données de la recherche : quèsaco ?

Jean-Blaise Claivaz, Université de Genève (UNIGE),
Jean-Blaise.Claivaz@unige.ch

Aude Dieudé, École Polytechnique Fédérale de Lausanne (EPFL),
Aude.Dieude@epfl.ch

Jan Krause, École Polytechnique Fédérale de Lausanne (EPFL),
Jan.Krause@epfl.ch

1. Introduction

Les données de la recherche (ci-après DR) peuvent être définies comme des « enregistrements factuels (chiffres, textes, images et sons), [...] utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche » [12]. En anglais, on parle de *research data* ou de *raw data*. L'Université d'Oxford propose la définition suivante: « *Research data [...] are defined as the recorded information (regardless of the form or the media in which they may exist) necessary to support or validate a research project's observations, findings or outputs.* »[13]

Autrement dit, les DR sont toutes les données nécessaires à la conduite d'un projet de recherche, indépendamment de leur nature. Citons par exemple les données statistiques, les mesures, les analyses, les enquêtes, les relevés topographiques, les données génomiques... Il faut aussi préciser que ces données peuvent être externes et préexistantes au projet, ou au contraire être produites par les chercheurs eux-mêmes au cours de leurs travaux.

Dans les institutions universitaires, l'importance du thème des DR grandit avec le fort accroissement de la quantité d'informations au sein de la plupart des disciplines scientifiques, ainsi qu'avec le développement de nouvelles méthodes d'analyse numérique. On parle désormais de *Data Science* [14], car toute information est appréhendée comme un élément d'un ensemble plus vaste, et la

---

[12] Une introduction à la gestion et au partage des données de la recherche. INIST. Source (le 28 janvier 2015):
http://www.inist.fr/donnees/co/module_Donnees_recherche_5.html

[13] Oxford University Research data management and open data policy, http://researchdata.ox.ac.uk

[14] Hey, T., Tansley, S., & Tolle, K. (Eds.). (2009). *The Fourth Paradigm: Data-Intensive Scientific Discovery*. ISBN 978-0982544204.

gestion de ces ensembles de données est devenue un défi majeur. Tous les champs de recherche sont concernés, mais à des échelles différentes. Certaines disciplines (génétique, météorologie, physique des particules...) travaillent avec des quantités phénoménales de données, et on parle dans ces cas précis de *Big Data* [15].

2. Problématique et enjeux

Les chercheurs sont les premiers concernés par les diverses questions que soulève le traitement des DR. Ils doivent régulièrement prendre des décisions sur les formats, les lieux de stockage ou les personnes autorisées à lire/écrire les informations. Ils sont également amenés à réfléchir au contrôle de version, à l'anonymisation des données sensibles, à l'archivage, à la diffusion/publication, ainsi qu'à la mise en œuvre de stratégies pour contrer l'obsolescence des formats et des logiciels. Ils doivent finalement s'approprier de nouveaux outils et de nouvelles technologies, par exemple les cahiers de laboratoires électroniques qui permettent de travailler collaborativement (cf. http://sv-it.epfl.ch/slims, http://fr.wikipedia.org/wiki/Système_de_gestion_de_l'information_du_laboratoire.

L'apparition du *cloud computing* apporte de nouvelles solutions, cependant celles-ci s'accompagnent d'une couche supplémentaire de problèmes, comme la question de la localisation (ou plutôt de la non-localisation) des serveurs, de la pérennité financière des sociétés du Web, de la sécurité, ainsi que la confidentialité des transactions [16].

Les agences de financement de la recherche, publiques ou privées, se positionnent de plus en plus clairement en faveur de la publication et de la réutilisation des données issues des divers projets qu'elles financent. Si les agences publiques estiment avoir aussi pour mission de rendre ouvertement accessibles les résultats financés par de l'argent public, toutes essaient d'améliorer leurs retours sur investissements, notamment en favorisant l'ouverture des données pour faciliter leur réutilisation par d'autres chercheurs et d'autres équipes. Le cas du télescope spatial Hubble illustre cette tendance, lui qui a vu le nombre de publications à son propos doubler depuis la mise à disposition des données sous licence ouverte [17]. Pour atteindre ces objectifs, les agences de financement ont ajouté une composante « gestion des données » dans les formulaires d'appel à projets. Ainsi, le Conseil Européen de la Recherche (European Research Council, ERC) a

---

[15] Smolan, R., & Erwitt, J. (2012). The human face of big data. ISBN 1454908270

[16] Hurwitz, J. (2010). Cloud computing for dummies. Wiley Publishing. ISBN 9780470484708.

[17] The Irish Times: When it comes to scientific data, sharing is caring. Source (le 22 janvier 2015): https://www.rd-alliance.org/irish-times-when-it-comes-scientific-data-sharing-caring.html

introduit dans le programme Horizon 2020 (http://ec.europa.eu/programmes/horizon2020/) le principe que 20% des projets doivent fournir un plan de gestion des données (*Data Management Plan*), et ce dans les six mois qui suivent leur acceptation. L'*ERC* se base sur l'expérience concluante qu'il a faite avec l'*open access* dans le cadre de son précédent programme de financement en matière de recherche et d'innovation [18]. Le Fonds national suisse (FNS) n'a pas encore émis de directive dans ce sens, mais son directeur, Martin Vetterli, professeur à l'EPFL, a annoncé dernièrement qu'il suivait attentivement la question [19].

La pression mise sur les chercheurs (surtout les jeunes générations) pour publier rapidement un grand nombre d'articles a un impact non négligeable sur la qualité des recherches et sur la vérification des résultats. Cette compétition augmente le risque de manipulations et de fraudes scientifiques. Comme les éditeurs et les comités de relecture peinent à détecter tous les cas frauduleux, la mise à disposition des DR en accompagnement des articles est perçue comme un moyen d'encourager l'intégrité scientifique des auteurs et de faciliter la reproductibilité des expériences [20], et donc d'en assurer la valeur scientifique [21]. Certains éditeurs demandent d'ailleurs l'accès aux DR dès le processus de peer-review [22].

Depuis que les DR sont considérées comme des objets scientifiques à part entière, leur préservation est devenue un souci pour les chercheurs ainsi que pour les gestionnaires de données (*data manager, data librarian, data curator)*. Pas très rassurante, une étude révèle que plus de 60% des liens vers les données publiées en ligne sont cassés après 10 ans [23]. Ceci souligne la nécessité impérative

---

[18] European Comission (2013). Guidelines on Data Management in Horizon 2020. Source (le 22 janvier 2013): http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[19] Présentation à la conférence *Open research data: the future of science* Mardi 28 octobre 2014.

[20] Mobley, A., Linder, S. K., Braeuer, R., Ellis, L. M., & Zwelling, L. (2013). A Survey on Data Reproducibility in Cancer Research Provides Insights into Our Limited Ability to Translate Findings from the Laboratory to the Clinic. *PLoS ONE*, *8*(5). doi:10.1371/journal.pone.0063221

[21] Nature Special: Challenges in irreproducible research. Source (le 22 janvier 2015): http://www.nature.com/nature/focus/reproducibility/

[22] Dryad. (2014). Data Archiving Policy - Dryad. Source (le 22 janvier 2015) http://datadryad.org/pages/jdap

[23] A. Pepe et al. (2014). How Do Astronomers Share Data? Reliability and Persistence of Datasets Linked in AAS Publications and a Qualitative Study of Data Practices among US Astronomers. PLoS ONE, 9(8). doi:10.1371/journal.pone.0104798

d'une meilleure gestion des données à long terme. Un marché entier s'ouvre pour de nouveaux acteurs spécialisés dans la publication de jeux de données, sous la forme de journaux spécialisés [24], de dépôts de données, ou sur le modèle d'administrations diffusant des données publiques (*Open Government Data*) (En Suisse : http://opendata.admin.ch/). Pour les DR, on parle d'*Open Research Data* (ORD). Un projet soutenu par SwissUniversities (anciennement la Conférence universitaire suisse ou CUS) les cible directement. Il s'agit d'ORD@CH (http://openresearchdata.ch) dont un des objectifs est d'indexer l'ensemble des DR diffusées en libre-accès en Suisse.

3. Besoins des chercheurs

En ce qui concerne la gestion des DR, les besoins des chercheurs sont extrêmement hétérogènes, à tel point qu'il est difficile de les généraliser, même par discipline. Chaque scientifique a ses propres pratiques, priorités et outils qui dépendent d'un certain nombre de critères, dont les principaux sont listés ci-dessous dans l'ordre alphabétique.

Critère	Commentaire
Complexité	Certaines données sont très simples et immédiatement compréhensibles, d'autres sont plus complexes et nécessitent une explication ou une description pour être appréhendées.
Confidentialité	Les données utilisées ou produites sont parfois publiques, parfois strictement privées. Cette décision ne dépend pas toujours du chercheur qui peut se voir imposer des conditions par le propriétaire des données ou par le commanditaire de l'étude.
Hétérogénéité	Les données d'un projet de recherche peuvent être homogènes (même type et même format) ou alors provenir de différentes origines et présenter des caractéristiques très diverses.
Pérennité	L'intérêt à long terme des données n'est pas toujours le même. Certaines sont irremplaçables et doivent être préservées, d'autres peuvent être générées à nouveau en cas de besoin.
Production	Il y a des expériences qui doivent produire les données nécessaires à leur réalisation, tandis que d'autres recherches s'appuient sur des données préexistantes.

[24] Par ex. *Journal of Data Science*, *Data in Brief*,...

Quantité	Certaines recherches s'appuient sur de petites quantités de données, d'autres sur des quantités phénoménales.
Réutilisation	Certaines données peuvent être utilisées dans de multiples recherches de plusieurs disciplines. Elles doivent être soigneusement décrites pour être compréhensibles en dehors du contexte de leur création.
Sensibilité	Certaines données sont sensibles et doivent être traitées de manière particulière, par exemple à travers un processus d'anonymisation, éventuellement irréversible.

Cette liste de critères n'est pas exhaustive, et il faudrait peut-être y rajouter les notions de :

- « re-production » : est-il possible de régénérer les données si elles sont perdues ?
- format : est-ce que les données sont analogiques ou numériques, et dans les deux cas, comment les encoder?
- coûts : à quel coût cette donnée a-t-elle été produite et combien coûte sa préservation ?
- aspects légaux : quelles règles institutionnelles, quelle législation nationale ou quelles licences s'appliquent, par exemple, *Creative Commons CC0* ?

4. Implication des bibliothèques dans la gestion des données – Data Life-Cycle Management (DLCM)

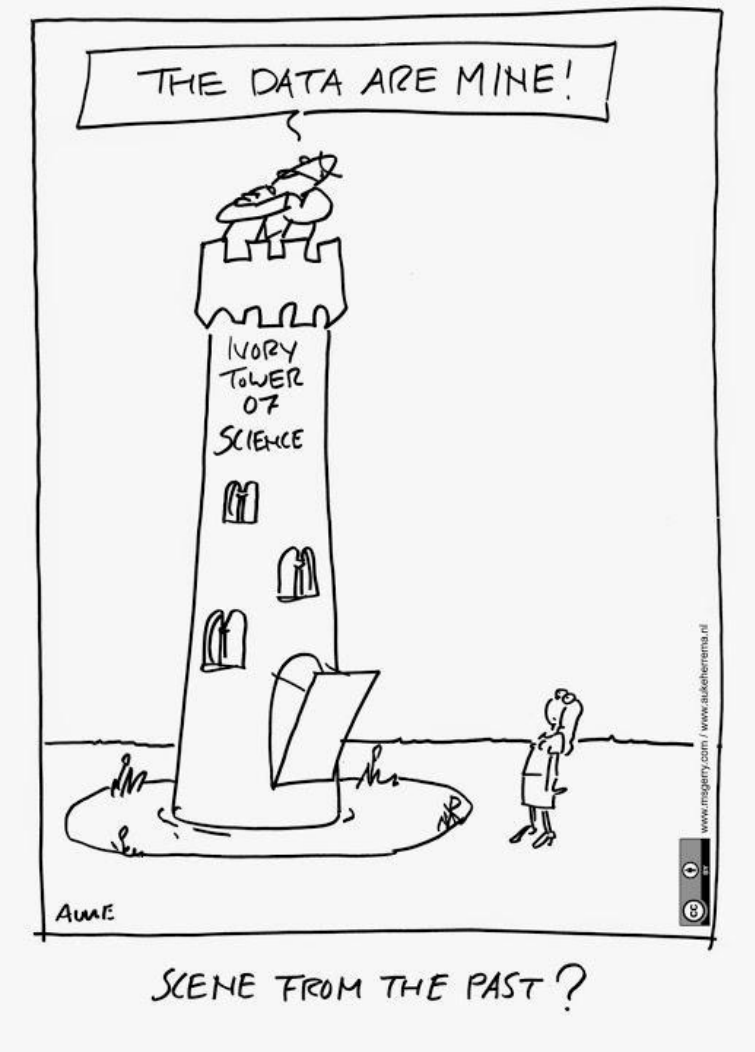
Si les besoins des chercheurs sont très hétérogènes, on ne peut pas en dire de même pour les institutions. En effet, elles sont globalement toutes confrontées aux mêmes impératifs et aux mêmes enjeux. Il paraît dès lors naturel qu'elles se regroupent pour élaborer des solutions communes telles que la planification de *Data Management Plans* (https://dmponline.dcc.ac.uk/), le travail et le partage des données actives (notamment les cahiers de laboratoires électroniques, déjà mentionnés ci-dessus), la publication et la préservation à long terme (dans des dépôts adéquats, voir http://www.re3data.org/), ainsi que la formation et la documentation des jeux de données.

Dans ce contexte, les bibliothèques peuvent apporter leur expérience en matière de gestion des métadonnées et de préservation des données sur le long terme. Pour rester dans le contexte helvétique, huit hautes écoles parmi lesquelles l'École Polytechnique Fédérale de Lausanne, la Haute École de Gestion (HEG), et l'Université de Genève, vont déposer cette année (en février 2015) dans le cadre

du programme P-2 de la CUS [25] une proposition de projet dénommée DLCM (*Data Life-Cycle Management*). Dans les grandes lignes, l'objectif du projet est de traiter de manière collaborative la question des DR dans toute sa complexité et sa singularité. Ce projet inclut une section de formation et de support aux professionnels de l'information.

5. Conclusion

Pour résumer, cet article introduit une question très complexe, qui est en plein développement et continuera certainement par la suite d'impliquer les bibliothèques universitaires. Par ailleurs, il apparaît de façon très nette, que les données de la recherche nécessitent une approche ouverte et collaborative avec les services informatiques, les directions des universités, et l'ensemble des autres acteurs institutionnels. La gestion des données de la recherche va sans aucun doute jouer un rôle de plus en plus important dans l'avenir des institutions, et a fortiori, des bibliothécaires. Affaire à suivre...



https://rd-alliance.org/plenary-meetings/fourth-plenary/plenary-cartoons.html

[25] CUS P-2 Information scientifique : accès, traitement et sauvegarde. Le programme est prévu pour les années 2014-2018.