Usage et gouvernance des données

Réflexions et préconisations

Groupe de travail du Collège « Données de la recherche »
du Comité pour la science ouverte (CoSO)

Sommaire

Composition du groupe :
Hélène Chambefort, INSERM
Juliette Hueber, CNRS
Claire Lemercier, CNRS
Kenneth Maussang, Université de Montpellier
Anne Vanet, Université Paris Diderot/Université de Paris

Octobre 2019

Contexte et objectifs

La Loi pour une République numérique et le Plan pour la science ouverte soutiennent d'une part l'ouverture des données publiques aux citoyen.nes, qui financent leur création, d'autre part le partage des données scientifiques, pour favoriser l'innovation. Si ces mesures sont légitimes, elles demandent des changements de pratiques importants voire de véritables changements de paradigme qui interrogent, malgré les pistes lancées par le plan pour la science ouverte.

L'objectif de notre groupe est **d'identifier ces changements nécessaires, du point de vue des pratiques scientifiques quotidiennes, afin d'anticiper des blocages possibles et de faire des préconisations pour les prévenir.** Nous avons mobilisé pour cela notre connaissance des pratiques dans les laboratoires de plusieurs disciplines (disciplines représentées : biologie, physique, histoire et histoire de l'art), du point de vue de plusieurs métiers (chercheur.se, enseignant.e-chercheur.se, documentaliste, archiviste – pour simplifier, la formule « chercheur.se » dans la suite de ce document inclut les universitaires).

Synthèse des discussions

L'ouverture des données : pourquoi, pour qui et comment ?

Il nous a semblé essentiel de nous interroger dans un premier temps sur les personnes susceptibles d'être intéressées par l'ouverture des données et sur l'objectif de cette ouverture. Les bénéfices évoqués par le Plan national pour la science ouverte (PNSO) : évitement des duplications d'efforts, intérêt pour l'innovation mais aussi les progrès économiques et sociaux, levier pour l'intégrité scientifique, ne reposent pas tous sur l'ouverture des mêmes types de données aux mêmes publics. Le but est-il de faciliter des **réplications** entre expérimentalistes spécialistes du même sujet, d'éviter des redondances, de permettre une réutilisation de données par une **partie différente de la même discipline** (théoriciens *vs.* expérimentateurs), par d'autres, voire de **nouvelles disciplines** (sociophysique par exemple), ou encore par des **entreprises (petites ou grandes), des associations, des administrations, le « grand public »**, etc. – sans oublier des destinataires plus classiques mais encore pertinents : **étudiant.es, professeur.es du primaire et du secondaire**, etc. ? Actuellement, ces objectifs sont souvent confondus dans une même intention, mais en pratique**, ce ne sont pas les mêmes données, les mêmes formats, les mêmes modes de mise à disposition** qui sont impliqués dans chaque cas.

Le PNSO évoque d'emblée une ouverture des données « autant que possible », posant ainsi le fait que certaines données ne peuvent pas être ouvertes, ou du moins pas immédiatement. Il s'agit ainsi de donner rapidement à cette formule un contenu concret.

Les obstacles à une ouverture totale et immédiate des données se posent en termes juridiques mais surtout en termes de moyens (ne serait-ce que d'espace de stockage sur les serveurs – avec les enjeux correspondants de sécurisation, de localisation, d'environnement, etc.) comme de curation des données. Les archivistes ont déjà largement travaillé sur les problèmes de conservation pérenne (ne serait-ce que pour 30 ans) des données. Le problème se pose en particulier pour les fichiers produits avec des logiciels propriétaires. Le coût en personnel et en infrastructures pour conserver et maintenir des jeux de données utilisables (conversions successives, changement de supports physiques, etc.) est considérable. **Il y aura nécessairement une allocation des moyens dans la mesure où tout ne pourra pas être bien stocké, bien décrit et bien communiqué** : il faut réfléchir à la manière de la réaliser.

On gagnerait dès lors à **dire que la science ouverte implique des choix –** comme l'archivage consiste largement en une sélection de ce que l'on va archiver, ou pas. Certaines données seront à mettre en ligne immédiatement, mais ne seront pas archivées ; d'autres seront conservées mais leur accès sera strictement limité pour un temps, etc.

Le PNSO prend en compte ces limites en évoquant non seulement les secrets encadrés par la loi, mais aussi le fait que l'obligation d'ouverture doive être « encadrée par les bonnes pratiques définies par chaque communauté scientifique, par exemple pour définir des durées d'embargo ».

Il est donc important, pour chaque discipline ou type de donnée, de définir ce qu'il est prioritaire d'ouvrir mais également le public destinataire (chercheurs.ses, grand public, etc.) ainsi que les conditions de réutilisation. En termes de travail concret, mettre à disposition des données afin qu'elles soient consultables

et réutilisables par le « grand public » grâce à un travail de structuration de l'information et d'éditorialisation (ce qui se rapproche des missions actuelles de diffusion de la recherche) n'est par exemple pas la même chose que fournir des jeux de données simplifiés pour l'enseignement secondaire, ou encore rendre une quantité importante de données « brutes » réutilisables par des collègues ou des entreprises.

Des formats communs de description de données : pour quoi faire ?

On évoque ainsi la question de l'élaboration de formats communs de description des données permettant leur réutilisation, en donnant **l'exemple de l'astronomie, qui est effectivement « en avance »** en la matière. On gagnerait toutefois à s'interroger sur les ressorts de cette avance et sur le caractère plus ou moins facilement transposable de cette expérience.

Si l'astronomie est « en avance », ce n'est pas seulement parce qu'elle a commencé plus tôt. Dans cette discipline, **le coût des instruments fait qu'il est intéressant pour tout le monde de réutiliser les données** (et sans doute que le coût de gestion des données est, en proportion, marginal). **Cela n'est pas transposable à toutes les disciplines, ni à tous les types de données.** Ainsi, en génétique, les formats de séquences sont standardisés, pour des raisons du même type ; mais rien n'est prévu pour le partage des données obtenues ensuite, à partir des séquences. Une partie des données en sciences humaines et sociales et en sciences de l'environnement est au contraire issue des observations de terrain de chaque chercheur.se – parfois au crayon sur un carnet –, ce qui tend à produire des incitations inverses en termes d'envies de réutilisation, de standardisation, de compétences et de coûts nouveaux nécessaires pour y parvenir.

Il faut prendre en compte ces situations plutôt que d'appliquer la même méthode partout. En particulier, préciser quelle réutilisation, par qui, de quel type de données parle-t-on, quel modèle pourrait s'appliquer dans chaque discipline. Cela nous semble constituer une première étape nécessaire avant de penser à des formats communs.

L'ouverture des données : quels sont les bénéficiaires ?

Rappelons ici que l'axe 2 du plan national pour la science ouverte vise à « rendre obligatoire la diffusion ouverte des données de recherche issues de programmes financés par appels à projets sur fonds publics. »

Pour atteindre cet objectif, il est également important de comprendre que **la motivation des collègues à partager leurs données ne sera pas la même selon les cas** : plus élevée si des pratiques de réutilisation existent et sont déjà valorisées dans leur discipline, moins élevées s'ils ont le sentiment de travailler uniquement pour une autre discipline (souvent déjà dominante) ou pour des entreprises comme les GAFAM. La communication autour des objectifs de la science ouverte devrait être adaptée à ces différences.

Les chercheurs.ses sont majoritairement conscient.es que leur travail est financé par de l'argent public et ne s'opposent pas en général au principe de l'ouverture des données ; cependant, **les questions de réutilisation commerciale font bien moins consensus**. L'idée que l'argent public permette de produire des données mises à disposition de tous, (entre autres de grandes entreprises du type GAFAM ou Unilever, Kraft, GE…) soulève de nombreuses questions éthiques chez certains personnels du monde académique. C'est d'autant plus vrai qu'il s'agit non seulement de fournir ces données sans contribution de la part des entreprises, mais au prix d'un coût additionnel pour la recherche, occasionné par l'ouverture des données (remise en forme, serveurs, etc.). Il en va de même lorsqu'on évoque la réutilisation par des entreprises ou même des laboratoires de pays qui ne pratiquent pas une ouverture symétrique.

Comment ouvrir les données ?

Les réticences qui apparaissent souvent lorsque nous présentons les enjeux de la science ouverte dans nos laboratoires tiennent au **sentiment d'être soumis à des injonctions contradictoires : science ouverte vs « valorisation »** de la recherche passant notamment par le dépôt de brevet ; **science ouverte vs respect du RGPD ; science ouverte vs concurrence accrue entre équipes** (recherche de financements, *publish or perish*, etc.).

Il paraît ainsi difficile de faire progresser la science ouverte sans faire évoluer les pratiques d'évaluation des personnels impliqués.

Il est également important de continuer à **investir dans l'accompagnement juridique** (voir préconisation C). En effet plusieurs guides juridiques existants ont été utilisés par des chercheurs.ses. Malheureusement, ne partageant pas un vocabulaire commun, les guides n'ont pas pu répondre de manière satisfaisante à leurs

besoins et une grande partie de leurs questions sont restées sans réponse. De plus chaque discipline doit affronter des questions juridiques très différentes qu'un unique guide peut difficilement résoudre.

La gouvernance des données

La gouvernance des données sous-tend un choix préalable des données qui seront publiées donc traitées en conséquence, puis, pour certaines, archivées. Cela implique un travail conjoint des chercheurs.ses avec les archivistes, documentalistes et/ou bibliothécaires.

Les discussions ci-dessus nous ont amené. es à constater d'importantes **différences de pratiques entre disciplines et des différences de traitements entre types de données et confirment la nécessité d'une politique pour la science ouverte qui dépasse le cadre des établissements** (en général divers en termes de disciplines et de types de données). Cela inciterait à discuter des préconisations sur l'ouverture des données à l'échelle nationale (ou plus : cf. les travaux d'EOSC à l'échelle européenne) par discipline ou par grands types de données.

Il faut se poser aussi la question de **l'expression possible des publics visés** sur ces choix : chercheurs.ses d'autres disciplines ou pays, entreprises, administrations, associations, enseignants.es du secondaire, « grand public », etc., mais aussi de publics actuellement impliqués dans la recherche et que l'on pense *a priori* moins favorables à l'ouverture des données (partenaires de la défense, de l'industrie). Dans quel cadre leur donner la parole sur ces choix ?

Le problème de la bonne échelle d'action se pose également en ce qui concerne les moyens importants, matériels et humains, qu'implique un partage des données qui réponde aux critères du *FAIR data*. Une réponse unique et globale à l'échelle d'un établissement – la création d'entrepôts et de services avec des personnels dédiés – doit être accompagnée par une offre de services à l'échelle nationale, mais également un réel accompagnement au niveau des laboratoires (avec l'importance d'avoir des personnes-relais entre la recherche au quotidien d'un côté et les spécialistes à temps plein des questions de science ouverte de l'autre). Le PNSO souligne ainsi l'importance d'infrastructures nationales et européennes dédiées à des thèmes ou disciplines : tout ne peut pas se faire à l'échelle des établissements.

Nous n'ignorons pas que des groupements et des infrastructures existent déjà tant au niveau national qu'européen, mais nous constatons qu'ils restent encore trop peu accessibles à de nombreux.ses chercheurs.ses ou ne répondent pas à leurs besoins.

Préconisations

Dans l'idéal, nous aimerions adresser ces préconisations non seulement au comité de pilotage du comité pour la science ouverte, mais aussi au comité pour la science ouverte en général – ou au moins à son collège « données » ; et tout particulièrement aux groupes de travail « kit pour établissements » et « Dataverse ». En effet, nous n'avons pas la prétention de proposer des solutions définitives, mais nous pointons des éléments sur lesquels il nous paraît prioritaire d'élaborer une position collective, d'une part pour faire avancer la science ouverte dans les laboratoires, d'autre part pour permettre très concrètement l'ouverture pérenne de plus de données.

Préconisation A – Définir ce qui doit être conservé

Il est important de distinguer différentes durées de vie et d'archivage des données, comme on le fait pour les archives en général (qu'elles concernent ou non la science). **Il n'est pas possible – et probablement pas utile – de tout conserver**.

Il est dès lors nécessaire de sélectionner les données à conserver. **Une bonne pratique devrait alors inclure une destruction programmée d'une partie des données, ce qui permet en contrepartie un traitement approprié des données conservées**. Cette réflexion doit intervenir dès le début (à l'élaboration du Plan de gestion de données), en prévoyant une durée de vie par type de données, quitte à autoriser un moment de révision juste avant la destruction. Dans le cas de données très massives, on peut conserver des échantillons (aléatoires ou non) **–** comme on le fait déjà pour les archives administratives.

Des lignes directrices en la matière devraient être définies par discipline au niveau national (au moins) : voir préconisation B.

Préconisation B – Décliner la science ouverte par discipline

Dans un premier temps, il faudrait **élaborer des guides de la science ouverte**, présentant les motivations, les principes et le cadre juridique, plus **adaptés à chaque discipline** (ou au moins, pour commencer, grand ensemble de disciplines ; ou encore à de grands types de données), avec des choix d'exemples et de focales adaptés. Le comité pour la science ouverte pourrait pour cela constituer des groupes de travail en son sein et/ou travailler avec des instances représentatives : sections du CNU, du CNRS, sociétés savantes (récemment regroupées en fédération), etc.

La mise en place d'instances ou de personnes mandatées pour donner le point de vue d'une discipline sur les grands choix à faire en matière de données – quoi ouvrir exactement, à qui, quand, pour quelle durée – **pourrait accompagner l'ouverture des données**. Ce travail doit-il entrer dans la mission des alliances (Athena, Aviesan, Allistene, etc.) ? Comment associer aux choix les différents métiers de la recherche et du soutien à la recherche, mais aussi des représentant.es des publics concernés par la réutilisation des données ?

Préconisation C – Donner aux chercheurs.es la possibilité de se former en matière juridique

Les chercheurs.ses n'ont jamais été formés quant aux liens qu'il pouvait y avoir entre les résultats de la recherche et les obligations/droits juridiques les concernant, concernant leur établissement, ou même l'état. Quid de la propriété des données coproduites dans des projets collectifs entre personnes de statuts différents ? Quid des doctorant.es et postdoctorant.es ? Est-ce que les cours et autres contenus liés à l'enseignement sont soumis au même régime que ceux liés à la recherche ? etc.). Qu'en est-il des données enrichies par les chercheurs.ses, mais provenant de musées, bibliothèques, maisons d'édition, administrations, entreprises, etc. (cas fréquent notamment en sciences humaines et sociales) ? Nous proposons donc de :

- clarifier le statut des différents personnels au regard du droit d'auteur ;
- clarifier les articulations entre brevetabilité et science ouverte et entre RGPD et science ouverte, travailler avec les personnels de la recherche à définir et faire connaître les exceptions de recherche au RGPD ;
- clarifier la situation des données qui sont mises en forme et enrichies dans les laboratoires, mais qui ont été originellement produites ailleurs et relèvent donc d'autres droits ;
- mettre en avant l'impératif d'avoir une licence sur tous les contenus partagés et travailler à une communication plus claire sur les licences possibles et recommandées ;
- gérer les tensions entre d'une part des guides juridiques rédigés de façon à favoriser la science ouverte, d'autre part des juristes d'universités/d'organismes de recherche ou des délégué.es à la protection des données privilégiant le plus souvent, de par leurs fonctions, une interprétation plus restrictive des textes.

Préconisation D – Donner aux chercheurs.es la possibilité de se former en matière de curation des données de la recherche

Afin de permettre aux chercheurs.es de s'approprier les questions liées à la gestion et à l'ouverture des données, il serait bon de généraliser les formations très concrètes avec un travail sur des cas précis, voire sur les jeux de données mêmes des chercheurs.ses, plutôt que des formations plus générales ou théoriques.

Ces formations ne doivent pas considérer comme acquises les phases en amont de l'archivage ouvert des données. Il existe pour l'heure trop peu de **formations sur les moyens et méthodes de stockage des données immédiatement après leur production**, pour soi ou pour une petite équipe de projet. Certains.es chercheurs.ses, par exemple, stockent des photos de textes parce qu'ils ou elles ne savent pas faire autrement. Les formations doivent aussi prendre en compte l'appropriation de savoir-faire liés à la mise en place et à la mise en œuvre de plans de gestion de données, toujours à partir de cas concrets des projets menés par les chercheurs.ses.

La science ouverte ne progressera pas sans **reconnaissance du travail correspondant dans les évaluations des chercheurs.ses**. On peut se reporter à ce sujet au texte du comité pour la science ouverte sur les « Types de documents, productions et activités valorisées par la science ouverte et éligibles à une évaluation ».

Plus précisément, en matière d'ouverture des données, il s'agirait de prendre en compte :

- le suivi de formations en matière de gestion et d'ouverture des données ;
- le temps passé à rendre ses données pérennes et partageables ;
- les efforts d'ouverture effective des données ;
- les efforts de prise de connaissance des pratiques d'autres disciplines, permettant de comprendre quelles données l'on pourrait réutiliser ou quelles données l'on pourrait permettre à d'autres de réutiliser. Il s'agit par exemple d'inciter les spécialistes de sciences humaines et sociales à assister à des colloques hors sciences humaines et sociales, les expérimentalistes à des colloques de théorie, et vice versa – ce qui est moins valorisé en l'état que l'interdisciplinarité au sens fort (projets communs).

Préconisation E – Organiser les soutiens humains nécessaire à l'ouverture des données

Pour la science ouverte comme pour la valorisation, il est nécessaire que les **personnels connaissent à la fois leur métier propre et comprennent suffisamment les questions « pointues » propres à chaque discipline.** Cela nécessite soit une double formation en amont (métier et discipline), soit un personnel de métier qui puisse apprendre à comprendre ces questions propres à chaque discipline.

La question de l'ouverture des données demande **une coordination entre personnels de différents métiers (systèmes d'information, documentation, archives, droit, protection des données, voire qualité, etc.) plutôt que l'invention d'un nouveau métier**. Il faut inventer des manières de faire travailler ces personnels ensemble, sans systématiquement les placer dans des « pôles » isolés des laboratoires, ce qui présente des risques d'éloignement avec les pratiques de recherche. On pourrait s'inspirer sur ces questions des expériences – réussies ou non – des services de valorisation, qui posent des problèmes à bien des égards similaires ; ce serait aussi l'occasion de réfléchir sur l'articulation entre valorisation et science ouverte.

Dans la situation actuelle, le plus urgent serait sans doute **d'indiquer aux chercheurs.ses une liste de personnes-ressources**. Plus précisément, il faudrait créer un lexique de tâches liées aux données (compréhensible par les chercheurs.ses et donc élaboré avec eux et elles) et indiquer des métiers, et dans chaque établissement/département/gros laboratoire des personnes, en face de chaque tâche. Par exemple : « vous souhaitez savoir quelles parties de vos données vous devez rendre publiques : aller rencontrer un e-archiviste -> nom et coordonnées ». Les guides évoqués dans notre préconisation B pourraient inclure ce lexique des tâches renvoyant aux métiers. L'élaboration de la liste de personnes-ressources dans chaque établissement pourrait être une tâche prioritaire des administrateurs.trices des données qui doivent être nommé.es dans le cadre du PNSO. Enfin, un accompagnement juridique semble indispensable (voir paragraphe ci-dessus « à qui appartiennent les données »). Ainsi il s'agit de structurer une offre de service claire pour les chercheurs.ses en connectant l'offre existante avec le besoin.

Préconisation F – Structurer les infrastructures pour répondre aux besoins de la science ouverte

L'ouverture des données de la recherche implique des changements qui ne sont pas uniquement quantitatifs, mais aussi qualitatifs. Au niveau des établissements, les infrastructures dédiées au stockage et à la mise à disposition des données sont insuffisantes et souvent trop restrictives, rendant souvent difficile le partage des données entre les membres d'un projet appartenant à des institutions différentes.

Par ailleurs, lorsqu'une ouverture est déjà prévue, stockage et ouverture (dans le cadre du financement sur projet) sont pensés pour un temps limité. Il existe ainsi bon nombre de sites web pensés pour donner un accès à certaines données de recherche pour un large public, mais leur durée de vie est souvent très limitée. Il convient de considérer le risque en termes de recrutement pour ces fonctions si le financement est lié à un projet : personnels temporaires (perte de savoir-faire à terme) ou sous-traitance sans garantie de pérennité.

Une réelle ouverture des données implique donc des infrastructures permettant d'héberger les données pour leur mise à disposition mais aussi pour un archivage pérenne. Il existe des centres nationaux tels que le CINES, mais il faudrait s'assurer qu'ils sont capables de monter en capacité en suivant le rythme de l'extension de la science ouverte. Le stockage paraît plutôt devoir se développer par établissement, mais cela peut poser des problèmes de duplication d'efforts ainsi que de repérage (chaque type de public visé pourrait-il vraiment chercher efficacement dans l'ensemble des entrepôts ? cela semble assez utopique, en tout cas coûteux et long à mettre en place). Huma-Num, pour les sciences humaines et sociales, apparaît comme un

bon modèle d'infrastructure nationale. Mais le modèle Huma-Num est-il transposable au-delà des SHS ? Peut-il monter en puissance, du moins suffisamment pour amortir la production scientifique nationale toutes disciplines confondues ?

Les infrastructures mises en place dans les établissements présentent souvent des limites notamment en termes de stockage et d'ouverture des données. Elles ne proposent par ailleurs pas de possibilité d'archivage, sauf à mettre en place un système d'archivage électronique. Les outils mis en place ne sont pas pensés sur le long terme et il arrive trop souvent que les données mises à disposition ne soient rapidement plus accessibles.