Les données de la recherche : *définition et enjeux*

Alors que la journée d'étude du dernier congrès de l'ADBU, en septembre 2013, portait sur « Données de la recherche : quel rôle pour la documentation ? », Thierry Fournier, responsable scientifique de celle-ci, retrace ici les contours d'un domaine encore en pleine expérimentation.

La définition la plus usuelle des données de la recherche (*research data*) nous vient de l'OCDE en 2006 : « *les données de la recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons), qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche* »[1].

Cette définition est construite dans une perspective particulière : celle de l'accès aux données de la recherche *financée sur fonds publics*. Elle se situe clairement dans la lignée de la *Déclaration de Berlin sur le libre accès à la connaissance* (2003) qui stipule que « *les contributions au libre accès se composent de résultats originaux de recherches scientifiques, de données brutes et de métadonnées, de documents sources, de représentations numériques de documents picturaux et graphiques, de documents scientifiques multimédia* »[2]. Il s'agit, en gérant et en permettant l'accès le plus large et le plus commode possible aux données de la recherche, d'éviter les phénomènes de privatisation et d'appropriation bien connus en ce qui concerne les publications classiques.

Dans leur acception la plus courante, les données de la recherche ne sont pas l'ensemble des documents/enregistrements produits par les acteurs de la recherche lors d'un processus de recherche, comme pourrait les définir l'approche archivistique classique. D'une part, les données peuvent être produites en dehors du processus de recherche : elles deviennent des données de la recherche dès qu'un chercheur les utilise et, en les utilisant, se les approprie. D'autre part, se retrouvent hors périmètre de la définition habituelle les données de type managérial ou gestionnaire : programmes ou projets de recherche, analyses préliminaires, documents financiers, correspondance…, tous documents ayant bien évidemment leur importance pour comprendre comment « la science se fait », mais qui ne sont pas « nécessaires pour valider des résultats de recherche ».

Hormis les travaux théoriques, la plupart des travaux de recherche peut être vue comme des élaborations et des traitements successifs de données, sur un *continuum* qui va des données « brutes » produites par l'instrument de recherche (que celui-ci soit un capteur… ou un entretien qualitatif !) aux résultats publiés dans des articles, des livres ou de la littérature grise. Au sein de ce *continuum* de traitements successifs, la détermination de certaines données comme « nécessaires pour valider les résultats de recherche » est elle-même un enjeu scientifique et épistémologique majeur : le positionnement du curseur pour la diffusion des données ne va pas de soi.

DIVERGENCES ET CONVERGENCES

La définition courante n'épuise aucunement les questions de la nature et du périmètre des données de la recherche, qui sont très dépendantes du cadre disciplinaire dans lequel s'effectue la recherche. Les données en sociologie (enquêtes, séries statistiques…) sont très différentes des données en archéologie (rapports de fouilles…) ou en linguistique (corpus de textes…). De même, il paraît difficile d'envisager les mêmes traitements pour des données d'observation en sciences de la terre et de l'univers, pour des séries de données médicales personnalisées ou pour des logiciels de simulation qui sont un des types de données produites par les sciences informatiques.

Au-delà de leur diversité disciplinaire, les données de la recherche ont tout de même des points communs, qui justifient qu'émergent des problématiques identiques à leur sujet. Elles sont numériques, ou au moins de plus en plus numériques, et ce caractère est la condition nécessaire pour envisager leur diffusion plus ou moins systématique (ce qui ne veut pas dire non réglée). Et elles sont massives, ou au moins de plus en plus massives, ce qui induit des coûts de traitement et de stockage non négligeables.

DES DÉFIS À RELEVER…

Les enjeux et défis posés par la question de la diffusion et de l'accès aux données de la recherche financée sur fonds public sont multiples. Pour être

[1] OCDE, *Recommandation du Conseil concernant l'accès aux données de la recherche financée sur fonds publics,* 14 décembre 2006 : http://acts.oecd.org/Instruments/ShowInstrumentView.aspx?InstrumentID=159&Lang=fr&Book=False

[2] http://openaccess.mpg.de/68042/BerlinDeclaration_wsis_fr.pdf

diffusées et rendues accessibles, ces données doivent tout d'abord être collectées : un plan de gestion des données (*data management plan*)[3] commence forcément par un état des lieux ou audit des données produites. Elles doivent ensuite être structurées et documentées par un certain nombre de métadonnées, qui dépendent forcément de la nature des données traitées, et être rendues interopérables. Un autre préalable à leur diffusion est la mise au clair de leur statut juridique, avec des questions complexes d'articulation entre le droit d'auteur, le droit des bases de données, le respect des données personnelles dans certains cas, et celle, centrale, des droits de réutilisation. Les interrogations liées à leur stockage, et encore plus à leur conservation pérenne, ne sont pas superflues non plus, et peuvent entraîner des coûts conséquents bien que difficiles à chiffrer. Enfin, une diffusion correcte des données ne peut faire l'économie d'une réflexion sur l'endroit où elles sont déposées et sur la manière dont celles-ci sont liées avec d'autres (des publications, des référentiels de gestion…).

… À DIFFÉRENTS NIVEAUX D'INTERVENTION

Ainsi, les enjeux posés s'adressent potentiellement à l'ensemble des acteurs du processus de recherche (financeurs, établissements, laboratoires avec leurs équipes de recherche et leurs infrastructures) et aux réseaux disciplinaires qu'ensemble ceux-ci constituent. Pour y voir un peu plus clair, il faut peut-être distinguer, à titre indicatif et non exclusif, plusieurs niveaux d'intervention.

• *Au niveau des laboratoires et de leurs infrastructures et plateformes,* un travail interne de gestion des données produites et/ou utilisées est nécessaire et nécessite, au premier chef, l'implication des chercheurs épaulés par les ingénieurs, les informaticiens et les personnels de la documentation et de l'information scientifique et technique (IST).

• *Au niveau des établissements* peuvent être menées à la fois la réflexion juridique sur le statut des données et l'élaboration de la politique de données, qui incluent la question des réservoirs légitimes (en et hors établissement) pour le dépôt et la diffusion des données de la recherche.

• *Au niveau des réseaux disciplinaires, nationaux ou internationaux,* le travail peut consister en l'élaboration de métadonnées structurées, de normes et de standards interopérables pour les types de données principaux de la discipline, ainsi qu'en une réflexion sur l'éthique des pratiques de diffusion.

Ces trois niveaux (laboratoires, établissements, réseaux disciplinaires) doivent bien entendu être articulés pour finalement déboucher sur le résultat souhaité, à savoir l'accessibilité des données avec le minimum de barrières sur le web, ce qui implique aussi de s'interroger sur la collaboration/compétition avec des acteurs privés[4], que ce soit des éditeurs (*publishers*) classiques ou des prestataires développant des outils spécifiques[5].

Photo et design : Virgile Laguin - Illustration extraite du catalogue du congrès

À l'assaut des données de la recherche !
Le congrès 2013 de l'ADBU s'est tenu au Havre avec une journée d'étude tout spécialement consacrée aux données de la recherche.

LES CASQUETTES MULTIPLES DU *DATA LIBRARIAN*

À tous ces niveaux d'intervention, le professionnel de la documentation ou de l'IST peut avoir sa légitimité. Comme le disait Alex Ball du *Digital Curation Center*[6] britannique lors du dernier congrès de l'ADBU, ce professionnel peut/doit tenir une multitude de rôles dans le processus de gestion et de diffusion des données de la recherche :

• *détective,* pour rechercher au sein de son établissement les données « pépites » présentes nulle part ailleurs ;

[3] Cf. par exemple : **www.dcc.ac.uk/resources/data-management-plans ou www.ed.ac.uk/schools-departments/information-services/research-support/data-library/research-data-mgmt**

[4] Cf. par exemple, entre plusieurs autres : **www.nature.com/scientificdata**

[5] Cf. par exemple : **http://datadryad.org**

[6] **www.dcc.ac.uk**

• *consultant,* pour sensibiliser les chercheurs à la question des données et leur diffusion, ainsi que pour être, avec ces derniers et les informaticiens, une cheville ouvrière de la mise en place de plan(s) de gestion des données ;
• *expert,* pour aider à la structuration et à la documentation des données diffusées et participer à l'élaboration des schémas de métadonnées ;
• *conseiller,* pour accompagner les chercheurs sur les questions de droit des données ;
• *éditeur de données,* pour assurer leur diffusion effective dans les bons répertoires et réseaux, avec des identifiants permettant une liaison correcte et libre des données aux publications elles-mêmes ;
• *formateur,* pour assurer au sein de son établissement les bonnes pratiques concernant leur gestion et leur diffusion.

La légitimité d'intervention du professionnel de la documentation sur ces terrains reste encore à conquérir. Elle ne s'acquiert, ou ne s'acquerra, que par le travail concret avec les chercheurs. Pour devenir un *data librarian,* il faudra accepter de sortir de sa bibliothèque, de s'intéresser de près – au sens disciplinaire – aux recherches menées pour les comprendre ; de s'ouvrir au traitement de données et métadonnées qui ne sont traditionnellement pas les nôtres et d'accepter la porosité entre ces données de la recherche et les publications traditionnelles ; de passer de la logique de l'acquisition/traitement à celle de la co-construction. Les compétences managériales et de médiation des professionnels de l'IST leur ont globalement permis d'apporter leur pierre à la mise en place progressive de l'accès libre aux publications, *via* les archives ouvertes notamment. Le chantier des données de la recherche est du même ordre, en plus complexe. Du pain sur la planche donc !

Thierry Fournier
SCD Université Rennes 1
Commission Recherche et documentation, ADBU
thierry.fournier@univ-rennes1.fr

DES ARCHIVES AUX DONNÉES DE LA RECHERCHE

Longtemps délaissées en France, les archives de la recherche bénéficient d'une attention croissante depuis une vingtaine d'années. Une succession fortuite de commémorations au sein des organismes de recherche a permis d'engager la réflexion sur leur collecte, leur classement et leur valorisation dans les années 1980. Depuis le début des années 2000, des services d'archives intermédiaires se créent progressivement dans les universités, ouvrant la voie à une gestion de proximité des archives administratives et scientifiques de ces établissements[1].

Souvent menée en concertation avec les bibliothèques, la collecte des archives scientifiques répond à plusieurs besoins exprimés par les chercheurs : un besoin de sources historiques, mais également un besoin de continuité scientifique : en astronomie, en médecine ou encore en sciences sociales, la préservation de bases de données et de documents parfois anciens peut se révéler stratégique.

Ces nouvelles problématiques ont conduit les archivistes à se réunir au sein du réseau Aurore (Archivistes des universités, rectorats, organismes de recherche et Cité des mémoires étudiantes), devenu section de l'Association des archivistes français depuis 2011. La constitution d'un groupe de travail interne sur les archives de la recherche a permis d'élaborer des outils communs (référentiel de gestion, fiches pratiques) pour rationnaliser la collecte des documents.

La gestion, la conservation et la communication des données de la recherche constituent aujourd'hui un enjeu important pour les archivistes. La massification des données impose davantage encore d'intervenir dès la production pour mettre en place de bonnes pratiques de gestion sur l'ensemble du cycle de vie. Les données doivent s'inscrire dans un contexte pour pouvoir être exploitées à nouveau : les notions de respect des fonds et de principe de provenance qui sont à la base du travail de l'archiviste ont toujours leur place dans le monde numérique. Il appartient donc aux bibliothécaires et aux archivistes de réfléchir ensemble à ces problématiques pour constituer le patrimoine numérique de la recherche et le rendre plus accessible.

Magalie Moysan
Archiviste à l'université Paris Diderot
Membre de la section Aurore de l'Association des archivistes français
magalie.moysan@univ-paris-diderot.fr

[1] Charlotte Maday, « Les archives en université, un si long chemin... », *Arabesques,* n° 69, janvier-février-mars 2013.