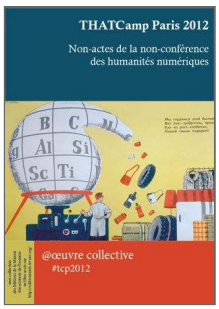

THATCamp Paris 2012
Non-actes de la non-conférence des humanités numériques

Éditions de la Maison des sciences de l'homme

Open data en SHS

Proposé par Cynthia Pedroja, Elifsu Sabuncu, Anne-Laure Stérin

Collectif

DOI : 10.4000/books.editionsmsh.364
Éditeur : Éditions de la Maison des sciences de l'homme
Lieu d'édition : Paris
Année d'édition : 2012
Date de mise en ligne : 1 octobre 2012
Collection : La Non-Collection
ISBN électronique : 9782735115273

Open data en SHS

Proposé par Cynthia Pedroja, Elifsu Sabuncu, Anne-Laure Stérin

Collectif

NOTE DE L'ÉDITEUR

Cet atelier résulte d'une fusion entre les propositions « Open data en SHS. À l'heure où l'on ne parle que de l'ouverture des données de la recherche, comment les chercheurs en SHS envisagent-ils cette question ? » et « Comment diffuser en ligne les résultats de la recherche. Questions de droit et d'éthique ».

1 Open data d'après Wikipedia : Une **donnée ouverte** (en anglais *open data*) est une information **publique brute**, qui a vocation à être **librement accessible** et **réutilisable**. La philosophie pratique de l'open data préconise une libre disponibilité pour tous et chacun, sans restriction de copyright, brevets ou d'autres mécanismes de contrôle.[1]

Introduction

> « Open data en SHS : À l'heure où l'on ne parle que de l'ouverture des données de la recherche, comment les chercheurs en SHS envisagent cette question ? »

2 La première proposition d'atelier concernait la diffusion en open data des données numériques issues de la recherche en sciences humaines et sociales (SHS). Quel lien peut-il y avoir avec le mouvement de l'open data, dans un contexte de volonté politique et sociale de mise à disposition des données publiques[2] ? Les données des SHS relèvent-elles du même statut ? Quel type de données diffuser ? Est-il pertinent de les rendre accessible à plus ou moins grande échelle ? En prenant l'exemple d'un travail sur un projet d'histoire des sciences, où les données sont principalement des enregistrements d'entretiens, quelle forme devrait prendre l'ouverture de ces données (publication en ligne, mise à disposition des données au laboratoire, etc.) ? Le chercheur peut-il conserver ses données lorsqu'il change de laboratoire, ou bien doit-il les laisser au laboratoire ? À qui appartiennent réellement les enregistrements d'entretiens ? Dans quelle mesure un

laboratoire peut-il revendiquer la propriété des données alors que le chercheur est tenu par un contrat moral avec l'interviewé ?

> « Comment diffuser en ligne les résultats de la recherche. Questions de droit et d'éthique ».

3 La deuxième proposition d'atelier recoupait la première, au sens où des questions de droit et d'éthique sont soulevées par la mise à disposition des données de la recherche en SHS. Depuis 2011, un groupe de travail transversal et ouvert a été créé à la MESHS de Lille. Ce groupe, initié par Véronique Ginouvès[3] et Pascal Garret de Crévilles, rassemble des chercheurs, des archivistes, des documentalistes, des juristes, etc., autour de ces questions. Des journées d'études valorisant les retours d'expériences ont déjà été organisées, et continueront à l'être régulièrement. Toutes les disciplines de SHS ainsi que tous les types de supports (sonores, visuels) sont étudiés. Une des principales difficultés soulevées par ce travail réside dans la définition même de ce qui constitue les données en sciences humaines et sociales.

4 Comme les enjeux de l'open data et de la publication des données en SHS sont à la source de nombreux questionnements, il est important de définir clairement l'objet de cet atelier, né d'une fusion entre ces deux propositions initiales : La première proposition consiste à déterminer *« à l'heure où l'on ne cesse de parler d'ouverture des données et d'open data, comment les chercheurs en SHS envisagent cette question ? »*. On se situe ici dans une logique où les chercheurs vont produire des données (c'est-à-dire mettre en place des protocoles qui vont permettre de produire les données). D'autres chercheurs peuvent vouloir avoir accès à ces données, alors de nouveaux protocoles sont à inventer et à mettre en place. La deuxième proposition concerne de ce fait *« la manière de diffuser en ligne les résultats de la recherche. C'est une question juridique et éthique »*. Dans le cadre du groupe de travail constitué à la MESHS de Lille, cette même problématique est abordée, mais à l'échelle de centres de recherche ou d'archives qui ont des fonds et se posent la question de leur mise en ligne.

5 L'atelier porte-t-il sur des problématiques juridiques concernant la mise à disposition de données en SHS ou pose-t-il plus généralement la question de la pertinence de la diffusion des données en open data ? À certains égards, les chercheurs étant des fonctionnaires publics, on pourrait considérer que la diffusion des données en SHS devrait être automatique. Cependant, tout le monde n'est pas partisan de l'open data. L'idée de cet atelier est aussi de se faire l'avocat du diable pour poser les bases d'un argumentaire : « Prouvez-moi que vous voulez avoir mes données et que vous allez en faire bon usage ! ». L'objectif final de cet atelier serait de construire un argumentaire qui permettrait de dire en quoi il est pertinent de diffuser les données des SHS et comment on peut le faire (en pratique et selon quels protocoles).

Quelle est la pertinence d'une diffusion des données SHS en open data ?

6 La discussion s'engage sur les intérêts et les inconvénients pour un chercheur en SHS à partager ses données et la possibilité de mise en place de barrières mobiles pour protéger le travail en cours. Dans certains cas, le choix de diffuser peut également représenter une prise de risque assumée en termes légaux (par exemple pour la publication d'un cliché dont on n'a pu retrouver toutes les personnes représentées).

7 Déterminer l'intérêt pour le chercheur de mettre à disposition ses données en open data est-il un point crucial ? Tous les freins juridiques mis en avant semblent souvent être des paravents. La question que se posent en fait les chercheurs est plutôt de savoir *qui* va exploiter les données diffusées ? La deuxième question est de savoir dans quel état d'esprit ouvrir les données ? Deux positions sont envisageables : La première est de le faire seulement dans le cas où toutes les garanties (juridiques) sont possibles. L'autre position consiste à mettre en ligne d'abord, et voir s'il y a un problème ensuite.

8 Lorsque la donnée est accompagnée de son contexte de production, un grand nombre des questions éthiques soulevées par la publication ne se posent plus. En effet, contextualiser les données permet de donner les clefs pour leurs utilisations futures. On évoque au cours de la discussion la possibilité de passer par la création de licences spécifiques, du type creativecommons[4], qui permettrait d'offrir un cadre pour la réutilisation des données et ainsi rassurer certains chercheurs. Cependant, même si les licences sont des dispositifs ingénieux, leur multiplication à l'infini sur chaque problématique juridique ne constitue peut-être pas une solution. L'enjeu de l'open data est précisément de permettre la réutilisation des données, et certaines licences peuvent par trop restreindre ce réemploi.

À qui appartiennent les données ?

9 Dans un certain nombre de domaines, comme le médical ou la biologie, les carnets de recherche appartiennent aux laboratoires. Il est frappant de constater qu'en SHS, on considère souvent que les données appartiennent aux chercheurs qui, bien souvent, quittent leur laboratoire avec. Dans bien des cas il n'y a pas de capitalisation au niveau du laboratoire sur le travail des chercheurs. Les générations successives d'un laboratoire d'histoire de l'art réaliseront par exemple successivement les mêmes clichés sans jamais les partager.

10 Plus généralement, les archives des chercheurs sont souvent considérées comme des données privées par les institutions de rattachement elles-même. Pourtant, d'un point de vue strictement légal, ces archives sont publiques ! L'institution qui paye le chercheur peut, théoriquement, se retourner contre le chercheur en cas de destruction.

11 Cependant, de réelles questions se posent concernant la diffusion de ces archives. Ainsi, les archivistes des archives publiques (départementales) sont en pleine discussion sur le fait de savoir ce qu'ils veulent/peuvent ou pas diffuser. Le risque est de se retrouver dans la même situation avec des chercheurs qui pourraient se trouver privés de statut. Lorsqu'on parle de partage et d'open data, une manière de sortir du problème consiste peut-être à considérer que la publication de sources est partie prenante du travail scientifique. Il faudrait donc distinguer le travail de recherche en train de se faire du travail de recherche abouti. Dans le cas des SHS, cela pourrait notamment permettre la mise en place d'une véritable critique des sources.

Comprendre l'open data au-delà de la seule recherche académique

L'importance de la restitution aux populations

12 Dans certaines disciplines, telle que l'ethnologie, il est important que les données soient restituées aux populations étudiées et qu'elles puissent éventuellement se les approprier. Cette demande de mise à disposition des données est d'ailleurs très récurrente lorsque les études suscitent une participation des citoyens. Une politique d'open data en SHS aurait l'avantage de répondre à cet enjeu.

Responsabilité sociale du chercheur et participation du public

13 Les chercheurs ont aussi une responsabilité sociale qui implique une diffusion de l'information. Quand on parle d'Open data, on parle également du grand public et il est important de ne pas rester trop « chercheur centré ». Comment transférer de l'information scientifique validée à d'autres communautés ? Comment permettre, et même favoriser des réappropriations ? Bien sûr, une telle ouverture des données peut soulever de nouvelles difficultés. Par exemple, sollicité au sujet du projet wikidata, le CNRS a répondu que les chercheurs avaient des droits d'auteurs sur leur contribution, or une telle réponse impliquerait de contacter un par un les 8 800 chercheurs du CNRS afin de pouvoir réutiliser leurs contributions. Lorsque ces mêmes chercheurs publient leurs données sous licence libre, dans PLOS[5] par exemple, on est obligé de passer par un intermédiaire privé (une fondation) pour récupérer des données financées sur des fonds public...

Un changement des mentalités nécessaire

14 Un travail de fond doit être mené pour faire évoluer les mentalités. Un tel changement passe sans doute par des évolutions concernant l'évaluation des thèses. Tant qu'un doctorant n'a pas publié ou remis son travail, il est normal qu'il ne souhaite pas mettre ses sources à disposition. Cependant, on pourrait tout à fait imaginer une sorte de contrat doctoral stipulant que les données soient rendues accessibles au terme de la thèse. Il serait intéressant que ces nouveaux modèles de diffusion commencent à prendre une réelle dimension dans plusieurs champs disciplinaires. Pour l'instant, la seule discipline mettant en pratique un tel principe est l'archéologie préventive. L'inventeur ou le découvreur d'un artefact archéologique original possède l'exclusivité pendant deux ans, cette exclusivité est éventuellement prolongeable sur cinq ans. Au terme de cette période de cinq années, et dans tous les cas, les données sur l'artefact doivent être rendues accessibles aux autres chercheurs.

La nécessaire définition de protocoles et de formats pour la diffusion des données

La diffusion des sources peut être sensible

15 Certaines données ne sont pas diffusables. Par exemple, on ne peut pas rendre compte de la même manière des choses lorsqu'il s'agit de périodes contemporaines avec des témoins de violences, génocides, etc. Pour les périodes anciennes, la diffusion des sources est beaucoup moins problématique.

La définition des protocoles

16 Dans le groupe de travail de la MESHS de Lille, des propositions de protocoles ont commencé à être envisagées avec l'idée qu'ils puissent être utilisés par d'autres. Par exemple, la prochaine journée d'étude sera consacrée à la mise à disposition de données de la recherche concernant les fichiers d'enquêtes produits par des organismes publics ou parapublics (INSEE, INED, Ministères, CNRS)[6].

17 L'idée est de permettre aux gens de savoir comment diffuser leurs données soit selon des modalités restreintes, soit en ayant recours à des modèles plus ouverts. L'idée est avant tout de décrire des situations concrètes, de partir des pratiques, et d'envisager les diverses solutions possibles, en présentant leurs avantages et leurs inconvénients.

Trouver des formes de collecte et des formats pour permettre à terme le partage des données

18 Définir des protocoles et des formats pour le partage de données suppose également de trouver des formes de collectes qui permettent à terme ce partage. Il est de ce fait important de partager les expériences de base de données collaboratives[7] et de valoriser le capital énorme de collecte individuelle de toutes les données primaires produites par le chercheur, stockées sur son ordinateur personnel et qui seront perdues après publication de travaux. Au bout de quelques années, certaines données ne sont même plus lisibles car les technologies utilisées pour les produire n'existent plus !

19 Pour diffuser les données, il faudrait d'abord s'accorder sur la manière dont on pourrait structurer ces informations pour pouvoir les partager, ce qui n'est pas sans difficultés, car l'apprentissage de nouvelles méthodes ou de nouveaux outils est souvent laborieux. Par exemple, dans un laboratoire CNRS, l'intervention d'ingénieurs spécialisés en traitement des données est souvent perçue comme une intrusion dans le travail des chercheurs. Lorsqu'on propose de passer à un nouvel outil, cela pose problème par rapport à un processus mental qui s'est mis en place depuis des années. Proposer un nouvel outil, c'est souvent remettre en cause la méthodologie du chercheur ; ce qui n'est d'ailleurs pas complètement faux.

20 L'autre difficulté provoquée par le changement des méthodes de travail relève d'une certaine peur quant à la mutualisation – même si en échange le chercheur pourrait profiter de ce qu'ont récolté ses collègues. La seconde étape, après la mutualisation, réside dans la publication sur Internet. Lorsque l'on rend les données accessibles, il se

peut qu'un autre groupe de recherche travaillant sur la même population puisse accéder à ces données et les rende visibles sur un site tiers – mais la réciproque est également possible. Comment gérer ce défi ?

21 Le groupe de la MESHS de Lille n'a pas encore vraiment abordé la question des formats. La question de la mise à disposition des données est cependant structurée en cinq temps : préparer, collecter, traiter, diffuser et réutiliser les données/documents/etc.

Création d'une plate-forme pour l'échange de bonnes pratiques

22 Il serait intéressant de disposer d'un lieu au niveau national, ou de la communauté des humanités numériques francophones, pour l'échange de bonnes pratiques en matière de production des données. L'objectif serait de se mettre d'accord sur la manière de les publier et de les rendre interopérables. On pourrait les mettre à disposition dans une sorte d'« entrepôt général ». Une des difficultés réside dans le fait que la diffusion des données n'est pas valorisée dans la carrière des chercheurs. C'est un véritable travail que de diffuser des données et cela prend énormément de temps. En terme de valorisation de ce travail de diffusion, il existe l'exemple d'une plate-forme qui s'appelle NUMES (abes/adonis)[8]. Cette plate-forme est mal connue mais est interopérable avec le projet patrimoine numérique[9]. L'objectif de NUMES est d'inventorier les différents corpus de données mis en ligne.

Former les étudiants au traitement des données

23 Afin de changer les mentalités, la proposition d'un « contrat » entre l'étudiant et son laboratoire semble envisageable. Cela pourrait prendre la forme d'un engagement réciproque : l'étudiant devrait libérer ses données dans un délai raisonnable après la soutenance de sa thèse, mais en retour l'école doctorale aurait l'obligation de fournir à l'étudiant un accompagnement et une formation pour préparer ses données (équipe informatique, formation sur encodage, logiciel, etc.) Plus généralement, cette idée de contrat doctorant/laboratoire soulève la question de l'apprentissage et de la formation des étudiants. Il faudrait même envisager des formations avant le niveau du doctorat. En effet, comme le soulignait Paul Bertrand lors de sa non-conférence, *Les digitals humanities sont-elles solubles dans le steampunk*, les étudiants d'aujourd'hui sont plutôt « facebook native » que « digital native ».

24 Changer les mentalités et les pratiques passe par une transformation radicale de tout le système universitaire. On pourrait imaginer n'accepter en thèse que des étudiants ayant deux diplômes ou, plutôt que de rallonger les parcours, penser à intégrer des formations de traitement et valorisation des données dans les cursus. Dans tous les cas, que cela soit pour un double cursus ou une formation spécifique proposée au cours des premiers cycles, il est important que des intervenants extérieurs au monde académique puissent assurer des cours afin de favoriser les échanges de compétences. Par exemple, dans les IUT, un minimum d'intervenants doivent venir de l'extérieur du monde académique, ce qui est important pour qu'un échange se crée avec l'extérieur. L'université ne doit pas rester un monde clos, et les industriels pourraient venir présenter les dernières nouveautés dans le domaine du numérique. Il est cependant important que les étudiants ne soient pas formés exclusivement sur des aspects techniques, ils doivent être sensibilisés à la qualité des données, aux aspects juridiques, etc.

Engager une réflexion d'ensemble dans les SHS sur les pratiques de collecte, la formation

25 Les discussions de cet atelier sur l'Open Data et les SHS ramènent sans cesse à la toute première discussion du THATCamp Paris 2010. On est en fait en train de changer plusieurs paradigmes à la fois dans le monde des sciences humaines et sociales. D'une part les données accessibles pour le chercheur sont aujourd'hui devenues des flux d'information. D'autre part, le travail classique du chercheur consistait à produire une somme de connaissances puis de la publier. Ce qui est en train de changer, c'est que cette publication devient en elle-même un flux aujourd'hui. D'où les difficultés exprimées par les chercheurs tout à l'heure, puisque cela ne correspond plus à la représentation que l'on se fait d'une publication scientifique.

26 La publication scientifique aujourd'hui n'est plus la somme des connaissances à un instant *t*. C'est faute d'une réelle prise en compte institutionnelle de cette dimension qu'on a du mal à appréhender les difficultés que pose la mise à disposition des données. Cette notion de flux de connaissances est très visible en histoire de l'art avec l'exemple de la rédaction des catalogues raisonnés. Les chercheurs souhaitent souvent absolument pouvoir figer la recherche sur le papier afin de pouvoir y mettre leur nom, quand bien même cette publication deviendrait rapidement obsolète. La production d'un objet numérique oblige une réelle prise de conscience par le chercheur du fait qu'il n'est pas le seul à avoir produit la donnée scientifique, et que beaucoup d'autres personnes ont travaillé les décennies précédentes sur ce sujet avant lui.

Conclusion

27 Le monde de la recherche en SHS est confronté au changement des règles de la publication scientifique et au changement même du sens de cette publication des données. S'il est pertinent de vouloir rendre accessibles ces données, on doit distinguer celles concernant les recherches en cours et celles concernant des recherches abouties. Afin de réduire la méfiance que peuvent éprouver les chercheurs en SHS par rapport à la diffusion de leurs données, les efforts des prochaines années devront sans doute porter sur la responsabilisation des chercheurs ou la contractualisation de ces questions dans le cadre des formations avec la mise à disposition de moyens pour les aider. Il sera également important de rassurer les producteurs de données à propos de la mise à disposition de ces dernières. La création de licence ou de barrière juridique pour limiter les usages possibles est un exemple de proposition. Une autre difficulté soulevée au cours de cet atelier concerne la difficulté à valoriser le temps passé à mettre à disposition les sources ou les données. Il faut donc aller vers un contrat, une responsabilisation de toutes les parties.

NOTES

1. http://fr.wikipedia.org/wiki/Open_data.
2. Par exemple, la plate-forme d'ouverture des données publiques du gouvernement français (Open Data) : http://www.data.gouv.fr/
3. Véronique Ginouvès, phonothèque de la MMSH : http://phonotheque.mmsh.univ-aix.fr
4. http://creativecommons.org
5. http://www.plos.org
6. http://www.meshs.fr/PUDL/
7. Par exemple, http://larhra.ish-lyon.cnrs.fr/Pole_Methodes/symogih_accueil_fr.php
8. http://www.numes.fr
9. http://www.patrimoineecrit.cultur.gouv.fr/Num.html

RÉSUMÉS

Question de la collecte des sources et de leur publication en ligne. Question des autorisations. Discuter dans l'atelier avec d'autres chercheurs de ces questions et sortir avec une check list pour discuter de ce qu'on doit faire, des risques, et si l'on décide de le faire savoir comment le faire. À l'heure où l'on ne parle que de l'ouverture des données de la recherche, comment les chercheurs en SHS envisagent-ils cette question ? Si la question des sources primaires (documents d'archive, bibliothèques numériques...) est désormais bien balisée, la question des données de terrain (entretiens, relevés d'observations...) est plus complexe : comment anonymiser les témoignages recueillis ? Des informations recueillies dans le cadre d'un protocole basé sur la confiance peuvent-elles être libérées sans contrainte ? Des entrepôts d'open data en SHS peuvent-ils être mutualisés, et si oui à quelle échelle ? Nous proposons de confronter les expériences et réflexions pour faire le point sur cette question épineuse.

INDEX

Mots-clés : open data, données, ouverture des données, données de la recherche