles il ne peut accéder facilement soit par des raisons de confidentialité, soit par l'obligation de passer par des formalités administratives longues et décourageantes.

DES MICRO-DONNEES D'ENTREPRISES SANS VOCATION A ETRE DIFFUSEES

Au égard de ce contexte, il est obligé de produire ses propres données à travers un travail de terrain en entreprise, c'est-à-dire, réalisation d'enquêtes quantitatives et qualitatives.

Pour ce faire, un travail de création de partenaires est essentiel, avec les entreprises bien évidement, mais aussi avec des cabinets de conseil.

En ce moment il travaille essentiellement sur les plans sociaux en entreprise via les procédés sus mentionnés. Les données collectées sont confidentielles, il ne peut les partager qu'avec les chercheurs qui travaillent sur le même projet. Ils créent ainsi des bases de données simples sur Excel, exploitent ces données le temps de la recherche et ensuite stockent localement ces données. Aucun système de réutilisation de ces données n'est envisagé car la grande difficulté est d'instaurer un système qui respecte la confidentialité. L'anonymisation excessive rendrait, d'un autre côté, ces données inexploitables.

UN ENJEU ESSENTIEL : FACILITER L'ACCES A DES DONNEES STATISTIQUES D'ENTREPRISES A GRANULARITE PLUS FINE

La situation des chercheurs travaillant avec des micro-données d'entreprises n'est donc actuellement pas satisfaisante. Il estime que l'accès à un certain type de données retenues par INSEE pourrait permettre d'avancer plus vite dans son travail de recherche. Actuellement son travail de collecte et de traitement des données prend facilement 80 % du temps de son activité scientifique. Ce temps est partagé entre la conduite d'entretiens sur le terrain et des procédures pour accéder à des micro-données concernant les entreprises qui sont éparpillées dans les unités territoriales. Là, aussi, aucun projet de centralisation de ces données au Ministère du Travail n'est en cours et cela est vraiment problématique. Il y a surement une sous-exploitation des statistiques publiques. Le grand enjeu est alors de rendre accessible plus facilement ces données aux chercheurs.

Dans son domaine, la pratique courante de diffusion des résultats est la publication d'articles dans des revues académiques de gestion et très rarement d'ouvrages. Quelques articles paraissent aussi dans la presse ou dans des quotidiens comme « Liaisons Sociales » ou «Entreprises et Carrières ».

ENTRETIEN N°42 (11/07/2014) – ECONOMIE

Modalité d'entretien

Face-à-face

Statut

Chercheur

Domaines de recherche

Économie du développement, emploi informel

Retours / Positions

CROISER DES DIFFERENTS TYPES DE DONNEES POUR L'ETUDE SUR L'EMPLOI INFORMEL

Retraité depuis peu, il continue son travail de chercheur. Ses recherches portent surtout sur l'économie de l'emploi informel en Afrique du nord et en Amérique Latine.

La question des données a toujours été au cœur de son activité de chercheur car la spécificité de son domaine de recherche, l'emploi informel, exige de croiser plusieurs types de données, ne pouvant pas se limiter aux données issues de la statistique publique des institutions comme Eurostat, INSEE ou INST (Tunisie) qui n'ont pas la granularité fine nécessaire à l'appréhension des activités informelles des ménages. En outre, les micro-données détenues par ces institutions concernant les petites entreprises ou les entreprises familiales sont souvent d'un accès compliqué.

Il a été mené ainsi à réaliser des entretiens qualitatifs et à construire des questionnaires visant des enquêtes plus larges. Les données collectées ont été mises dans une base de données personnelle, mais jamais vraiment partagées en vue d'une réutilisation.

DES ENJEUX IMPORTANTS POUR LES DONNEES A PLUSIEURS NIVEAUX

Il voit actuellement le sujet des données de la recherche selon plusieurs perspectives plus ou moins problématiques :

- D'une part l'INSEE et d'autres organismes de statistique publique devraient réaliser un immense travail de rapprochement de leurs données et des catégories socio-professionnelles aux standards internationaux ;

- Des bases de données accessibles, créés par des initiatives de chercheurs et institutions dans son domaine, seraient souhaitables, mais actuellement la compétition entre ces acteurs est grandissante ;

- Les économistes publient majoritairement en anglais dans les revues anglo-saxonnes ; or, celles-ci, à sa connaissance, commencent à exiger à présent que l'article soit

accompagné des données. Ceci annonce sûrement des changements importants à venir pour l'édition scientifique dans les autres disciplines des SHS.

- Il retient, en tout cas, qu'en Sciences Sociales et en Economie la notion de la « réplicabilité » de l'expérience est fondamentale. Dans ce sens il ne suffit pas de mettre à disposition les seules « données brutes », mais il faut aussi rendre disponible la méthodologie documentée de leur collecte en parallèle aux résultats de la recherche qui sont un travail d'interprétation (livre, article).

- Et, enfin, l'idée de multiplier les types de diffusion des données peut avoir un effet positif de transparence et de reconnaissance des « coulisses » de la recherche. Il ne faut pas oublier qu'un travail de terrain est un travail collectif. Ainsi par exemple, le rôle du guide qui a conduit l'anthropologue au terrain de sa recherche ou celui de l'interprète qui a permis les entretiens reste très souvent à l'ombre. Et que dire alors de celui des statisticiens ? Il suffit de se rappeler de l'ouvrage de François Furet et de Jacques Ozouf, Lire et écrire, publié en deux tomes, dont le second comportait la méthodologie et le travail sur les données des statisticiens alors que le tome I faisait état des résultats interprétatifs de Furet et Ozouf. Or, le tome I a été un succès de vente alors que le tome II restait sur les étagères.

ENTRETIEN N°43 (11/07/2014) – LINGUISTIQUE

Modalité d'entretien

Téléphone

Statut

Chercheur/professeur d'université

Domaines de recherche

Bilinguisme, langues et migration, politiques linguistiques et éducatives.

Retours / Positions

DES DONNEES DE TERRAIN ET D'ANALYSE

Elle travaille actuellement sur un projet portant sur le rapport dynamique entre les langues et la migration en France. Ce projet, collectif, implique la participation de linguistes,

ethnolinguistes, sociologues et anthropologues. Pour les besoins de sa recherche elle produit trois types de données :

- Enregistrements d'observations participantes ;
- Entretiens enregistrés ;
- Analyse des interactions suite à la transcription des enregistrements.

DONNEES PARTAGEES EN DEHORS D'UN PROGRAMME FORMEL DE PARTAGE

Les données sont partagées entre les membres travaillant sur le même sujet et sont aussi à disposition des autres membres travaillant sur des sujets divers, mais ne sont pas objet d'un programme formel de partage. Celui-ci se fait en fonction de besoins, de la demande de certains chercheurs ou dans le cadre des colloques et journée d'études.

Une diffusion plus large est difficile à envisager à défaut d'un travail collectif de traitement des données qualitatives en vue de respecter le droit à la vie privé et à la confidentialité des personnes interviewées.

UNE GESTION DES QUESTIONS DEONTOLOGIQUES ET ETHIQUES AVANT UN PROGRAMME FORMEL DE PARTAGE

Il y a un intérêt certain à engager des programmes allant dans ce sens, à condition de bien définir au départ les objectifs d'une telle diffusion et de fixer des règles de déontologie de la recherche pour la réutilisation de ces données. Les jeunes chercheurs, surtout ceux qui ont débuté leur activité de recherche déjà dans l'environnement numérique, sont plus sensibles à ces questions et ont des propositions plus militantes vers une ouverture des données. Mais cela ne peut pas se faire, à son avis, sans une véritable gestion de toute problématique éthique qui en découle.

ENTRETIEN N°44 (16/07/2014) – ARTS/PHOTOGRAPHIE

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Esthétique de la photographie

Retours / Positions

DEUX TYPES DE DONNEES : MATERIAUX DE LA RECHERCHE ET DONNEES ACADEMIQUES

Pour elle il y a toute de suite deux types de données de la recherche. D'une part les données diverses dont elle se sert au cours de sa recherche : références d'articles, travaux des collègues, images documentées, communications des colloques, dialogue avec les collègues.

D'autre part, les données de son terrain de recherche, par exemple, un corpus de 6000 photos cédées par un photographe amateur. Bien qu'elle ne les ait pas produites elle-même, elle les considère comme les données propres à sa recherche, matériaux sur lesquels elle va élaborer sa démarche interprétative.

LE CADRE JURIDIQUE FRANÇAIS DOIT EVOLUER EN FAVEUR D'UN « FAIR USE » DES IMAGES PHOTOGRAPHIQUES

Le problème principal que rencontrent les chercheurs dans son domaine concerne l'utilisation des images photographiques protégées par le droit d'auteur et le droit de l'information. Dans son cas, les problèmes relèvent plus du droit d'auteur, car elle travaille avec des photographies de paysages.

De manière générale, les chercheurs utilisent deux types d'images pour ses recherches : produites par des amateurs et produites par des professionnels. Surtout dans le premier cas, il est extrêmement compliqué et contre-productif de remonter vers les auteurs de ces images, les contacter et demander des autorisations pour les utiliser.

Face à cette situation, la solution retrouvée par un collectif de chercheurs de son laboratoire a été de revendiquer pour l'utilisation des images le statut du « fair use » ou de l' « exception pédagogique », usage non commercial destiné à la recherche et à l'enseignement. Il s'agit d'un acte politique des chercheurs, vu que stricto sensu ils ne sont pas protégés par une loi leur donnant un statut à part.

Dans ce sens, il est urgent à son avis que le cadre juridique du droit de l'information évolue en France et fasse une nette distinction entre les usages à but commercial et non commercial. En effet, dans la pratique, et comme l'ont constaté les chercheurs, rares voire nuls, sont les auteurs qui portent plainte pour usage non autorisé de ces images.

DIFFUSER RAPIDEMENT DES DONNEES DANS LES BLOGS

Elle diffuse les étapes intermédiaires de sa recherche dans son blog personnel hébergé dans une plateforme collective de chercheurs et pense que la plupart de jeunes chercheurs en font autant. Mais cela reste incompris par quelques-uns de ses collègues qui préfèrent les canaux traditionnels de diffusion des résultats de leurs recherches, comme les revues. Parmi les raisons invoquées de ces préférences, le plagiat revient de forme récurrente.

A ses yeux, le risque de plagiat est présent dans toute forme de diffusion, pas plus dans le support numérique que dans une publication papier. Elle pense, au contraire, que la diffusion signée et rapide de ses travaux en cours sur son blog évite le risque de plagiat.

D'après son expérience personnelle, les bénéfices d'une diffusion de la recherche en cours sont nombreux : elle a tissé des liens durables avec des chercheurs et des photographes qui ont abouti à son intégration dans l'équipe du laboratoire.

ENTRETIEN N°45 (16/07/2014) – DROIT

Modalité d'entretien

Face-à-face

Statut

Chercheur/maître de conférence

Domaines de recherche

Relations employeur/salariés dans l'emploi public

Retours / Positions

UNE AMELIORATION DES CONDITIONS DE CONSULTATION DES SOURCES JURIDIQUES

En tant que juriste elle ne produit pas, à proprement parler, des données mais utilise celles mises à dispositions sur des bases de données juridiques, comme Legifrance.fr. Ces données sont essentiellement des textes juridiques, lois, conventions collectives. Un autre type de données dont elle se sert sont les textes des « doctrines » - ensemble des opinions (écrits, commentaires, théories, etc.) réalisés par les universitaires et les juristes - et des textes de jurisprudence. Les technologies du web ont beaucoup amélioré les conditions de travail de chercheurs en droit et des juristes, leur permettant d'accéder rapidement à un nombre de ressources importantes à travers Internet.

UN TRAVAIL DE TERRAIN QUI NE VISE PAS UNE DIFFUSION EN AMONT DE LA PUBLICATION

En tant que chercheure dans un laboratoire interdisciplinaire, elle produit un certain nombre de données issues d'enquêtes qualitatives. Ce type de travail de terrain vise le plus souvent à illustrer des problèmes d'application des lois du travail dans des contextes précis. Les données collectées sont traités d'un point de vue juridique - et non pas sociologique - et confrontées aux résultats d'une analyse juridique de sources de la jurisprudence. Dans le cas présent, une série d'entretiens enregistrés a été menée auprès d'individus salariés, employeurs et inspecteurs du travail. Ces entretiens sont anonymisés et stockés ensuite pour un usage personnel. Les données ne seront pas réutilisées par d'autres chercheurs, mais les résultats de son analyse seront partagés entre les membres du laboratoire travaillant sur le même projet et venant d'autres disciplines, la sociologie notamment.

EN DROIT, DES PRATIQUES ET UN RYTHME DE PUBLICATION DIFFERENTS

Le partage des données entre des chercheurs juristes est rare, voire inexistante. La question d'une éventuelle diffusion de ces données, en amont ou en parallèle à une publication, ne se pose pas vraiment.

Il faut comprendre la spécificité de cette discipline dans laquelle les chercheurs ont une pratique et un rythme de publication très différents à ceux d'autres disciplines des SHS. Tout d'abord, les auteurs en Droit sont rémunérés pour ses publications et soumettent leurs travaux à des éditeurs spécialisés et reconnus dans le domaine. D'autre part, les juristes ne sont pas soumis à la même pression de publication que les chercheurs d'autres disciplines en SHS, leur carrière ne dépendant pas du nombre de publications pour avancer. Une autre particularité de son domaine est la priorité encore accordée au format papier par rapport au format numérique.

Dans son cas particulier, son activité de recherche menée avec d'autres chercheurs sociologues l'a incitée à déposer des articles en version pré-print sur HAL, mais cela reste exceptionnel.

ENTRETIEN N°46 (16/07/2014) – SCIENCES DE L'EDUCATION ET SIC

Modalité d'entretien

Face-à-face

Statut

Chercheur/Professeur d'université

Domaines de recherche

Conception et usage des technologies issues de l'informatique dans l'éducation

Retours / Positions

DES MODES DE TRAVAIL COLLECTIFS IMPLIQUANT LE PARTAGE DES METHODES ET DES DONNEES DE LA RECHERCHE

Ses recherches procèdent selon deux types d'approches, qualitative et quantitative. L'approche qualitative consiste dans l'élaboration de questionnaires d'entretien avec les acteurs de l'enseignement secondaire et universitaire. Cette étape de travail est menée avec toute une équipe, toutes les données sont partagées dans le laboratoire ainsi que les tâches de mise en format de ces données.

Pour la démarche quantitative, des enquêtes sont réalisés auprès des établissements scolaires et universitaires et les données seront collectées automatiquement à l'aide de logiciels. Le traitement de ces données implique la participation de spécialistes en données quantitatives, à savoir, des statisticiens.

Dans son laboratoire, la mutualisation des données n'est pas seulement occasionnelle mais est, depuis quelques années, un objectif sciemment poursuivi par les chercheurs.

LA QUESTION DE L'OUVERTURE DES DONNEES POSE TOUT D'ABORD DES QUESTIONS POLITIQUES

En effet, il estime qu'actuellement l'enjeu de l'ouverture et du partage des données a deux implications fortes. La première est d'ordre politique : l'absence des données ou la rétention des données par certains organismes est une barrière à la science mais aussi aux transformations des sociétés. Par exemple, lors d'un projet de recherche développé par lui sur les ressources utilisées par les enseignants, il s'est vu refuser l'accès aux données détenues par les éditeurs scolaires qui réalisent périodiquement des grandes enquêtes auprès des enseignants. De ce fait, les pratiques effectives des enseignants concernant les ressources demeurent largement inconnues ; or, ce sujet intéresse les sociétés et devrait pouvoir bénéficier d'une transparence.

OUVRIR LES DONNEES EST AUSSI IMPORTANT, EPISTEMOLOGIQUEMENT PARLANT, QUE CITER SES SOURCES

La deuxième implication est d'ordre épistémologique et pragmatique : la mise à disposition des données à côté d'une publication est aussi importante que la citation de ses sources. En conséquence, les chercheurs ne doivent pas hésiter à entreprendre la démarche de cultiver leurs données. Les raisons lui paraissent évidentes : justification de la méthode, reproductibilité des résultats (pouvoir tester, par exemple, un algorithme) et coopération dans la construction d'un champ de recherche commun.

Mais il est évident que cela impliquerait dans un travail supplémentaire pour le chercheur : il faudrait contextualiser et rendre transparente la finalité des données (pourquoi elles ont été collectées ou produites) et traiter ces données au niveau du « code » pour les rendre interopérables et exportables.

Il estime qu'il faut œuvrer dans ce sens et transformer le modèle éditorial des revues en réalisant des « data publications ». Les données ne devraient pas être un simple « appui » aux hypothèses développées dans les articles mais être elles-mêmes objet d'un traitement éditorial.

ENTRETIEN N°47 (17/07/2014) – ECONOMIE

Modalité d'entretien

Téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Analyse économique des élections, participation électorale, modes de scrutin, opinion à l'égard des migrants

Retours / Positions

LES MICRO-DONNEES DES BUREAUX DE VOTE NE SONT PAS FACILEMENT ACCESSIBLES

Elle utilise les données du Ministère de l'Intérieur, de l'INSEE, de Quetelet et d'instituts de statistique de certains pays à l'étranger, par exemple, l'Afrique du Sud. L'accès aux micro-données des bureaux de vote peut poser un certain nombre de problèmes, obligeant à réaliser des démarches assez longues auprès du ministère. Elle souhaiterait que ces données deviennent plus accessibles mais les chercheurs à travailler sur ce type de données sont peu nombreux, ce qui rend improbable une transformation de cette situation dans le court terme.

EN SCIENCES ECONOMIQUES : PUBLIER EN ANGLAIS POUR ETRE LU ET SOUMETTRE LES DONNEES

Elle publie principalement en anglais et dans des revues américaines spécialisées dans son domaine et réputées dans le milieu scientifique. En sciences économiques, c'est d'ailleurs la seule façon d'être lu et cité.

Le contrat passé avec ces revues engage le chercheur à transmettre aussi les données. Elle ne sait pas comme cela se passe dans d'autres disciplines, mais en sciences économiques cela devient la règle générale, certaines revues refusant de publier un article en absence des données.

LE PEER-REVIEW RESTE LA MEILLEURE FAÇON DE GARANTIR LA QUALITE DE LA PUBLICATION SCIENTIFIQUE

Elle n'a pas l'habitude de déposer des articles sur des plateformes comme HAL par exemple ou SSRN car le temps à passer pour un simple dépôt lui semble très long. En revanche, elle partage ses articles dès que la revue l'autorise directement sur son espace personnel dans le site d'Academia.

Par ailleurs, le fonctionnement de ces plateformes d'auto-archivage ne lui semble pas satisfaisant. Personnellement, elle estime que le peer-review est toujours la meilleure manière de garantir la qualité de la publication scientifique.

Mais des initiatives de chercheurs pour partager des données seraient bienvenues en économie et en sciences sociales. Cela pourrait effectivement mener à des transformations dans les pratiques de chercheurs et favoriser des rapports moins compétitifs et plus collaboratifs, comme ceux actuellement en cours. Pour l'instant cela semble plutôt discret mais, à sa connaissance, le LEO de l'Université de Orléans développe un projet LABEX orienté vers la valorisation des données du laboratoire.

ENTRETIEN N°48 (17/07/2014) – SCIENCES DE L'EDUCATION

Modalité d'entretien

Téléphone

Statut

Maître de conférences / chargé de recherche

Domaines de recherche

Processus d'apprentissage auto-dirigés des langues étrangères en milieu universitaire

Retours / Positions

DES DONNEES EXPERIMENTALES PARTIELLEMENT DIFFUSEES DANS LES COMMUNICATIONS SCIENTIFIQUES

Ses recherches portent sur le processus d'apprentissage autodirigé des langues étrangères dans le Centre de ressources de langues de l'université.

Les recherches, de type exploratoire, visent à décrire ce processus et à évaluer le degré d'autonomie des étudiants. Les résultats de ces recherches peuvent aboutir à des préconisations en matière de ressources ou des dispositifs de conseil aux étudiants.

Elle travaille principalement avec trois types de données :

- Carnets de bord d'étudiants : les étudiants notent leurs actions et usages des ressources, leurs difficultés ;

- Données d'observation non participante

- Entretiens enregistrés et transcrits

Elle développe cette recherche depuis deux ans avec une collègue qui est dans ce projet depuis six ans. Chaque étape de ce suivi réalisé auprès de différentes populations d'étudiants est l'objet d'un article, rédigé en parallèle. Ceci permet de comparer des processus, réajuster des décalages observés dans divers dispositifs, proposer des conseils aux étudiants, prendre en main les difficultés et suivre l'évolution des processus d'apprentissage.

Les données, anonymisées et stockées en local, sont partagées avec sa collègue mais pas du tout dans un périmètre plus large. Elle ne verrait aucune objection à le faire mais personne jusque-là n'a démontrée d'intérêt particulier par ces données. D'autre part, l'anonymisation rend difficile, à son avis, une réutilisation.

D'une certaine manière elle estime diffuser partiellement ces données lors des colloques et conférences et dans ses articles. Mais ces données sont alors soumises à des traitements particuliers quant à leur contenu avant diffusion. Par exemple, elles sont regroupées en thématiques ou accompagnées d'approfondissements à travers des commentaires et références scientifiques.

DIFFUSER LES DONNEES A LARGE ECHELLE EN SHS EST PEUT-ETRE UNE UTOPIE

La question des données de la recherche et leur éventuelle diffusion à plus large échelle dans les SHS suscite son intérêt même si cela représente, à son avis, une sorte d'utopie. Il faudrait du temps, des motivations concrètes, des projets communs, tout un ensemble de conditions qui lui semblent difficiles à réunir dans l'état actuel de l'environnement scientifique.

En tout cas si cela devait se faire, elle estime que les chercheurs devraient prendre en main l'initiative plutôt que se laisser diriger par des superstructures. Il faudrait développer pour cela une véritable communauté, avec des objectifs communs et cela représente peut-être la plus grande difficulté.

ENTRETIEN N°49 (18/07/2014) – HISTOIRE

Modalité d'entretien

Face-à-face

Statut

Chercheur/professeur d'université

Domaines de recherche

Histoire politique et culturelle de l'Amérique latine au XXe siècle.

Retours / Positions

DONNEES « OBJECTIVES » ET DONNEES D' « INTERPRETATION » EN HISTOIRE

Les données pour les historiens sont essentiellement les sources archives, des données qu'on pourrait appeler, à un but de simplification, « objectives ». Il y a, aussi, les données produites par les historiens, à savoir des données d' « interprétation », transformations des premières par le travail herméneutique de l'historien.

UNE MANIERE TRADITIONALISTE DE TRAVAILLER CHEZ LES HISTORIENS…

Lorsqu'ils sont les premiers à traiter une source, ils se l'approprient et ne souhaitent pas les diffuser avant l'aboutissement de leurs travaux de recherche.

En parallèle à cela, il y a chez les historiens un côté très traditionnaliste de culte du « matériel ». Le contact physique avec leurs sources reste, même face aux nouvelles technologies, de première importance.

Pour ces raisons, il estime que les historiens ne se posent pas assez la question des données. Ils travaillent de façon assez individuelle, sont attachés à l'aspect matériel du contact à la source et continuent à vouloir les revues en format papier en plus de la version numérique.

...AVEC QUELQUES TRANSFORMATIONS CHEZ LES JEUNES CHERCHEURS

Bien évidemment, il faut prendre en considération les différences générationnelles et nuancer ce tableau. Pour les jeunes chercheurs qui ont réalisé ses études et débuté leur carrière dans l'environnement numérique, les revues électroniques suffisent largement et ils sont plus disposés à communiquer dans les réseaux sociaux des réflexions qui jalonnent leur démarche scientifique.

Pourtant, même face à ces transformations il est difficile de parler, en histoire, d'un partage de données comme il serait possible en sociologie.

LE PARTAGE EN VUE D'UNE REUTILISATION DES DONNEES BRUTES DE L'HISTORIEN EST-IL PLAUSIBLE ?

Dans les réflexions des historiens sur les données de la recherche, il y voit transparaître plutôt une préoccupation dominante qui relève de l'archivistique, à savoir, la conservation de sources.

Pour lui une question importante à être débattue est : est-il vraiment possible de réutiliser des données de la recherche produites par un autre chercheur en histoire ? Par exemple, dans son cas, il a photographié une sélection d'archives consultées au Chili selon des critères inhérents à l'objet de sa recherche. Cette sélection, qui compose un corpus d'archives, est déjà son interprétation, étroitement liée aux objectifs de sa recherche. Il s'agit d'une interprétation parmi d'autres possibles.

A supposer qu'il les mette en ligne dans une base de données, un chercheur ne pourrait plus les réutiliser en tant que « données brutes ». De là l'importance de bien documenter ces données, le contextualiser, avant de les diffuser, pour éviter une réutilisation inappropriée de ces sources. Toutefois, il est visible aussi par-là, qu'une telle réutilisation sera limitée.

FAIRE DE L'HISTOIRE A DISTANCE GRACE AU NUMERIQUE : QUELS PROBLEMES EPISTEMOLOGIQUES ?

Une question qui l'interpelle actuellement en tant qu'historien, réside dans les changements d'un point de vue épistémologique et pratique, opérés par la numérisation massive des archives et archives de presse, comme on y procède actuellement au Brésil et au Chili. Dans un avenir proche, un historien pourra réaliser ses recherches sur les années 20 au Brésil sans y mettre le pied, uniquement à l'aide des ressources en ligne. Quels problèmes épistémologiques cela pose-t-il?

ENTRETIEN N°50 (21/07/2014) – LINGUISTIQUE

Modalité d'entretien

Face-à-face

Statut

Chargé de recherche

Domaines de recherche

Les corrélats acoustiques et articulatoires de la parole expressive (expressions d'émotions et d'attitudes)

Retours / Positions

EN PHONETIQUE DES DONNEES ACOUSTIQUES ET PHYSIOLOGIQUES NUMERISEES SYSTEMATIQUEMENT

Les recherches en phonétique et en linguistique produisent principalement des données acoustiques et des données physiologiques. Les premières sont le plus souvent des enregistrements audio et les deuxièmes des audio et des vidéos.

Ces données sont numérisées systématiquement et cela est la pratique récurrente bien avant qu'on commence à réaliser les numérations massives d'autres types de données en d'autres disciplines.

Les technologies utilisées sont variées, pouvant même être empruntées au domaine médical (notamment pour la collecte de données physiologiques) et vont de la simplicité de l'enregistrement en laboratoire à des logiciels de traitement acoustique de pointe.

Le processus de collecte des données susdites comprend des étapes de prétraitement technique et de notation du flux de la parole (contextualisation de chaque plage des enregistrements, association des signaux acoustiques à des données symboliques). Ce n'est qu'ensuite que le travail interprétatif d'analyse peut avoir lieu et donner place à un article.

Ces données sont produites par des tests des réactions perceptives à un certain nombre de sons. Dans ce cas il s'agit souvent de données quantifiables, par exemple, enregistrement des récurrences de confusion entre certains sons chez des individus.

LA LINGUISTIQUE ET LA PHONETIQUE A CHEVAL ENTRE DEUX APPROCHES

Le but des recherches dans son domaine est l'étude du fonctionnement de la parole, soit dans des aspects très fondamentaux des théories linguistiques, soit dans des applications plus pratiques, comme par exemple l'apprentissage des langues étrangères. Cette discipline est pour ainsi dire au croisement d'une approche très objective de collecte des données et d'une approche interprétative qui la rapproche aux autres disciplines des SHS. Sa méthodologie est, en tout cas, celle des sciences expérimentales.

LES TYPES DE DONNEES COLLECTEES DANS UNE APPROCHE EXPERIMENTALE NE PERMET PAS UNE REUTILISATION

Il estime que la création des corpus de données mutualisées ayant pour but une réutilisation n'est possible que si ces données sont à visée très généralistes, comme les méthodologies de recherche, par exemple.

En linguistique phonétique, la collecte des données ne va pas sans un nombre considérable d'aléas et d'imprévus qui sont partie intégrante de la recherche elle-même, sans compter l'importance du contexte dans lequel cette expérience a été réalisée, impossibles de reproduire surtout lorsqu'il s'agit des collectes de la parole spontanée. Il voit mal comment ces données pourraient être réutilisées dans d'autres recherches semblables car celles-ci ne peuvent pas opérer d'une autre façon sinon à travers d'une approche expérimentale. Ces données ne sont absolument pas de « données brutes » comme le seraient peut-être les données issues d'une approche morphologique comparative des organes vocaux. Il pense d'ailleurs qu'il est extrêmement difficile de parler des « données brutes » dans son domaine, car toute collecte est ciblée au départ par un objectif et une expérience qui va la provoquer.

Il y a néanmoins des initiatives de partage dans son domaine surtout dans des projets collaboratifs et soutenus par le CNRS. Mais dans son laboratoire le stockage se fait en local et en cloud la plupart du temps.

EN LINGUISTIQUE, LA DIFFUSION LA PLUS PROBABLE EST EN AVAL AUX PUBLICATIONS

Une diffusion des données ne serait probable qu'après avoir finalisé la recherche associée. Les chercheurs sont compétitifs et peu disposés à partager des données qui leur ont coûté beaucoup de temps.

Les linguistes déposent parfois leurs corpus dans la plateforme ELRA (European Language Resources Association) dont l'accès est en partie payant.

Les linguistes en phonétique dépendent beaucoup de la publication en revues spécialisées et actes des conférences et publient encore très peu en Open Access. Mais cela va évoluer à son avis.

Ces revues ne proposent que très rarement une diffusion audio ou vidéo pour des extraits de données, cela pourrait être une évolution intéressant à développer en plus large échelle.

EN SHS LA TENDANCE EST D'ALLER VERS LA MUTUALISATION DES DONNEES

De manière générale, il pense néanmoins que la tendance est d'aller vers la mutualisation des données en SHS.

Cette question est l'objet de nombreux débats actuellement et le CNRS soutient plusieurs projets de valorisation de la recherche à travers le traitement et mise à disposition de ces données. Il est un peu dubitatif concernant une possible normalisation de ces données et une description de métadonnées en Dublin Core, comme le prône Huma-num. Dans le cas des corpus phonétiques cela serait compliqué pour l'intégralité des données qui est d'une nature trop hétérogène.

ENTRETIEN N°51 (22/07/2014) – SIC

Modalité d'entretien

téléphone

Statut

Chercheur/maître de conférence

Domaines de recherche

Médiation culturelle et scientifique, communication culturelle, espaces publics et arts de la scène

Retours / Positions

DES DONNEES DE TERRAIN SOUMISES A LA METHODE DE « CONFRONTATION »

En Sciences de l'information et de la communication (SIC), il n'existe pas une méthodologie propre de production des données, celle-ci étant la plupart du temps empruntée à la sociologie, aux sciences du langage, à l'économie. Les SIC sont par définition une science interdisciplinaire qui utilise souvent des données économiques, des données d'entretiens, des données d'observations.

Dans son cas, travaillant principalement sur la médiation culturelle, elle produit trois types de données « matériaux de la recherche » :

- Données d'entretiens : les entretiens sont semi-directifs et conduits dans une approche anthropologique. Cette méthode, assez classique, ne peut éluder la vraie question : comment traiter ces données ? Question qui, à son avis, n'est pas abordée de façon systématique avec le sérieux qu'elle mérite en France. Dans sa conduite d'entretiens avec des médiateurs culturels elle opère selon une méthodologie qui inclut trois étapes : analyse des entretiens par observation des récurrences de thèmes, préoccupations et allusions, élaboration de questions et des grilles d'analyse et discussion focus en mode groupe, c'est à dire confrontation des questions à des médiateurs culturels qui lui permettent de mettre en perspective et ajuster sa démarche du départ.

- Données d'observation de groupes, acteurs et publics divers : ces données sont panélisées et donnent lieu à des colloques où différents acteurs prennent la parole, dans une méthodologie similaire à la précédente.

- « Traces médiatiques » : matériaux récupérés sur Internet donnant lieu à des analyses anthropologiques des dispositifs des offres culturelles. Par exemple, elle a travaillé sur l'offre culturelle des musées qui proposent leurs espaces pour des événements comme les fêtes d'anniversaire d'enfants.

PAS DE PARTAGE AVANT L'ECRITURE SCIENTIFIQUE

Tous ces matériaux collectés ne sont pas partagés en tant que données brutes et, même après anonymisation, elle estime que les entretiens ne doivent pas être diffusés. Elle transforme toutes ces données en les « thématisant », créant des tableaux et fiches descriptives, ce qui lui permettra de travailler par comparaison et passer à l'écriture scientifique. Partager ces étapes intermédiaires ne lui a jamais effleuré l'esprit car elle pense que c'est une méthode très personnelle de travailler qui n'aurait aucune utilité à être diffusé.

ENTRETIEN N°52 (28/07/2014) – PSYCHOLOGIE

Modalité d'entretien

E-mail

Statut

Chargé de recherche CNRS

Domaines de recherche

Psychologie de la perception et psychologie cognitive expérimentale, intuition mathématique chez les enfants et nouveaux nés

Retours / Positions

DES DONNEES « CLASSIQUES » EXPERIMENTALES NUMERISEES ET ENREGISTREES

Elle travaille uniquement sur des données très « classiques » issues de situations expérimentales, et ceci même lorsque ces expériences sont menées sur le terrain, à savoir, dans des écoles ou en maternité. En règle générale, ces expériences se déroulent en présentant à chaque enfant ou bébé diverses situations, bien contrôlées, dont les réponses seront enregistrées. Les réponses sont principalement des clics sur des boutons de réponse, orientation vers un coin ou l'autre de la pièce, temps de regard vers une diversité d'images ou films pour les plus jeunes. Plus rarement, des réponses verbales ou aux mouvements sont collectées et dans ces cas elles seront enregistrées en audio ou vidéo. Ceci dit, d'autres psychologues vont sur le terrain pour réaliser des observations; et dans ce cas-là ils doivent prendre des notes et réaliser des enregistrements.

Les données sont en général numérisées dès leur collecte et leur utilisation par la suite est très réglementée par les comités d'éthique. D'abord, l'anonymisation de toutes ces données est obligatoire. Elles sont stockées pendant 10 ans suivant la publication, et ensuite doivent être détruites.

UN PARTAGE ENTRE CHERCHEURS

Le partage est possible entre les chercheurs, à l'exception des données audio et vidéo qui représentent un cas à part et qui demandent l'autorisation des parents des enfants ainsi que le détail des modalités de diffusion acceptées ou non de ces enregistrements. En général, seuls les chercheurs impliqués dans la recherche ont le droit d'avoir accès à ces données.

Une fois que les parents ont accordé leur autorisation, les données vidéo/audio peuvent être diffusées dans le cadre des cours ou de présentations scientifiques.

En ce qui concerne les autres données numériques collectées, elles sont partagées immédiatement avec les collaborateurs de sa recherche. Après les publications des résultats des recherches, elles peuvent être mises à disposition d'autres chercheurs qui en font la demande. Pour l'instant ces données n'ont pas été diffusées en parallèle à des publications, mais elle constate que de plus en plus de revues le demandent.

Les résultats de ses recherches sont le plus souvent publiés sous forme d'article et diffusés aussi dans des présentations orales.

A son avis, la documentation des données et leur diffusion généralisée seraient très intéressante, notamment d'un point de vue méthodologique et à des fins de réplication de l'expérience.

Dans ses recherches, elle n'a utilisée qu'une seule fois des données produites par d'autres chercheurs et qui étaient déjà publiées. Autrement elle a rarement envisagé de le faire car cela peut poser des problèmes de droit d'auteur.

Personnellement, elle n'a jamais utilisée des procédés de « content mining », mais ces pratiques existent dans le champ de la psychologie et surtout dans la psycholinguistique.

ENTRETIEN N°53 (09/08/2014) – DROIT

Modalité d'entretien

Email

Statut

Directrice de recherche/chercheur

Domaines de recherche

Normes, droit, monde arabe, Égypte, constitution, famille, justice, démocratisation, charia.

Retours / Positions

DES DONNEES DU WEB 2.0 COLLECTEES MANUELLEMENT

Ses recherches ont la double particularité de porter sur des questions d'actualité (évolutions juridiques et politiques en cours) et sur le terrain étranger (monde arabe et plus particulièrement Egypte). Il lui est donc impératif de disposer rapidement de «données de la recherche» sur les évolutions politiques et juridiques dans les pays arabes. L'histoire s'étant «accélérée» depuis 2011, cette exigence est devenue encore plus primordiale.

Personnellement, elle ne produit pas de « données de la recherche » en tant que telles, si l'on comprend par là le matériau brut, non analysé de manière scientifique. Elle utilise plutôt celles auxquelles elle peut avoir accès via le web et contribue à leur rediffusion dans les réseaux sociaux. Par exemple, elle utilise des données brutes trouvées surtout grâce à Twitter. Il s'agit de liens vers des articles de presse, des vidéos sur You-tube, des décisions de tribunaux, des émissions de télévision, des sites spécialisés sur des questions particulières (sites sur les travaux de l'Assemblée constituante en Egypte ou en Tunisie, sites sur les élections législatives, etc.).

Lorsque ces données lui semblent particulièrement utiles, elles les « retweete » ou les met sur sa page Facebook. L'objectif est aussi de faire connaître des points de vue particulièrement originaux qui peuvent mieux éclairer la perception d'une question spécifique, comme par exemple, les violations des droits de l'homme en Egypte.

EXPLOITER LES DONNEES AVANT PUBLICATION ET LES RENDRE ACCESSIBLES EN AVAL DE CELLE-CI

Concernant une possible diffusion des données collectées, elle aurait plutôt tendance à garder «égoïstement» certaines données jusqu'à leur exploitation dans une publication, par exemple, les décisions de justice qui sont particulièrement difficiles à obtenir. De manière générale, elle estime que la publication des données en aval à celle des résultats, permettrait d''approfondir certains points particuliers ou de vérifier l'exactitude d'une analyse. Par exemple, l'accès aux textes juridiques n'est pas toujours aisé pour le profane et il arrive assez fréquemment que des erreurs de traduction ou de compréhension soient commises

par des chercheurs non juristes. Il est donc essentiel de pouvoir revenir au texte originel arabe.

FOISONNEMENT DE DONNEES SUR LE WEB EN TEMPS REEL

Le « data mining » et la « fouille de textes » ne sont pas des procédés qu'elle utilise aujourd'hui dans sa recherche, bien qu'ils suscitent vivement son intérêt. Cela pourrait peut-être simplifier son analyse des données du Web.

Les technologies du numérique ont effectivement à la fois simplifié et compliqué sa propre pratique. Depuis son inscription sur Twitter et Facebook, assez récemment, elle a accès à une multitude d'informations dont elle ignorait l'existence et est informée en temps réel de toute nouvelle actualité dans la zone géographique sur laquelle elle développe sa recherche. Mais en contrepartie, le temps passé à recenser et à lire toutes ces informations est assez considérable. Les avantages restent tout de même supérieurs aux inconvénients.