Préface

Renaud FABRE
Professeur émérite, Université Paris 8 Vincennes-Saint-Denis, Paris, France

Aujourd'hui, les bases de véritables politiques de la donnée scientifique sont devenues visibles pour les chercheurs comme pour les utilisateurs. C'était donc le bon moment pour proposer, en version française et anglaise, un ouvrage de référence pour les travaux de recherche comme pour l'enseignement supérieur, qui fasse le point des principes, des projets et des réalisations en cours, en appliquant à nouveau la ligne éditoriale de l'encyclopédie *Sciences*, dans cette collection dédiée au « management des connaissances » : fournir un ouvrage ayant vocation à occuper une place centrale dans les références de formation initiale et continue de l'enseignement supérieur, où la « stratégie data » se développe actuellement sous l'impulsion de tous les acteurs[1].

Dans ce contexte, en effet, l'intérêt primordial de l'ouvrage, coordonné par Violaine Rebouillat et Joachim Schöpfel, est double :

– il vient au bon moment pour découvrir les projets de référence et en connaître le montage ;

– il expose, dans un large éventail de disciplines et de situations, les facettes caractéristiques de l'objectif d'une politique de la donnée, et de tout le travail de science à partir de données, de *data scholarship*, comme on le qualifie désormais (Borgman 2020).

Ces deux éclairages sont fournis par l'ouvrage de façon pertinente et complète, par et pour ceux qui construisent ces politiques de la donnée, et qui en sont en même temps les

1. Voir par exemple le DU de Sorbonne Université en formation continue : www.data-strategie.sorbonne-universite.fr.

Partage et valorisation des données de recherche,
coordonné par Joachim SCHÖPFEL et Violaine REBOUILLAT.

acteurs et utilisateurs de premier rang : les projets structurants disciplinaires ou thématiques, les formules et dispositifs nationaux de partage actuels et futurs sont passés en revue, tout comme sont abordées les problématiques du partage des données, tant pour les usages scientifiques que pour ceux de la valorisation. Des recommandations sont formulées pour l'avenir et l'on mesure ainsi que cet ouvrage nous entraîne au cœur des nouveaux enjeux d'un travail scientifique où se renouvellent, simultanément, le partage des connaissances et l'enrichissement de ses modalités.

L'accent mis par l'ouvrage sur l'entreposage des données de recherche doit être bien compris d'emblée : la fonction de l'entrepôt est en effet une fonction essentielle à tous égards dans le processus de construction du partage des données ; pas de visibilité, de comparaison, de positionnement, de curation, d'échantillonnage, de gestion du stock et de réplication sans entrepôts. On l'aura compris, l'entrepôt est la forme canonique actuelle, globale ou locale, de tout processus construit de politique de la donnée. Mais il y a plus encore : l'entrepôt est, du fait de son importance fonctionnelle évidente, le lieu même d'un apprentissage collectif des chercheurs en faveur du partage : que serait la perspective, en effet, si chaque expérience de partage, chaque pratique, devait se trouver bornée par les murs d'une approche cloisonnée à chaque laboratoire et à chaque utilisation ? On l'aura compris, l'entrepôt est le véhicule d'une nouvelle pratique scientifique, d'où l'extrême importance donnée aux pratiques d'immatriculation et de gestion des entrepôts, comme le montre le développement de la publication à ce sujet (Downs 2021) : on y reviendra plus loin dans cette introduction.

Quelques remarques sur le contexte de publication de cet ouvrage et un éclairage sur sa portée nous permettront d'illustrer notre propos.

Le contexte de publication de cet ouvrage est celui d'une politique de la donnée « à la française » qui est riche d'une dynamique de science publique, portée elle-même par un esprit de partage et d'ouverture des données scientifiques, adressé à tous les bénéficiaires et usagers de la science.

Dans ce sens, de nombreuses étapes ont d'ores et déjà été franchies ou sont en train de l'être, avec la construction d'analyses et de dispositifs sur la possibilité de partager, sur la nécessité de le faire, sur le besoin de standards de partage et d'échange. C'est le sens des grands tournants législatifs comme celui de la loi du 17 juillet 1978, qui confère aux productions assurées par la puissance publique et sous son contrôle, le caractère de « données publiques » et qui sont, sauf exceptions limitatives, susceptibles d'être communiquées sur demande. Le principe d'accessibilité, première caractéristique de l'Open Data, crée une obligation de communication et fonde la science ouverte sur une démarche d'accès au bien commun qui est, d'emblée, un droit d'accès à la donnée.

La France reste en la matière marquée par la simplicité et l'universalité de son régime d'accès à la donnée, fondé, peut-on le rappeler, sur l'article 15 de la Déclaration des droits de l'homme et du citoyen (droit à la communication des données de l'action publique), et sur la directive de 2003 du Parlement européen sur la réutilisation des informations du secteur public[2] ; l'expérience nationale française s'est récemment approfondie, avec un dispositif législatif complet de structuration de l'Open Data : la loi Valter, du 28 décembre 2015, sur la gratuité et la réutilisation des données du secteur public, puis les articles 30 et 38 de la loi pour une République numérique d'octobre 2016, pour ne citer que quelques-uns parmi les nombreux textes européens et nationaux récents sur ce thème.

Une conception stimulante de la science à l'heure numérique s'est ainsi mise en marche, parfois en partant de loin, avec des pionniers et des précurseurs ayant parfois vingt ans d'avance, et tous étant animés d'une même conviction : les données sont des objets singuliers. Les données, en effet, sont à la fois une et multiples en ce qu'elles contiennent une observation, enregistrée dans un contexte, et qu'elles peuvent, tout en même temps, appartenir à des ensembles plus vastes, ou tout simplement parallèles, où elles acquièrent une autre signification et participent à une autre réalité observable. Les données sont également singulières en ce qu'elles peuvent, pour partie, relever des pratiques d'Open Data, et pour une autre relever par contrat de dispositifs de valorisation par l'entreprise ou par le tissu associatif non marchand. Cette plasticité des données et de leur usage façonne les modalités de leur mise à disposition.

Bien entendu, l'ensemble de ces opérations de partage scientifique ont chacune leur sens, qui est celui du protocole auquel ce partage de données se rattache, mais ces partages sont aussi l'occasion d'opérations d'exposition, de vérification, de reproduction, qui correspondent aux besoins croissants de reproductibilité et de réplicabilité (Fineberg 2020).

Un ou des services nationaux en réseau pour les données (Catherine 2020) ? La réflexion s'organise, et cet ouvrage y participe, au fond, avec des options riches et positives.

Stocker est l'opération matériellement la plus facile au monde et, en même temps, scientifiquement, l'une des plus complexes. C'est vrai pour tout dépôt d'information, de toute nature, mais c'est plus difficile encore à concevoir pour les données : c'est une attitude exigeante et vigilante qui est ici requise et dont les règles sont de plus en plus clairement partagées et admises (Jeffery *et al.* 2021).

2. 2003/98/CE www.legifrance.gouv.fr/jorf/id/JORFTEXT000000521881.

Ce qui différencie l'entrepôt de données et l'« auberge espagnole » (pardon pour nos voisins, mais l'expression en français est, hélas, bien connue), c'est la présence ou l'absence de services et d'accompagnement en profondeur des besoins du chercheur. Là aussi, de nombreuses réflexions avancent avec un niveau d'exigence qui garantit un net progrès vers la maturité des projets (Suhr *et al.* 2020), qui doivent tous compter avec les progrès inégaux de l'*open access* (Hahnel *et al.* 2020) et les progrès inégaux de la formation aux outils numériques (Klebel *et al.* 2020).

En conclusion, la vertu panoramique de l'ouvrage promet un regard sûr, donné sur une progression en marche vers sa maturité, dans un domaine où aujourd'hui toute la science est sollicitée et interrogée par des entrées multiples. Reste bien entendu, comme le remarque, avec d'autres, Joachim Schöpfel dans une très précieuse publication antérieure, quelle réponse donner à la question : « Comment penser ces *data documents* qui font partie de la science en devenir et qui, de ce fait, tout naturellement, "brouillent les frontières" ? » (Schöpfel *et al.* 2020).

Rassurons-nous toutefois et trouvons dans les dynamiques actuelles d'excellentes raisons d'espérer : ce brouillage n'est que l'effet temporaire d'une recomposition du paysage de l'information scientifique dans lequel à l'amont, les entrepôts de données occupent une place stratégique fondatrice, au premier rang des sources organisées d'une nouvelle façon de faire de la science. Dans notre article commun récent, consacré aux plateformes d'information scientifiques en devenir au niveau européen et international, nous insistons sur les diverses formes que prend l'exigence de traçabilité des données et de la structuration de leur flux (Fabre *et al.* 2021). Il est temps, il est vrai, comme l'observe la Commission européenne[3], d'approfondir les approches de la science ouverte qui mettent autour d'une même table tous les acteurs et tous les utilisateurs de l'information scientifique, dans une approche qui concilie tous les usages, et qui fasse vivre ainsi tous les aspects de la science ouverte dans un déploiement de la science des données qui reste une phase critique (Davenport et Malone 2021).

Cette phase est d'autant plus critique que le développement des projets scientifiques « multi-usagers » est aujourd'hui extrêmement vigoureux, tout comme d'ailleurs celui des projets d'apprentissage collectif (He *et al.* 2020), dont un nombre fortement croissant de disciplines montrent l'usage de plus en plus répandu. Ce développement s'effectue ailleurs en ligne avec celui d'outils de partage et d'entreposage, comme les *knowledge graphs*, qui ont un développement extrêmement vigoureux dans le domaine scientifique et qui s'orientent avant tout vers les utilisations partagées de données hétérogènes, combinant ainsi des sources variées de documents et de données.

3. https://projectescape.eu/news/launch-initial-escape-esfri-science-analysis-platform-discovery-data-staging.

Bibliographie

Borgman, C.L. (2020). *Qu'est-ce que le travail scientifique des données ? Big Data, Little Data, No Data*. OpenEdition Press, Marseille.

Catherine, H. (2020). Etude comparative des services nationaux de données de recherche : facteurs de réussite. MESRI comité pour la science ouverte [Online]. Available at: https://www.ouvrirlascience.fr/etude-comparative-des-services-nationaux-de-donnees-de-recherche-facteurs-de-reussite/

Davenport, T. and Malone, K. (2021). Deployment as a critical business data science discipline. Harvard Data Science Review [Online]. Available at: https://doi.org/10.1162/99608f92.90814c32

Downs, R.R. (2021). Improving opportunities for new value of open data: Assessing and certifying research data repositories. *Data Science Journal*, 20(1), 1.

Fabre, R., Egret, D., Schöpfel, J., Azeroual, O. (2021). Evaluating scientific impact of research infrastructures: The role of current research information systems. *Quantitative Science Studies*, 1–25 [Online]. Available at: https://doi.org/10.1162/qss_a_00111

Fineberg, H., Stodden, V., Meng, X.-L. (2020). Highlights of the US National Academies Report on "Reproducibility and Replicability in Science". *Harvard Data Science Review*, 2(4) [Online]. Available at: https://doi.org/10.1162/99608f92.cb310198

Hahnel, M., McIntosh, L.D., Hyndman, A., Baynes, G., Crosas, M., Nosek, B., Shearer, K., van Selm, M., Goodey, G. (2020). The State of Open Data 2020. Digital Science Report [Online]. Available at: https://doi.org/10.6084/m9.figshare.13227875.v2

He, C., Li, S., So, J., Zeng, X., Zhang, M., Wang, H., Wang, X., Vepakomma, P., Singh, A., Qiu, H. *et al.* (2020). *FedML: A Research Library and Benchmark for Federated Machine Learning*. Preprint. ArXiv [Online]. Available at: https://arxiv.org/abs/2007.13518

Jeffery, K., Wittenburg, P., Lannom, L., Strawn, G., Biniossek, C., Betz, D., Blanchi, C. (2021). Not ready for convergence in data infrastructures. *Data Intelligence*, 3(1), 116–135.

Klebel, T., Reichmann, S., Polka, J., McDowell, G., Penfold, N., Hindle, S., Ross-Hellauer, T. (2020). Peer review and preprint policies are unclear at most major journals. *PLoS ONE*, 15(10), e0239518.

Cette bibliographie est identique à celle de l'ouvrage correspondant en anglais publié par ISTE.

Schöpfel, J., Farace, D., Prost, H., Zane, A., Hjørland, B. (2020). Data documents. *Encyclopedia of Knowledge Organization*, 48(4), 307–328 [Online]. Available at: https://www.isko.org/cyclo/data_documents

Suhr, B., Dungl, J., Stocker, A. (2020). Search, reuse and sharing of research data in materials science and engineering – A qualitative interview study. *PLoS ONE*, 15(9), e0239216.