Libre accès et données de recherche. De l'utopie à l'idéal réaliste

Bernard Jacquemin, Joachim Schöpfel, Renaud Fabre

Libre accès et données de recherche. De l'utopie à l'idéal réaliste*

Open Access and Research Data. From Utopia to Realistic Ideals

Bernard Jacquemin[1], Joachim Schöpfel[1] et Renaud Fabre[2]

[1]Univ. Lille SHS, EA 4073 – GERiiCO, F-59000 Lille, France.
{Prenom.Nom}@univ-lille.fr
[2]Univ. Paris 8 Vincennes–Saint-Denis, Laboratoire d'Économie Dionysien,
F-93200 Saint-Denis, France.
renaudfabre@orange.fr

Résumé

L'ouverture des données de la recherche, née à la fois de la mouvance *Open* (logiciel, données publiques, publications scientifiques) et de l'incitation des commanditaires institutionnels nationaux et internationaux, soulève des interrogations non seulement techniques, mais aussi épistémologiques, éthiques, politiques. Cet article étudie les questions posées lors de l'élaboration de l'action de recherche et de la planification de l'ouverture des données, lors de la collecte des données nécessitant leur description en vue de leur conservation, et lors de leur mise à disposition depuis les entrepôts de stockage. Ce texte revient également sur les initiatives d'incitation institutionnelles et les mesures d'accompagnement qui soutiennent la politique d'ouverture des données de la recherche, et fait le point sur l'attitude volontariste des décideurs au plus haut niveau de l'État.

Mots-clefs : Ouverture des données, données de la recherche, politique des données, politique d'ouverture description, éthique, diffusion.

Abstract

Open research data, resulting both from the Open movement (software, public data, scientific publications) and from the encouragement of national and international institutional sponsors, raises not only technical but also epistemological, ethical and political questions. This article examines the questions raised when developing the research action and planning the opening of data, when collecting data requiring a description for storage, and when making them available from storage warehouses. It also discusses institutional incentive initiatives and accompanying measures that support the policy of opening up research data, and reviews the proactive attitude of decision-makers at the highest level of government.

Keywords: Open data, research data, data policy, openness policy, description, ethics, dissemination.

* Jacquemin, Bernard, Schöpfel, Joachim, Fabre, Renaud, 2019. Libre accès et données de recherche. De l'utopie à l'idéal réaliste. *Études de communication*, vol. 52, p. 11-26.

Introduction

L'apparition de l'informatique au milieu du XX^e^ siècle, puis la diffusion de plus en plus rapide des techniques, outils et contenus numériques, a fait de l'univers digital un nouvel incubateur d'utopies et un lieu d'expérimentation et de développement pour des systèmes basés sur des principes souvent résolument démocratiques, libertaires, humanistes (Boullier & Ghitalla, 2004). On a vu par exemple naître et croître la mouvance du logiciel libre, dont de nombreuses productions – comme *PMB* ou *Zotero* pour le monde de l'information documentaire – s'affichent désormais comme des succès incontournables (Zimmermann & Foray, 2001). Dans l'univers info-communicationnel, l'ouverture aux publications scientifiques (*Open Access*, archives ouvertes) s'est imposé au point que les grands acteurs commerciaux de l'édition scientifique eux-mêmes développent une offre *Open* et adoptent des modèles économiques liés à cette ouverture. Depuis la fin du siècle dernier, et plus encore au tournant du millénaire, un mouvement politique et social visant à la libéralisation de l'accès aux données (*Open Data*), principalement scientifiques et institutionnelles, se fait sentir. La *Déclaration de Budapest* (*Budapest Open Access Initiative*, BOAI, 2002) propose ainsi le libre accès aux publications de recherche, tandis que la *Déclaration de Berlin (2003)* élargit la perspective et propose de garantir le libre accès à la connaissance en ouvrant l'accès à la fois à la littérature scientifique mondiale, aux données de cette recherche et aux outils qui ont permis de collecter ces données.

Parallèlement à ces initiatives, la pression s'accentue sur les autorités pour que les données collectées par les pouvoirs publics soient elles aussi mise à disposition, au point que le Parlement européen vote en 2003 la directive *Public Sector Information* (PSI) visant à l'ouverture des données institutionnelles dans une perspective de réutilisation de celles-ci. Côté français, la création en 2011 d'*Étalab* permet de concrétiser la politique gouvernementale en terme d'ouverture et de partage des données publiques issues des établissements publics à caractère administratif, industriel et commercial, ainsi que des entreprises publiques, délégataires ou associatives (publication *via* le portail *data.gouv.fr*). Faisant le pont entre les données liées aux savoirs universitaires et les données collectées par le domaine institutionnel, la *Loi pour une République numérique*, votée en France à l'automne 2016, institue la publication systématique des données produites par les pouvoirs publics, tant au niveau administratif qu'au niveau scientifique, y compris les publications issues de financements publics.

Au sein de ce mouvement réclamant l'ouverture informationnelle et scientifique, la demande de mise à disposition des données de la recherche (*Open Research Data*) est un courant plus récent, et jusqu'à présent à la fois moins médiatisé et moins revendiqué que la mise à disposition des publications scientifiques ou des informations administratives. Pour autant, cette mouvance s'affirme de plus en plus massivement au niveau institutionnel, notamment avec la nécessaire mise en place d'un plan de gestion des données (*Data Management Plan*, DMP) dans les programmes de recherches européens H2020 ou nationaux ANR. La publication des résultats de la recherche ne suffit plus, et il faut désormais pérenniser et dans la mesure du possible partager les données, brutes ou pré-traitées, collectées au cours des projets de recherche (Dillaerts & Boukacem-Zeghmouri, 2018). Le monde de l'édition scientifique suit aussi ce courant et demande de plus en plus souvent aux auteurs de garantir la mise à disposition des données décrites dans les publications à fin de preuve. La plupart des universités et grands établissements de recherche mènent actuellement une réflexion pour assurer une politique de conservation et partage des données de la recherche menée en leur sein et pour amener les chercheurs à faire évoluer leurs pratiques dans ce domaine.

Comme il est souvent de mise dans l'univers numérique, c'est par la clef de la technologie informatique – et notamment les infrastructures – que les données de la recherche sont généralement abordées. Pour le chercheur en Science de l'information et de la communication, ce point d'entrée techno-centré est peu productif et trop focalisé sur la faisabilité, les outils et les procédures. Nous préférons les traiter sous l'angle à la fois informationnel – puisque les données sont constitutives de l'information – et communicationnel, car l'ensemble du mouvement est communication, entre les chercheurs, les institutions, les citoyens, les éditeurs, l'industrie… Ces dimensions essentielles méritent d'être mises en lumière (Chartron & Schöpfel, 2017).

Cette mouvance de la science ouverte et des données ouvertes suscite bien des espoirs, notamment dans la perspective d'une amélioration de la confiance dans les résultats de recherche publiés, mais aussi dans le développement de nouvelles approches de recherche et donc dans la construction de nouvelles connaissances. Elle soulève également des interrogations techniques – comment gérer quelles données ? – et épistémologiques – quelle articulation entre les connaissances issues et à construire à partir de ces données ? – mais aussi éthiques – quelles conséquences éthiques dans la collecte, la conservation, la mise à disposition et la réutilisation des données de la recherche ? L'impact de l'ouverture des données sur les pratiques des chercheurs et sur les performances de recherche semble donc particulièrement important (Jacquemin, Schöpfel, Chaudiron, & Kergosien, 2018). Cet impact et ces questionnements sont d'autant plus prégnants dans les Sciences de l'information et de la communication que les données de recherche sont à la fois une ressource productrice de connaissances, comme c'est le cas pour toutes les disciplines scientifiques – et donc qu'elles sont susceptibles d'être valorisées et réutilisées – mais aussi un terrain de recherche particulièrement intéressant puisqu'il s'agit d'informations qu'il faut gérer, conserver et communiquer. En cela, il est important d'étudier cet univers des données de la recherche et de ses acteurs pour rendre ces ressources disponibles, accessibles, exploitables et durables.

Dans cet article, nous allons explorer trois angles d'approche des données de recherche, trois moment dans leur cycle de vie, pour nous pencher particulièrement sur les questionnements que leur mise en place soulève et sur les solutions pratiques et intellectuelles qui peuvent être choisies : planification (et principe d'ouverture), description et gestion, organisation de la diffusion et de la réutilisation. Nous nous penchons ensuite sur l'accompagnement proposé par les pouvoirs publics pour assurer l'ouverture des données. En guise de conclusion, nous revenons sur l'attitude volontariste des politiques d'ouverture menées au sommet de l'État.

1 Planifier l'ouverture des données

Si l'ouverture de la science et des résultats de recherche, notamment sous forme de publication, tend aujourd'hui à remporter une adhésion massive, les chercheurs hésitent encore, et restent réticents lorsqu'il s'agit de donner accès à leur données de recherche (Kaden, 2018).

La planification de la gestion des données de recherche passe initialement par l'adoption du principe de leur ouverture, donc de leur communication à terme à des tiers. Cette ouverture ne va pas de soi, et est même à rebours des pratiques de recherche traditionnelles. Les chercheurs manifestent en effet un réel attachement pour une liberté professionnelle qu'ils revendiquent, et se montrent souvent jaloux des données de recherche qu'ils collectent, peut-être davantage en sciences humaines que dans des domaines où les données sont partagées depuis longtemps, comme la biologie animale, la météorologie ou l'astrophysique. Pourtant, de nombreux chercheurs sont des fonctionnaires ou fondent leur ac-

tions de recherche sur des financements publics, ce qui rend légitime la mise à disposition du produit de leur travail à leur employeur ou commanditaire. Par exemple, en France, l'organisation de l'accès libre aux données scientifiques fait partie des objectifs de la recherche publique (Code de la Recherche, article L112-1 alinéa e). La volonté d'ouvrir les données de la recherche a été confirmé par le plan d'action national 2018-2020 *Pour une action publique transparente et collaborative* dont l'engagement 18 vise à construire un écosystème de la science ouverte dans lequel « la science sera plus cumulative, plus fortement étayée par des données, plus transparente, plus intègre, plus rapide et d'accès plus universel (et qui) induit une démocratisation de l'accès aux savoirs, utile à la recherche, à la formation, à la société » (Etalab, 2018, 57). Ceci passera entre autre par l'incitation à une ouverture des données produites par les programmes de recherche publics, à partir de 2019. L'Union européenne n'est pas en reste, qui incite à mettre en œuvre dans les projets Horizon 2020 de bonnes pratiques scientifiques compatibles avec les *FAIR Guiding Principles* de la gestion et du pilotage des données de la recherche (European Commission, 2017 ; Wilkinson et al., 2016) : les données doivent être faciles à (re)trouver, accessibles et si possible ouvertes, interopérables et réutilisables.

On voit par ailleurs se développer dans les universités et établissements de recherche une offre liée directement à la planification de la collecte, gestion, conservation et diffusion des données de la recherche via des portails permettant d'établir un DMP : *DMP OPIDoR* (INIST CNRS en France), *EasyDMP* (Projet européen EUDAT), *DMP Online* (Digital Curation Centre, Edinburgh), *DMP Tool* (University of California)… Des portails de dépôt des données de la recherche permettant leur gestion et leur mise à disposition commencent également à voir le jour sous l'impulsion de différents acteurs (voir leur recension dans le *Registry of Research Data Repositories* [1]).

Il est finalement nécessaire de prévoir les conditions de l'ouverture des données, qui devrait être systématique dès lors que rien ne s'y oppose. D'un point de vue éthique, les données à caractère personnel ne doivent pas être communiquées, et il en va de même pour les données collectées auprès de personnes qui n'autorisent pas cette diffusion. La réglementation de déontologie interdit également de donner accès aux données de santé sans qu'une anonymisation complète ait pu être opérée (Lauber-Rönsberg, 2018). Par ailleurs, les secrets militaires, judiciaires, économiques ou industriels peuvent également justifier de limiter la communication des données de la recherche. Il est intéressant de noter l'asymétrie de ce compartimentage, les univers de l'industrie et de la sécurité étant de gros consommateurs des contenus ouverts. On veillera par ailleurs, lors de l'ouverture de ces données, à garantir l'affichage de la paternité attachée aux données, car la propriété intellectuelle est inaliénable et ne peut donc pas être dissoute par le recours à l'*Open Science*.

2 Produire et décrire les données

Il est délicat, au début d'un projet de recherche, d'envisager toutes les dimensions de ce projet, et de prendre en compte la richesse, la nature, la variété des données qui seront collectées, produites, transformées. Il est parfois difficile d'anticiper sur les données, plus encore sur les résultats produits aux différents stades de l'analyse de données, et donc de prévoir ce que seront ces données et comment on pourra les rendre compréhensibles. Pourtant, dans le cadre d'une action de recherche définie, il est essentiel d'avoir accès à la fois technique et intellectuel aux données, d'en comprendre la teneur et la structure, d'en percevoir le niveau de traitement et éventuellement de transformation depuis la collecte initiale. C'est l'ensemble du contexte de production et d'élaboration qu'il faut appréhen-

1. https://www.re3data.org/.

der. Par exemple, la réplication reste une phase essentielle de la démarche scientifique, que seule la reproduction fidèle du processus initialement suivi permet de valider. Il s'agit donc de décrire à la fois l'action complète (hypothèses, cadre théorique et méthodologique, outils…) et le terrain de collecte, mais aussi la méthode de collecte et les traitements effectués sur les données (marquages, tests statistiques, transformations…). C'est donc d'abord l'ensemble de ce contexte de recherche qui nécessite une description systématique et précise : cadre, objectifs et finalité, moyens d'opérer et processus suivis.

La collecte et la conservation des données demande elle-même à suivre une démarche favorisant non seulement leur conservation et les traitements qui vont être nécessaires à l'action en cours, mais également la préparation à leur ouverture : conservation et mise à disposition en bonne intelligence pour une utilisation ultérieure, sans préjuger de cette destination. Les données de la recherche doivent donc être stockées dans des types de fichiers dont le choix n'est pas anodin : le fichier numérique doit être dans un format à la fois adapté aux données collectées et aux traitements que ces données vont avoir à subir, mais dans la perspective de leur ouverture, il faut envisager leur conversion (ou leur préservation initiale) dans un format à la fois ouvert et pérenne – ce qui garantit que n'importe qui puisse y avoir accès sans se préoccuper par exemple de disposer d'un matériel trop spécifique ni d'un logiciel difficilement accessible – n'importe quand. Il s'agit enfin de veiller à ce que le format choisi ne consente pas de perte d'information (par exemple en fréquence d'échantillonnage ou en définition visuelle), en tout cas dans des limites qui n'hypothèquent pas leur utilisation à fin de preuve ou de nouvelle recherche.

Ce risque de perte d'information concerne également le niveau de traitement des données à privilégier pour la conservation et l'ouverture. En effet, la Science ouverte n'implique pas l'archivage systématique et définitif de toute l'information relative à la recherche, et donc toutes les données ne sont pas destinées à être conservées définitivement. Ainsi, il n'est pas question de conserver tous les états de toutes les données produites par des actions de recherche indéfiniment : la planification et la description de l'action de recherche doit tenir compte des niveaux d'élaboration des données depuis leur collecte jusqu'à leur exploitation, et décrire les étapes de traitement des données. Il faut ici déterminer le niveau de traitement adapté à la conservation et à la communication pour le jeu de données (*dataset*) conservées, de manière à proposer celui qui est le plus à-même d'être utilisé pour la réplication et la validation de l'action de recherche initiale, mais aussi dans une perspective de réutilisation pour d'autres actions. Cette réflexion doit être menée sans préjuger de la nature de la réutilisation des données, et veiller à conserver granularité informationnelle explicite au niveau individuel, afin de préserver le maximum de possibilités dans la réutilisation des données.

Enfin, organiser et décrire les données collectées et traitées est un cheminement qui doit être effectué autant d'un point de vue externe qu'au niveau interne. Décrire l'action de recherche, les fichiers contenant les données et leur structuration, les traitements appliqués, en un mot faire ce qu'on pourrait appeler le catalogage des données, tout cela est déterminant. Il faut toutefois également procéder à une description des contenus des jeux de données : la structure des objets doit être soigneusement dérivée des catalogues de thèmes de recherche correspondant (comme par exemple la nomenclature des domaines scientifiques du *European Research Council* [2]), afin de permettre le plus largement le couplage entre article et données. L'expérience des *Data Verse* peut aussi apporter des idées de solutions.

2. ERC (2018). *2019 ERC evaluation panels and keywords*. Bruxelles, European Research Council. Disponible depuis `https://erc.europa.eu/funding/starting-grants` (page consultée le 11 avril 2019).

Finalement, la description et la structuration des données doit se faire autant au niveau interne qu'externe : si les jeux de données, leur état de traitement, la structuration de fichiers qui les contiennent et leur implication en collecte et en transformation dans l'action de recherche doivent être précisément décrits, il en va de même pour les données elles-mêmes, à l'intérieur des jeux de données. Les données collectées, pour être traitées, doivent généralement être formatées, structurées, affichées ou étiquetées pour les rendre lisibles, compréhensibles et exploitables méthodiquement, tant par un chercheur humain que par un outil informatique. Il faut donc dans un premier temps établir l'organisation des données, leur structuration permettant leur appréhension et leur traitement systématique dans le cadre de l'action de recherche prévu. On privilégiera ici encore les structures libres et ouvertes bénéficiant d'un grand pouvoir expressif, et accessibles sans effort technique insurmontable (p.ex. le recours aux tableaux tabulés ou au marquage XML). C'est la logique, la nature et éventuellement la norme attachées à cette structure interne des données qui doivent être rendues accessibles à l'utilisateur, et donc un descriptif précis sur l'organisation et la signification internes des données qui doit être réalisé. Il faudra veiller dans ces choix non seulement à établir une structuration fonctionnelle et adaptée à l'action de recherche en cours et à la discipline dans laquelle elle s'inscrit, mais aussi envisager une interopérabilité des données conservées le cas échéant avec d'autres données similaires ou différentes. La connaissance et le choix des normes et standards pratiqués s'avèrent donc ici particulièrement importants pour garantir à la fois l'utilisation initiale des données sans pour autant dépasser le niveau de compétences techniques nécessaires au-delà du champ disciplinaire concerné, ni s'imposer des pratiques qui sortiraient des méthodologies de recherche de ce champ.

D'un point de vue éthique, les points d'attention oscillent entre la nécessité de protéger les éléments sensibles relatifs aux personnes, à la santé, au secret industriel, commercial ou militaire, et la nécessité de fournir les données le plus complètes et les plus interprétables possible. Il faut dès lors envisager différents dispositifs permettant de résoudre cette tension, qui vont de la fermeture totale ou partielle de l'accès au jeu de données (embargo avant ouverture, accès limité à certains usagers) à l'expurgation d'éléments plus ou moins nombreux de manière à pouvoir anonymiser, voire anonymer, les données déposées pour protéger certains contenus [3]. Un marquage préalable des informations sensibles lors de la démarche de recherche initiale permet d'automatiser le processus pour le dépôt de ces données sensibles. Lorsque les données de recherche sont exploitées à fin de validation de la démarche scientifique, par exemple pour une publication, il convient de veiller à ce que la protection des données ne rende pas leur vérification impossible : dans tous les cas, si l'ouverture des données est sélective, la question de cette sélection doit obligatoirement être posée.

3 Conserver les données pour les diffuser

Si l'ouverture des données de la recherche passe par leur dépôt effectif sur un entrepôt en ligne, il ne s'agit pas pour autant d'un espace libéré de toute contrainte, ni accessible à tous sans autre forme de contrôle. Un entrepôt de données répond en effet à des objectifs, des missions qu'un simple espace en ligne n'envisage pas nécessairement. Outre l'espace de stockage nécessaire pour le dépôt des données, il faut prévoir aussi leur préservation technique à l'abri de toute erreur issue de la détérioration ou de la destruction des matériels

3. L'*anonymisation* consiste à masquer les éléments personnels contenus dans des données de recherche au moyen d'un code d'identification accessible au seules personnes autorisées. L'*anonymation* est un processus similaire, mais pour lequel le masquage n'est pas indexé, ce qui rend le processus irréversible à quiconque.

de conservation, donc une redondance des fichiers sauvegardés et un contrôle régulier et systématique de l'intégrité des données dans chaque fichier. Les erreur n'étant pas dues qu'aux aléas techniques, la sécurisation des fichiers doit également passer par une identification, voire une gestion de versions successives, pour toute modification ou altération des données concernées. Cette identification, de toute manière nécessaire initialement pour permettre l'attribution de paternité, et donc la propriété intellectuelle, voire l'attribution d'une licence, liée aux données, est impérative pour toute tâche d'écriture concernant les données ou leurs métadonnées, mais pose la question de la conservation automatique et de l'affichage de données personnelles liées au producteur ou au propriétaire de ces données, et de tous ceux à qui on donne le droit de les modifier, ainsi que des traces de leur action sur les données. La dimension de préservation des données semble dès lors primer sur celle de la protection des personnes concernées. Cette identification peut aller plus loin si l'ouverture des données et limitée chronologiquement ou à des usagers déterminés : la limitation d'accès passe évidemment par une identification personnelle de tous les usagers, mais implique également un processus de décision pour autoriser ou refuser cet accès.

Cependant, la conservation et la préservation des données de la recherche n'est qu'un des aspects des missions que les entrepôts de données ont à assurer. Sous le terme générique de diffusion transparaissent différentes tâches complexes qu'il faut envisager pour que l'ouverture de ces données puisse se faire dans de bonnes conditions. Il convient d'abord d'assurer une certaine visibilité à ces données, de faire en sorte que les publics qui pourraient y voir un intérêt soient en mesure d'en avoir connaissance et de les trouver. Il s'agit ensuite que cette découverte puisse avoir lieu dans des conditions de confiance pour que les données puissent être abordées en toute sécurité, et testées ou réutilisées sans appréhension ni arrière-pensée stérile. Il faut enfin veiller à ce que l'accès aux données ne soit pas conditionné à l'utilisation d'un dispositif donc l'usager final pourrait ne pas disposer aisément, ou ne pas disposer du tout, dans un environnement numérique où les progrès techniques et logiques sont tels que l'incertitude est difficilement dépassable.

Dès lors, les plates-formes de dépôt de données doivent elles-mêmes bénéficier d'une réelle visibilité en ligne, associée à une notoriété positive, de manière à inciter la recherche et la découverte des données de la recherche par les publics intéressés. Ce sont généralement les établissements de recherche ou d'innovation, liés ou non à des institutions nationales ou internationales, qui bénéficient de la meilleure visibilité dans ce domaine, couplée aussi à un capital confiance à la fois issu d'un contrôle de la qualité de la recherche menée en leur sein par le modèle du contrôle par les pairs (l'institution tire parti de la confiance placée dans la recherche menée par ses chercheurs), et de l'aura positive que l'institution elle-même peut avoir dans la communauté scientifique, voire dans la société civile (les données jouissent de la confiance placée dans l'institution à la base de leur collecte et traitement). L'entrepôt de l'établissement aura donc une visibilité et bénéficiera d'un *a priori* positif à la mesure de ceux de cet établissement. Il importe toutefois de dépasser le niveau institutionnel, voire le niveau national, pour profiter de toute la visibilité possible sur un réseau numérique mondialisé. En cela, l'exemple des archives ouvertes et de l'auto-archivage des publications scientifiques peut servir de référence : le référencement centralisé des entrepôts de données d'une part (voir *supra re3data*), et l'utilisation de métadonnées descriptives structurées et encodées selon des protocoles compatibles pour un moissonnage des catalogues de l'autre, sont deux bons outils permettant à la fois une visibilité des jeux de données pour tous les entrepôts renommés selon des standards internationaux reconnus dans la communauté scientifique, et une découverte des données disponibles via des outils de recherche qui ont montré leurs capacités.

D'autre part, pour assurer l'accessibilité à un ensemble de données, il faut s'assurer que chaque jeu déposé dans un entrepôt soit individuellement associé à un identifiant

unique et pérenne qui non seulement donne une forme d'immatriculation à un ensemble de données bien précis dans un état donné sans risque d'ambiguïté même si cet ensemble est altéré ultérieurement, mais qui permette également de localiser ce jeu de données par simple utilisation de cet identifiant unique. Ce type d'identification pérenne existe dans l'univers documentaire numérique, sous la forme de Handle, de DOI, d'ARK, de PURL…, mais il est absolument nécessaire que cette identification soit réalisée au moment du dépôt des jeux de données sur la plateforme de conservation et de diffusion. Il revient donc à la plateforme de dépôt de proposer cet identifiant pour chaque jeu déposé.

Enfin, les entrepôts de données doivent idéalement être en mesure de donner accès aux contenus qu'ils gèrent et conservent une fois le jeu de données sélectionné. Cela signifie que les fichiers doivent être lisibles et exploitables. Or si le choix des formats et des structures de données et de fichiers revient aux déposants, qui doivent s'attacher à privilégier l'ouverture jusque dans la technique numérique, il est illusoire de leur demander d'anticiper sur les progrès dont la technologie informatique bénéficie, et sur les transformations pratiques qui s'ensuivent. Il convient donc que les plateformes liées aux données de la recherche soient en mesure de proposer non seulement une documentation suffisante sur les formats, structures, normes et outils à privilégier au moment du dépôt, mais également les ressources nécessaires pour que des données devenues anciennes, et éventuellement dépassées au niveau des formats et outils, restent lisibles et utilisables, soit en offrant les outils anciens, soit en proposant une conversion dans les formes nouvelles pour assurer leur pérennité.

4 Analyser, accompagner, agir

Dans le cadre du Plan national pour la science ouverte, le Gouvernement a créé un *Comité pour la Science Ouverte* (CoSO) pour la coordination et l'accompagnement de cette politique d'ouverture. La démarche du CoSO vise aussi bien le niveau national et international que local. Ainsi, son « collège données » a élaboré des recommandations pour les appels à projet de l'Agence Nationale de la Recherche (ANR) et contribue au recensement des dispositifs de données mais s'intéresse également aux stratégies institutionnelles et projets locaux. Fin mars, l'ANR a lancé un appel « Flash science ouverte » afin « d'accélérer la maturation des diverses communautés disciplinaires face aux enjeux de la structuration, de l'accessibilité, de la réutilisation, de l'interopérabilité, de la citation, du partage et de l'ouverture des données de la recherche ». Au niveau européen, l'adoption de la nouvelle *Directive du Parlement européen et du Conseil sur le droit d'auteur dans le marché unique numérique* [4] également en mars ouvre la voie pour l'application de l'exception du *text and data mining* en France, avec de nouvelles perspectives pour l'exploitation des résultats de la recherche.

Tout cette dynamique autour des données de recherche présente un triple enjeu pour les Sciences de l'information et de la communication (Schöpfel, 2018) :

Analyser les données de recherche au sein du nouvel écosystème de la science ouverte constituent un défi scientifique multidimensionnel. Comment définir « données de recherche » par rapport aux modèles de l'information, de la communication et des savoirs ? Quelles sont les pratiques et compétences (*data literacy*) des chercheurs ? Quels sont les nouveaux dispositifs et infrastructures, et quels sont les fonctions et rôles qui se développent autour de ces dispositifs ? Comment décrire, structurer et organiser les données ? Comment les communiquer, et dans quel but ?

Accompagner pour assurer l'efficience de l'action publique, le développement de nouveaux dispositifs et la mise en œuvre de stratégies institutionnelles ont besoin d'un

4. http://www.europarl.europa.eu/doceo/document/A-8-2018-0245-AM-271-271_EN.pdf?redirect.

accompagnement scientifique, par l'évaluation des différentes initiatives et par l'élaboration d'indicateurs pour le monitoring à plus long terme. Les Sciences de l'information et de la communication disposent d'un certain nombre de méthodes et d'outils particulièrement intéressants pour cet accompagnement, dans le domaine de l'évaluation des dispositifs et pratiques aussi bien que dans celui de la scientométrie.

Agir le troisième enjeu concerne les Sciences de l'information et de la communication en tant que discipline ou communauté. Comment se positionnent-elles par rapport aux données de recherche ? Quelles sont leurs pratiques et infrastructures ? Y a-t-il une stratégie disciplinaire ? Quels sont ses propres projets de recherche ?

5 Retour à la politique

Les données relevant de l'*Open Data* nécessitent des mises en réseau multiples, scientifiques, économiques et techniques. Ces mises en réseau rendent indispensable une traçabilité « systémique » permettant l'application de la *Loi Valter* qui pose un principe de gratuité de la réutilisation des données publiques (2015) [5] et de la *Loi pour une République Numérique* (2016) [6] qui prévoit que l'échange d'informations au sein du périmètre de l'État, pour l'exercice des missions de service public, ne peut donner lieu à redevance, aux actions de la nouvelle pratique scientifique et de la nouvelle économie relevant de l'*Open Data*.

Ainsi, la Cour des Comptes dans un référé adressé au Premier Ministre par courrier du 11 décembre 2018, après avoir constaté « des difficultés récurrentes et multiples pour se conformer au droit » (= la loi numérique) au sein des trois établissements publics IGN, Météo-France et CEREMA, a rappelé les enjeux de la politique d'ouverture des données publiques et a recommandé de « clarifier la doctrine et les conditions d'application des règles relatives à l'ouverture des données et des codes sources des logiciels, ainsi que celles afférant à la gestion des licences » et à « redéfinir les modèles économiques des opérateurs en tirant les conséquences de l'ouverture des données publiques et de l'attrition des ressources propres correspondantes ».

Dans sa réponse du 4 mars 2019, le Premier Ministre considère que « la donnée doit désormais être vue comme une infrastructure essentielle et critique du fonctionnement de l'économie et de l'État. La maîtrise de la production de la donnée, de son utilisation et de sa valorisation relève d'enjeux que l'on peut qualifier de souverains. (…) La construction d'une (…) infrastructure adaptée nécessite ainsi la mobilisation de plusieurs leviers – financiers, contractuels, juridiques et techniques – ainsi que l'adaptation des modèles de gouvernance (…) ». Il plaide pour une « bonne et rapide transition », soutenue si besoin par des investissements complémentaires (plan d'investissement d'avenir, fonds de transformation de l'action publique…) et accompagnée par des actions de suivi et d'évaluation d'impact.

La politique d'ouverture, en effet, tend à distinguer un amont gratuit et « public », identifié par la puissance publique et donnant ainsi lieu à des valorisations traçables dissociant l'allocation et l'utilisation, qui font ainsi apparaître un périmètre économique et taxable (Fabre, 2017). Ce lien a été très clairement décrit par un rapport parlementaire sur les données géographiques : « La gratuité de la diffusion et de la réutilisation des données géographiques souveraines implique de faire financer leur production par la subvention, à défaut de pouvoir le faire par la vente de ces données. Si le modèle d'affaires de l'*open*

5. https://www.legifrance.gouv.fr/jo_pdf.do?id=JORFTEXT000031701525.
6. https://www.legifrance.gouv.fr/jo_pdf.do?id=JORFTEXT000033202746

data se vérifie empiriquement, un retour vers les caisses publiques s'effectuera par la fiscalisation des richesses supplémentaires créées grâce à la libération des données ». Et de rappeler d'une manière empirique qu'au Royaume-Uni, « une étude fréquemment citée fait par exemple état d'une croissance du produit national brut comprise entre 13 et 28 millions de livres, dont 4 à 8 millions sous forme de taxes, liée à la mise à disposition gratuite de [données publiques]. Enfin, la libération des données associée à une meilleure coordination dans leur production et à une montée en qualité doit permettre de diminuer les coûts liés à l'utilisation de données inexactes, à l'entretien de bases redondantes et aux doubles saisies ou encore aux coûts de transaction liés à la recherche, à l'acquisition et au traitement de ces données » [7].

Les organismes de recherche et les établissements scientifiques publics sont acteurs de cette nouvelle économie de l'*Open Data* et porteurs de la politique d'ouverture des données, confrontés à ce titre aux mêmes enjeux juridiques, techniques, organisationnels et économiques que les autres opérateurs. Pour la recherche, comme le montre notre numéro par ses coups de sonde sur ces thématiques, il y a un champ de considérations nouvelles à analyser, selon les disciplines de l'information et de la communication : cette approche est génératrice d'applications fécondes vers tous les domaines concernés.

Bibliographie

BOAI. (2002). *Budapest Open Access Initiative.* Consulté le 2019-04-11, sur https://www.budapestopenaccessinitiative.org/read

Boullier, D., & Ghitalla, F. (2004). Le Web ou l'utopie d'un espace documentaire. *Revue I3. Information Interaction Intelligence*, *4*(1), 173–189. Consulté le 2019-04-11, sur https://www.irit.fr/journal-i3/volume04/numero01/revue_i3_04_01_11.pdf

Chartron, G., & Schöpfel, J. (2017). Libre accès aux publications et sciences ouvertes en débat. Dossier. *Revue française des sciences de l'information et de la communication*, *11*. Consulté le 2019-04-11, sur http://journals.openedition.org/rfsic/2868 doi: 10.4000/rfsic.2868

Dillaerts, H., & Boukacem-Zeghmouri, C. (2018). Information scientifique et diffusion des savoirs : entre fragmentations et intermédiaires. *Revue française des sciences de l'information et de la communication*, *15*. Consulté le 2019-04-11, sur http://journals.openedition.org/rfsic/4543 doi: 10.4000/rfsic.4543

Déclaration de Berlin. (2003). *Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities.* Consulté le 2019-04-11, sur https://openaccess.mpg.de/Berlin-Declaration

Etalab. (2018). *Pour une action publique transparente et collaborative : plan d'action national pour la France 2018-2020* (Rapport technique). Paris : Secrétariat d'État chargé de la Réforme de l'État et de la Simplification. Consulté le 2019-04-11, sur https://www.etalab.gouv.fr/wp-content/uploads/2018/04/PlanOGP-FR-2018-2020-VF-FR.pdf

European Commission. (2017). *H2020 Programme. Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. Version 3.2* (Technical report). Bruxelles : European Commission. European Research Council. Consulté le 2019-04-11, sur http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

7. Faure-Muntian V. (2018). *Les Données géographiques souveraines*. Rapport au gouvernement, Conseil national de l'information géographique, La Défense. Disponible sur http://cnig.gouv.fr/?p=18055 (page consultée le 11 avril 2019).

Fabre, R. (2017). *Les nouveaux enjeux de la connaissance* (ISTE Editions éd.). London. Consulté le 2019-04-12, sur https://iste-editions.fr/products/les-nouveaux-enjeux-de-la-connaissance

Jacquemin, B., Schöpfel, J., Chaudiron, S., & Kergosien, E. (2018). L'éthique des données de la recherche en sciences humaines et sociales. Une introduction. In L. Balicco, E. Broudoux, G. Chartron, V. Clavier, & I. Pailliart (Eds.), *L'éthique en contexte info-communicationnel numérique. Déontologie, régulation, algorithme, espace public. Actes du colloque « Document numérique et société »* (pp. 71–86). Échirolles : De Boeck Supérieur.

Kaden, B. (2018). Warum Forschungsdaten nicht publiziert werden. *LIBREAS. Library Ideas*, *33*. Consulté sur https://libreas.eu/ausgabe33/kaden-daten/

Lauber-Rönsberg, A. (2018). Data Protection Laws, Research Ethics and Social Sciences. In F. M. Dobrick, J. Fischer, & L. M. Hagen (Eds.), *Research Ethics in the Digital Age : Ethics for the Social Sciences and Humanities in Times of Mediatization and Digitization* (pp. 29–44). Wiesbaden : Springer Fachmedien Wiesbaden. Consulté le 2019-04-12, sur https://doi.org/10.1007/978-3-658-12909-5_4 doi: 10.1007/978-3-658-12909-5_4

Schöpfel, J. (2018). Hors norme ? Une approche normative des données de la recherche. *Revue COSSI - Revue Communication, Organisation, Société du Savoir et Information*, *5*. Consulté le 2019-04-11, sur https://revue-cossi.info/numeros/n-5-2018-processus-normalisation-durabilite-information/730-5-2018-schopfel

Wilkinson, M. D., Dumontier, M., IJsbrand, J. A., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016, mars). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*(160018). Consulté le 2018-08-31, sur https://www.nature.com/articles/sdata201618 doi: 10.1038/sdata.2016.18

Zimmermann, J.-B., & Foray, D. (2001). L'économie du logiciel libre. Organisation coopérative et incitation à l'innovation. *Revue économique*, *52*(1), 77–93. Consulté le 2019-04-12, sur https://www.persee.fr/doc/reco_0035-2764_2001_hos_52_1_410277 doi: 10.3406/reco.2001.410277