



CLASSIQUES  
GARNIER

COTTE (Dominique), « Les données de la recherche. Un objet de la recherche en sciences humaines et sociales ? », *Études digitales*, n° 2, 2016 – 2, *Le gouvernement des données*, p. 23-39

DOI : [10.15122/isbn.978-2-406-07064-1.p.0023](https://doi.org/10.15122/isbn.978-2-406-07064-1.p.0023)

*La diffusion ou la divulgation de ce document et de son contenu via Internet ou tout autre moyen de communication ne sont pas autorisées hormis dans un cadre privé.*

© 2017. Classiques Garnier, Paris.  
Reproduction et traduction, même partielles, interdites.  
Tous droits réservés pour tous les pays.

COTTE (Dominique), « Les données de la recherche. Un objet de la recherche en sciences humaines et sociales ? »

RÉSUMÉ – Les chercheurs de toutes disciplines sont soumis, dans le cadre d'un mouvement général vers la transparence, à l'injonction de l'ouverture des données et de leurs publications dans des canaux différents des canaux traditionnels. Il conviendrait de s'interroger sur ce que signifient ces données, notamment dans le champ des sciences humaines et sociales. Les sciences de l'information et de la communication sont bien armées pour penser la question des données de la recherche.

ABSTRACT – Researchers from all disciplines are subject, as part of a general movement towards transparency, to the injunction to make data and their publications open—in channels that are different from traditional channels. We should wonder about what these data mean, especially in the field of human and social sciences. The information and communication sciences are well equipped to think about the question of research data.

# LES DONNÉES DE LA RECHERCHE

## Un objet de la recherche en sciences humaines et sociales ?

Soumis à des injonctions concernant le traitement et la mise à disposition de leurs « données », les chercheurs en Sciences humaines et sociales (SHS) sont en même temps particulièrement armés pour penser le phénomène et introduire une dimension réflexive.

Nous empruntons à Jean Davallon<sup>1</sup> la notion de construction d'un *objet scientifique* à partir de la relation qui se crée entre les *objets concrets* et leur manipulation comme *objets de recherche* par les chercheurs. Le discours général des instances de pilotage de la recherche et de la pratique scientifique tend à réifier la notion de « données de la recherche » en présentant ces dernières comme des choses acquises, déjà-là, « données » et auxquelles il conviendrait d'appliquer un certain nombre d'opérations, afin de les mettre à disposition dans une logique de « science ouverte ». Or, il n'est rien moins évident de définir avec précision ce que sont les « données de la recherche » en SHS<sup>2</sup>. Cette contribution constitue une première proposition pour créer un cadre d'étude et d'analyse, à la fois de ce que sont les « données de la recherche » en SHS et du positionnement que cela implique pour les chercheurs.

---

1 Jean Davallon, *Objet concret, objet scientifique, objet de recherche*, *Hermès* n° 38, p. 30-37, 2004.

2 Le propos, qui se veut ici général, doit bien évidemment être précisé et complété en tenant compte des spécificités disciplinaires, ce qui exige des enquêtes complémentaires. On peut trouver une première approche utile dans le travail de Francisca Cabrera (Intd, 2014) qui a interviewé des chercheurs de différentes disciplines sur leur compréhension de ce que sont pour eux les « données de la recherche ».

LA « DONNÉE »,  
QU'EST-CE QUE C'EST AU JUSTE ?

En reprenant le titre d'un article de blog de Sylvie Fayet nous pouvons dire que les « données » sont des « mal nommées<sup>3</sup> ». Bruno Latour proposait d'employer le terme d'« obtenues<sup>4</sup> » afin d'insister sur le processus de production qui encadre la fabrication de la donnée. Les chercheurs ne sont pas des « chasseurs-cueilleurs » qui prennent ce qu'ils trouvent déjà là. Ils prélèvent des éléments au sein de la réalité, selon un canevas qui constitue au préalable leur hypothèse de recherche et organisent ces prélèvements pour leur donner sens ; un *regard* donc pré-existe et toute collecte de données est orientée au préalable, ne serait-ce que par le dispositif mis en place pour les « récolter ». Par parenthèses, on voit ici à travers le vocabulaire (collecte, récolter...) à quel point le mot « donnée » induit par lui-même cette représentation de quelque chose déjà construit et « donné » alors que tout le processus de recherche consiste à produire l'objet scientifique. Pour prendre un exemple simpliste, il faut savoir où braquer le télescope avant d'observer. Mais c'est bien là en partie le cœur du sujet, dès lors que certains évoquent le passage d'une science basée sur une logique hypothético-déductive, à une « *data driven science*<sup>5</sup> ». On considère que, dès lors que la donnée est plus facile, moins coûteuse à produire de manière massive (à travers le « Big Data »), par le seul jeu des machines de captation, l'angle de vue n'a plus d'importance et que la détection et la hiérarchie des phénomènes ne se fait qu'*a posteriori*. On pourrait appeler cela le modèle du « radar » ou de la caméra de surveillance, il suffit de le faire tourner en permanence, et tout événement surpris dans le champ sera automatiquement signalé. Cette vision devient totalitaire si elle suppose que l'on supprime alors tout angle mort et que tout peut être détecté ou surveillé de manière préventive, l'interprétation venant après. Cette vision scientifique, d'une science qui ne vit que par ses appareils s'exprime parfaitement dans l'expression

3 URL : <https://urfistinfo.hypotheses.org/2581>

4 Maryse Carmes, Jean-Max Noyer, L'irrésistible montée de l'algorithmique, Méthodes et concepts en SHS, Les cahiers du numérique, 2014/4, vol. 10, p. 63-102.

5 Chris Anderson, The end of theory : the data deluge makes the scientific method obsolete, *Wired*, URL : <https://www.wired.com/2008/06/pb-theory/>, 2008.

« données brutes » (*raw data*) qui suppose un matériau offert que le travail du chercheur consiste à raffiner a posteriori.

Le terme « donnée » par conséquent, nous bloque, certes. Néanmoins, même si nous savons que le langage est politique, nous devons « faire avec », car l'injonction qui nous est faite, en tant que chercheurs, « d'ouvrir nos données de recherche » suppose justement que cet objet existe. Dès lors que toutes les institutions politiques (OCDE, Ministères...) et scientifiques (CNRS, Académie des sciences...) organisent leur discours et leurs actions en prenant appui sur cette notion, ils lui donnent un corps qui nous oblige à nous en emparer, fût-ce de manière critique et à en interroger le sens. Force est pourtant de constater, comme point de départ, que les nuances proposées dans les différentes définitions qui suivent soulèvent de vraies questions dès qu'il s'agit de cerner précisément ce dont on parle.

Ainsi pour l'OCDE<sup>6</sup>, les données sont des « enregistrements factuels (chiffres, textes, images, sons) utilisés comme sources principales pour la recherche scientifique et généralement reconnus par la communauté scientifique comme nécessaires pour valider les résultats de la recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche ».

Dans le monde anglo-saxon, l'Université de Bristol propose : « Les données, ou unités d'information, qui sont créées au cours d'une recherche, subventionnée ou non, et qui sont organisées ou formatées de telle sorte qu'elles soient communicables, interprétables et adaptées à un traitement souvent informatisé. »

Quant à la Royal London Society, elle avance cette définition : les données sont « des informations qualitatives ou quantitatives (...) qui sont factuelles. Ces données peuvent être brutes ou primaires (directement issues d'une mesure), ou dérivées de données primaires, mais ne sont pas encore le produit d'analyse ou d'interprétation autre que de calculs. »

Pour le CNRS<sup>7</sup> (2016), les « données de la recherche » sont un sous-ensemble des « données de la science » qui, elles, incluent également

6 Marie-Christine Jacquemot-Perbal, Françoise Cosserat, Gestion et diffusion des données de la recherche, INIST, Nancy, 2015, URL : [http://www.inist.fr/IMG/pdf/urfistrennes\\_20150616.pdf](http://www.inist.fr/IMG/pdf/urfistrennes_20150616.pdf)

7 CNRS-DIST, Une science ouverte dans une république numérique, Livre blanc, 2016, 237 p., URL : <http://www.cnrs.fr/dist/z-outils/documents/2016%2003%2024%20Livre%20blanc%20Open%20Science.pdf>

les « résultats de la recherche » distingués à leur tour entre résultats publiés et résultats non publiés. L'idée d'une complémentarité données/publications est ici prégnante, ce qui laisse penser que la reproduction d'un canal de publication pour ces objets spécifiques que sont les données et les résultats intermédiaires, à côté du canal de publication des articles, et le tout dans une logique de disponibilité liée au concept de « science ouverte » est la solution. Plus problématique, à notre sens, est la définition spécifique aux SHS, fournie dans ce livre blanc et qui évoque des objets d'étude aussi disparates que « une chanson, un rapport d'activité d'entreprise, l'architecture d'un monument ». Il est clair qu'on est ici bien au-delà de la « donnée » *produite* dans le cours de l'activité de recherche, et qu'on englobe des notions qui sont des objets ou du matériau *pour* la recherche.

On voit bien ici que ce qui pouvait être identifié, dans une acception première et héritée des sciences dites « dures » de données comme séries quantitatives – et donc calculables – est considérablement élargi à une liste d'objets qui n'a potentiellement pas de fin, et qui mêle allègrement la notion de *données* et ce que, dans certaines de nos disciplines, nous appelons les *corpus*, les *contenus* ou simplement les *documents*. Par ailleurs, rien ici ne parle de ce que la science documentaire appelle les « métadonnées » et qui sont précieuses pour qualifier (nous dirons pour notre part « documentariser ») les jeux de données.

#### LOGIQUES DE DONNÉES ET EFFETS DE RETOUR SUR LES PRATIQUES DE LA RECHERCHE

Les demandes qui nous sont faites aboutissent à mettre en visibilité une pratique qui est à la base et à la condition même d'un travail de recherche : collationner un matériau d'étude à partir duquel des hypothèses pourront être travaillées, testées, validées ou infirmées. Comme on l'a vu ci-dessus, selon les disciplines et la nature même des travaux de recherche, ce matériau d'étude peut prendre des formes très variées, formes que la doxa va subsumer sous la notion généraliste de « données ». Dès lors elle donne à ce matériau un statu *ex ante* et oblige à se

positionner en prévision de ce qui va advenir. On peut en prendre pour preuve la notion de DMP *Data management plan* ou plan de gestion des données (PGD) qu'il est demandé de réaliser en amont et qui finira (qui finit ?) par devenir un des éléments de la sélection des projets et donc de leur financement. Ceci oblige à avoir une vision réflexive sur ses propres pratiques, car ce qui était un matériau-pour-faire devient dès lors un objet singulier, un composant spécifique du cycle de la recherche sur lequel nous reviendrons plus loin.

Cependant, avant de creuser plus avant une macro typologie de ces « données de la recherche » en SHS, il convient de rappeler quels sont les principaux arguments qui sont utilisés par les partisans de l'ouverture des données, qu'ils soient eux-mêmes chercheurs, commanditaires de la recherche (agences, organismes de financement) ou détenteurs du pouvoir politique. Nous pouvons en recenser trois principaux :

- Le premier est celui de la ré-utilisation d'un matériau dont la collecte, la collation et l'organisation a mobilisé des moyens humains et financiers et qui pourrait être réutilisé. On peut ici se poser la question sur la pertinence d'un argument qui dissocie, de cette façon, le matériau « donnée » de ses conditions de production et surtout de ses *intentions* de production. C'est à nouveau accréditer l'idée qu'il existe des données « brutes » que l'on pourrait ensuite accommoder spécifiquement selon les besoins. Nombre de disciplines des sciences physiques ou la biologie ont déjà intégré le fait que la fourniture des données est un élément de l'évaluation à part entière et que tout article (envisagé comme la synthèse finale) de la recherche ne peut être soumis qu'accompagné des « jeux de données » qui ont été produits au cours de l'investigation.
- Le deuxième argument est celui de la mise à disposition de ces données à d'autres publics que les chercheurs, notamment le « grand public », dans une perspective qui hésite entre la notion « d'ouverture » de la science et la logique de vulgarisation. Là encore, cela impliquerait de dissocier les données et leur interprétation, et supposerait que ce matériau puisse « dire » de nouvelles choses s'il est manipulé par de nouvelles mains. En même temps cela soulève la question de

la compétence pour l'interprétation de ces données. La plupart des commentaires et notamment ceux de la conférence d'Amsterdam d'avril 2016<sup>8</sup> insistent sur le fait que des collectifs « citoyens » puissent s'emparer des résultats de la science qui ont des implications sociétales. Ceci sans qu'il y ait, sur le fond, de réflexion sur le statut réciproque de l'expertise et des « savoirs amateurs » ni que soit remise en cause la division du travail dans la recherche qui aboutit le plus souvent à une hyper-spécialisation. L'appel pour l'action sur la science ouverte d'Amsterdam attend de l'ouverture les effets suivants : « Mettre fin au cercle vicieux qui force les chercheurs à publier dans les seules revues prestigieuses, et renforcer la reconnaissance pour les autres formes de communication scientifique ; une plus grande dissémination d'un plus grand volume d'information scientifique qui ne bénéficie plus seulement à la science seule, mais à *la société comme un tout*, y compris le monde des affaires (*business community*). » Ce qui est décrit ici n'est rien moins que le passage d'une écriture pour les pairs à une écriture pour le public. Par ailleurs cet argument laisse complètement de côté la question *politique* de l'organisation de la recherche et de son rapport à la société.

- Enfin, un troisième argument est utilisé dans une logique de la preuve, pour lutter contre la fraude en matière scientifique (la notion stricte de « fraude », comme comportement délibéré peut être élargie à l'idée de vérifier la validité et surtout la rigueur des résultats présentés).

Mais donc, à supposer que les chercheurs en SHS soient effectivement disposés à :

- Réunir leurs « données » dans des ensembles clairement identifiables et *lisibles*
- Ouvrir ces données à toute communauté désireuse d'en prendre connaissance ou possession
- Soumettre ces données aux instances d'évaluation,

---

8 URL : <http://français.eu2016.nl/a-la-une/actualites/2016/04/05/plan-d%E2%80%99action-europeen-pour-la-science-ouverte>

... encore faut-il être capable de les isoler, d'en faire des ensembles cohérents et d'en documenter les usages possibles. Or, encore une fois nous risquons de nous perdre si nous nous contentons d'une approche de recensement des supports. Car à supposer qu'une image ou une photographie soit assimilable à une donnée, son usage est évidemment complètement différent pour un archéologue, un historien d'art, un sémioticien ou un analyste politique<sup>9</sup>.

Par conséquent notre proposition sera d'élargir la catégorisation des « données de la recherche » en Sciences humaines et sociales en prenant en compte « ce que ces données font à la recherche », en fonction de leur type général et surtout de ce qu'elles appellent comme « traitement » ou plutôt comme *travail*, de la part des chercheurs. La situation des SHS est évidemment particulière en ce qu'elles manipulent des objets de sens qui ne sont pas toujours réductibles à des ensembles calculables. Une des questions récurrentes qui se pose est notamment le sort à réserver à tout un matériau préparatoire qui prend la forme d'objets textuels singuliers comme les blogs, les carnets de recherche, les « conversations » dans des forums scientifiques et ainsi de suite... Certains organismes, dans leur définition de la donnée, excluent explicitement ces éléments en les identifiant à des « archives », et d'autres, au contraire, en font une partie intégrante du matériau préparatoire de la recherche.

Nous envisageons alors d'examiner plusieurs types de données qui engagent des manières différentes de travailler et qui supposent des statuts différents de la « donnée ». Elles ne sont pas exclusives les unes des autres et ne semblent pas non plus caractéristiques de l'une ou l'autre des disciplines composant les SHS, du moins de manière catégorique.

---

9 Marie Després-Lonnet, *Temps et lieux de la documentation : transformation des contextes interprétatifs à l'ère d'internet*, Mémoire d'Habilitation à Diriger des Recherches (HDR), Université de Lille, 2014.

## LES DONNÉES DE PRÉLÈVEMENT

Identifier dans une masse de faits des indices signifiants à recouper et à traiter revient à opérer dans le réel une série de prélèvements – organisés selon une intention directrice – afin de construire une représentation. C'est sans doute l'acception la plus classique de ce que nous pouvons identifier comme « données » en sciences humaines et sociales. Mais les sciences humaines et sociales « prélèvent » aussi dans la réalité à partir d'autres méthodes, comme l'enquête ou l'entretien (selon la distinction que nous n'interrogerons pas ici entre les méthodes dites « quantitatives » et « qualitatives »). À travers ces pratiques, il est question de produire des éléments d'analyse pour interroger les hypothèses émises et inscrire les résultats de ce travail dans une démarche objective. Interpréter ici l'injonction qui est faite de « l'ouverture » de ces données, pose la question de ce qui doit être mis à disposition et sous quelle forme : s'agit-il des éléments premiers (terme préféré à « bruts ») recueillis lors d'entretiens, d'observations ? S'agit-il de mettre à disposition directement des fichiers sons d'entretiens, des transcriptions complètes, des synthèses, des *verbatim* ? *Quid* des aspects juridiques et éthiques vis-à-vis des personnes qui auront été interrogées ?

Pense-t-on que, parce que ces éléments sont dans des formats plus « lisibles » que des séries de chiffres ou des données mathématiques, ils sont *ipso facto* plus facilement interprétables par des publics différents ?

Il est clair que la question ouverte de la « Big Data » interroge les sciences humaines et sociales, car ce qu'elle nous (pré)dit c'est le fait que par le jeu de capteurs, par le biais de mise en « traçabilité » des actions et des pratiques sociales dès lors qu'elles sont numériquement appareillées (et elles tendent à le devenir de plus en plus), il n'existe plus d'effort nécessaire pour produire cette donnée. Le monde et son interprétation appartiendraient alors aux *data scientists* et aux seules sciences quantitatives. C'est ne pas voir ici que ces attitudes, pratiques, comportements et usages n'ont rien de naturel et qu'ils sont socialement déterminés, c'est ne pas voir non plus que la façon d'en capter les « traces » est elle-même puissamment déterminée par les technologies utilisées, lesquelles à la fois produisent, façonnent et recueillent les

éléments qu'ils sont censés expliquer. Les sciences humaines et sociales, et singulièrement les sciences de l'information et de la communication sont alors fondamentalement requises pour dénouer les auto-justifications, et fournir une analyse réflexive et critique.

### LES DONNÉES DE DESCRIPTION OU MÉTADONNÉES

Par ailleurs, si l'on se réfère au mouvement de « l'open data » qui est une autre forme d'injonction faite aux acteurs publics : ministères, collectivités..., on voit bien que les résultats de ces dépôts sont le plus souvent décevants car ils correspondent à une satisfaction minimale de la demande, sans plus d'efforts que le simple dépôt de fichiers. Or un tableau Excel ou une liste d'éléments dans un fichier PDF, sans commentaires, ni explications, ni éclairage n'apportent finalement pas grand-chose pour compléter l'exposé des résultats de la recherche.

Or, même si la mise en forme de ces objets premiers les rend assez souvent plus « lisibles » pour un public extérieur que les données des sciences quantitatives, il y manque des éléments d'interprétation qui peuvent rendre ces jeux de données pratiquement inopérables et irrécupérables pour autrui. *A contrario*, la volonté de les rendre opérables et récupérables exige un gros effort de « documentarisation » dont il faut se demander à qui il incombe : aux chercheurs eux-mêmes, aux professionnels de l'information scientifique et technique ?

La question devient bien alors celle de la façon de « donner à ses données » non pas la valeur scientifique qu'elles ont intrinsèquement – dès lors qu'on suppose qu'elles ont été recueillies et travaillées correctement – mais une sorte de valeur « méta-scientifique » qui les rende prêtes à une exploitation enrichie, voire à un croisement avec d'autres données du même type. Par exemple si des chercheurs de différentes disciplines étudiant un même objet arrivent à des conclusions identiques à partir de méthodologies disciplinaires différentes, chacun explicite ses résultats, soit dans le corps des articles, soit dans les dossiers rendus aux instances d'évaluation de l'agence de financement. Mais ce qui serait vraiment intéressant serait qu'un lecteur, quel que

soit le public auquel il appartienne (autres chercheurs, professionnels du domaine, grand public...) puisse bénéficier de liens qui mettent en regard à la fois les résultats, mais aussi les méthodes et le détail des éléments recueillis. Cela supposerait donc de s'être mis d'accord sur des éléments de description, des référentiels communs... C'est tout à fait possible, mais à condition d'être effectivement en contact et d'être au courant des recherches effectuées par d'autres. On pourrait donc imaginer que les organismes de dépôt et les systèmes d'archives ouvertes ne jouent pas simplement un rôle de mise à disposition mais permettent de produire des connaissances par le croisement des informations recueillies.

#### LES DONNÉES D'ENVELOPPEMENT

Cette question est au cœur des méthodes qui sont en discussion dans le cadre des « Humanités numériques » : une bonne part de ce qui est utilisé comme « terrain » par les chercheurs l'est sous la forme de contenus numériques qui ne sont accessibles que par le biais de dispositifs techniques spécifiques qui constituent en même temps une « enveloppe » indispensable des contenus<sup>10</sup>. Qui cherche par exemple à analyser telle ou telle tendance ou comportement sur le web ou dans les réseaux sociaux est obligé d'y être impliqué et de s'intéresser à la fois au dispositif et à ce qui s'y passe. Quelles sont alors les « données » à extraire pour les conserver ? Que signifie, dans une optique disons, Poppérienne, l'idée de reproductibilité des résultats de la science dans un contexte où l'objet d'étude n'est jamais exactement le même parce qu'il est basé sur des technologies dynamiques qui en modifient les contours à chaque instant. Contrairement au premier cas où le chercheur ou l'équipe ont en quelque sorte la main sur la production de « leurs » données, ils sont ici dépendants d'autres dispositifs qui influent eux-mêmes sur le sens des « données » qu'ils proposent à l'analyse.

---

10 Sabine Loupien., *Bibliothéconomie des archives audiovisuelles : les archives sonores à l'heure des *digital humanities**, Thèse de doctorat, sous la direction de Imad Saleh, Université Paris 8, 2016.

Cette question est particulièrement étudiée par les disciplines de l'archive, dont la vocation est de favoriser la préservation à long terme des productions intellectuelles d'une société donnée. Dans le cas des contenus numériques, la question cruciale qui se pose est celle de savoir comment pourront être « rejoués » les documents (au sens large, ou encore les œuvres) numériques, lisibles à un instant T avec une technologie, sachant que les cycles d'obsolescence des supports (pensons à la disquette par exemple) et des logiciels sont de plus en plus courts. Cela suppose la mise au point de formats descriptifs qui consignent des informations portant à la fois sur les contenus, sur les conditions de production, sur les logiques de préservation<sup>11</sup>. Ces données font à leur tour l'objet de descriptions selon des référentiels reconnus par les différentes communautés qui les manipulent<sup>12</sup>.

#### LES DONNÉES D'EXTRACTION

Un autre cas de figure concerne l'application directe sur le support d'étude (un document ou un corpus...) de techniques qui relèvent de l'analyse de données, comme le *Text and Data Mining* (TDM). Ceci peut d'ailleurs avoir plusieurs objectifs : soit « faire parler » les contenus différemment en automatisant leur analyse, et donc produire une sorte de double du fonds initial avec des éléments d'analyse complémentaires (par exemple des calculs d'occurrence de termes, des relevés d'apparitions d'images...). La question du TDM a été un des éléments cruciaux portés par la DIST du CNRS dans son livre blanc pour que la discussion parlementaire sur la loi numérique intègre cet élément et que les éditeurs scientifiques soient contraints d'ouvrir leurs contenus à ces techniques de fouille de données, ce qu'ils se refusaient à faire jusqu'à présent.

Certaines équipes en médecine ou dans d'autres disciplines scientifiques pratiquent une sorte de « méta-science » en compilant les résultats

---

11 Franck Cormerais, *Traitement des textes, sens et logique des formats, Études Digitales*, 1,1, 2016.

12 Par exemple le standard PREMIS recense les différents types de formats de préservation des données (URL : <http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf>).

de dizaines et dizaines d'enquête épidémiologiques pour rapprocher des éléments qui auraient nécessité des années de lecture pour être comparés. Dans les exemples fournis par Rémi Gaillard<sup>13</sup>, certaines disciplines produisent plus de résultats par la ré-ouverture et ré-exploitation de données déjà accumulées que par l'exploitation initiale des premiers jeux de données recueillis.

### LES DONNÉES DE BALISAGE

Dans ce mode opératoire, il y a une dialectique entre modèle et données, le balisage des contenus et corpus étant conçu pour y intégrer des éléments meta-descriptifs (métadonnées) ou opératoires qui permettront de guider un sens de lecture, de recomposer des textes, ou d'en fournir une métalecture par le jeu des annotations<sup>14</sup>. Tous les schémas d'encodage des contenus visent à organiser une exploitation des textes qui peut être à vocation scientifique ou plus largement de conservation et de pérennisation.

Il est clair que, selon les degrés d'imbrication contenus/données que reflètent ces différents types de traitement, la capacité à « ouvrir ses données » ne relèvera pas des mêmes logiques et induira des projets et des méthodologies diverses.

Les différents modes de production ou de traitement des données que nous avons évoqués ne se situent pas dans la même temporalité par rapport au travail de recherche. La logique de la recherche est passée, on le sait, en quelques décennies, d'un rythme d'investigation de long terme, qui pouvait être celui d'une vie, à une logique plus collective et surtout à un rythme guidé par le « projet », selon un modèle éprouvé dans le cadre du management de la firme, de l'entreprise commerciale. Qui dit projet dit séquence, ouverture et clôture, début et fin, orientation générale et étapes intermédiaires, et ceci a été illustré par de nombreux commentateurs sous la forme d'un cercle, défini comme un cycle dont les

---

13 Rémi Gaillard (2014), *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?*, Mémoire, Enssib, Université de Lyon, 104 p.

14 Clémence Jacquot (2016), *Du texte aux données, Texte digital, philologie numérique et dispositif d'attention*, *Études Digitales*, I, n° 1, p. 41-67.

étapes s'enchaînent de manière logique et harmonieuse, depuis l'émission d'une hypothèse jusqu'à la publication. Les différentes étapes de ce cycle sont marquées par des jalons de production documentaire bien délimités, et qui constituent, depuis plus de deux siècles le paradigme dans lequel s'inscrivent la production et surtout la validation des résultats de la science. La publication est le point d'orgue de la recherche et la *revue* l'organe à la fois de validation et de diffusion. On voit assez bien ici que la proposition qui est faite aux chercheurs consiste à redoubler, par des dépôts à travers d'autres canaux que ceux existants, la mise à disposition de ce que le CNRS appelle les « résultats de la science » : données et résultats de la recherche.

Or, rien n'empêche aujourd'hui, à travers la mise à disposition sur les réseaux, les blogs, les *data journals*, les entrepôts d'archives ouvertes de semer, tels des petits cailloux, les traces de toute cette activité, autrement dit de rendre public ce que Latour et Woolgar<sup>15</sup> allaient chercher dans les activités de laboratoire au moment de sa recherche sur « la science en train de se faire ». Mais cela peut-il avoir lieu de manière incidente, comme une sorte de cheminement parallèle à l'activité traditionnelle de la recherche, ou bien cela va-t-il modifier la conformation du cycle lui-même et l'engagement des activités ? La parallélisation des processus de publication des résultats via les canaux classiques de la communication scientifique : revues, articles de colloques, posters... et de publication des données via de nouveaux canaux (entrepôts de données, archives ouvertes, blogs...) induit nécessairement une nouvelle façon de travailler qui inclut en permanence, à chaque étape du cycle de recherche, une préoccupation pour la production et le traitement de la « donnée ».

Est-ce que ceci implique *ipso facto* que l'on s'achemine ainsi vers une autre manière de faire (de) la science ? C'est la thèse – et la position militante – défendue par les tenants de la « science ouverte » ou « science 2.0 » à l'instar du mouvement « Hack Your PhD<sup>16</sup> » qui appelle à une transparence totale des efforts de la recherche durant le processus de recherche lui-même. Cette position est basée notamment sur le fait que les frontières entre le provisoire et le définitif, le « en cours » et l'achevé, l'informel et le formel, le dedans et le dehors sont rendues de

---

15 Bruno Latour, Steve Woolgar, *La vie de laboratoire, la production des faits scientifiques*, Fayard, 1988.

16 URL : <https://hackyourphd.org/> ; <https://hackyourphd.wordpress.com/>

plus en plus poreuses par l'usage des technologies numériques. Là où le chercheur prenait des notes dans un calepin, il nourrit un blog, là où il collait des résultats dans un cahier de laboratoire, il rentre des données dans un cahier de laboratoire électronique, là où il faisait un croquis, il prend une photo numérique. Il est donc très tentant de dire que, puisque c'est là, il suffit de le mettre à disposition. À ce stade, l'enjeu est celui de la circulation et de l'accès à des objets documentaires dont les formes restent inscrites dans la logique canonique de la publication scientifique : *l'article*, signé, validé par les pairs, publié dans une revue classée selon les critères de la discipline, etc. L'autorité scientifique au sens de validation de la recherche, n'est ici absolument pas modifiée, ni dans sa forme ni dans son contenu.

L'idée de la mise en transparence des « données de la recherche » à travers le renouveau des formes documentaires rencontre à notre sens trois obstacles :

- En premier lieu, il serait illusoire de penser (cela renvoie encore une fois au mythe du matériau « brut ») que ce qui est ainsi produit et mis en ligne est exempt d'un regard, d'une vision, qu'il conviendrait alors de renseigner, ce qui est précisément le travail de synthèse qui est réalisé lors de la production d'un rapport, d'un article ou d'une thèse.
- En deuxième lieu, il est déjà admis que la seule sphère de la publication scientifique « officielle » validée (articles, communications, posters...) excède de très loin les capacités d'absorption d'un chercheur, même dans les limites étroites de sous-sous disciplines très spécialisées. Comment alors imaginer absorber en plus toute cette littérature parallèle, que l'on qualifiait autrefois dans les écoles de documentation de « grise » ?
- En troisième lieu, il n'y a pas pour l'instant mise en adéquation des procédures de validation et d'évaluation avec ces nouvelles pratiques. Cet effet de l'œuf et de la poule n'encourage pas les chercheurs – en dehors d'une posture militante – à faire les efforts supplémentaires<sup>17</sup>.

---

17 16 Joachim Schöpfel, Open access – the rise and fall of a community-driven model of scientific communication. *Learned Publishing* 28 (4), 321-325. URL : <http://dx.doi.org/10.1087/20150413>, 2015.

Si les données et les résultats sont des objets logiques (et que nous étudions comme objets d'écriture et de communication), la science elle, est un objet social ; on ne peut la réduire à l'ensemble des projections intellectuelles qu'elle secrète à travers ses travaux, et qu'il suffirait « d'ouvrir » pour en faire des biens communs. Les questions soulevées derrière, sur la commandite, sur le temps de la recherche fondamentale, sur les logiques de financement, les questions juridiques sont autant d'éléments qui relèvent plus du gouvernement de la science que de son format. On ne saurait éluder la question politique derrière celle des conditions techniques de mise à disposition des résultats de recherche et de leurs matériaux préparatoires.

## CONCLUSION

C'est une conclusion en forme de boucle qui sera proposée ici. On voit bien qu'il n'est pas neutre d'envisager la production scientifique comme un travail fondé sur l'acquisition – voire même la production – de « données », c'est-à-dire de « déjà-là ». Et puisque c'est censé être là, le coût de l'effort pour mettre ces objets à la disposition des communautés scientifiques est considéré comme négligeable. La logique de la « donnée » rejoint celle de la « transparence » selon laquelle la simple mise à disposition suffirait à faire sens. Or, tout comme la recherche scientifique elle-même est un travail de médiation (entre le réel et sa représentation consciente), il faut un travail de médiation pour rendre les résultats de la science lisibles et réutilisables.

Puisque ces « données » nous collent aux doigts comme le sparadrap du capitaine Haddock, interrogeons-nous de manière plus large sur ce que sont les *matériaux* que nous accumulons pour produire de la connaissance. La question de fond est alors de savoir comment nous transformons des *matériaux-pour-nous* en *matériaux-pour-autrui*. Comme on l'a dit, ceci ne peut se faire sans un processus de réécriture, ou des écritures multiples, au-delà de la contradiction entre les fameuses « données brutes » et les données travaillées. Or, dans les injonctions qui nous sont faites par les différents organismes concernés, les « cibles » de cette

réutilisation sont multiples : autres chercheurs du champ, chercheurs d'autres champs dans une perspective d'interdisciplinarité, journalistes ou vulgarisateurs, grand public, et le monde nébuleux de l'entreprise ou du « business ». On ne peut pas penser qu'il soit possible, sans un coût cognitif et économique considérable, de satisfaire à toutes ces exigences, dont certaines peuvent apparaître contradictoires. Il semble qu'il soit donc urgent de s'emparer de ce sujet, non pas seulement sous l'angle opérationnel du « comment faire », mais sous l'angle scientifique du « quoi faire » ? Le « quoi » représentant ici à la fois un objet (c'est quoi, finalement ce qu'il faut mettre à disposition) et un objectif (qu'est-ce que nous, chercheurs, *voudrions* faire de nos « données » ?).

Dominique COTTE  
GRIPIC, Université Paris-Sorbonne

## BIBLIOGRAPHIE

- ANDERSON, Chris, The end of theory : the data deluge makes the scientific method obsolete, *Wired*, 2008.
- BLANCHARD, Antoine, Ce que le blog apporte à la recherche, in Dacos Marin, ed. *Read/Write Book*, Open Edition Press, 2010, p. 157-166.
- CABRERA, Francisca, Les données de la recherche en Sciences humaines et sociales : enjeux et pratiques Enquête exploratoire, mémoire, INTD, CNAM, 2014.
- CARMES, Maryse, NOYER, Jean-Max, L'irrésistible montée de l'algorithmique, Méthodes et concepts en SHS, *Les Cahiers du numérique*, 2014/4, vol. 10, p. 63-102.
- CORMERAIS, Franck, Traitement des textes, sens et logique des formats, *Études Digitales*, I, 1, 2016, Paris, Classiques Garnier, 2016, p. 25-40.
- CNRS-DIST, Une science ouverte dans une république numérique, Livre blanc, 2016, 237 p., URL : <http://www.cnrs.fr/dist/z-outils/documents/2016%2003%2024%20Livre%20blanc%20Open%20Science.pdf>
- DAVALLON, Jean, Objet concret, objet  tifique, objet de recherche, *Hermès* n° 38, 2004, p. 30-37.
- DESPRES-LONNET, Marie, *Temps et lieux de la documentation : transformation des contextes interprétatifs à l'ère d'internet*, Mémoire d'Habilitation à Diriger des Recherches (HDR), Université de Lille, 2014.
- FAYET, Sylvie, Les données, ces mal-nommées, 2013, URL : <http://urfistinfo.hypotheses.org/2581>
- GAILLARD, Rémi, *De l'Open data à l'Open research data : quelle(s) politique(s) pour les données de recherche ?*, Mémoire, Essib, Université de Lyon, 2004, 104 p.
- JACQUEMOT-PERBAL, Marie-Christine, COSSERAT Françoise, *Gestion et diffusion des données de la recherche*, INIST, Nancy, juin 2015.
- JACQUOT Clémence, (2016). Du texte aux données, Texte digital, philologie numérique et dispositif d'attention, *Études Digitales*, I, n° 1, Paris, Classiques Garnier, 2016, p. 41-67.
- LATOUR B., WOOLGAR S., *La vie de laboratoire, la production des faits scientifiques*, Paris, Fayard, 1988.
- LOUPIEN, Sabine, *Bibliothéconomie des archives audiovisuelles : les archives sonores à l'heure des digital humanities*, Thèse de doctorat, sous la direction de Imad Saleh, Université Paris 8, 2016.
- SCHOPFEL, Joachim, Open access – the rise and fall of a community-driven model of scientific communication. *Learned Publishing* 28 (4), 2015, 321-325.