Sur le terrain des données de la recherche au SCD Lille 1

Engager, en 2013, une démarche autour de l'*open access* au sein d'une université implique de lier dans un même mouvement l'enjeu des publications à celui des données de la recherche. Ce constat est à l'origine d'une étude menée par le SCD de Lille 1 auprès de laboratoires rattachés à son université.

Une première enquête réalisée en 2012 portait sur les pratiques des chercheurs en matière de dépôt de leurs publications dans une archive ouverte. La mission débutée en février 2013 se situe dans son prolongement : il s'agit de mener une étude prospective sur les données de la recherche et d'y inclure un premier état des pratiques et des besoins des chercheurs de Lille 1.

APPROCHE MÉTHODOLOGIQUE

L'université Lille 1 se caractérise par sa pluridisciplinarité. Université des « Sciences et technologies », elle intègre également les sciences économiques et la gestion, la géographie et la sociologie. Ces différentes disciplines sont étudiées au sein de 39 équipes de recherche labellisées : il paraît difficile, dans le délai imparti, de mener une étude exhaustive sur l'ensemble des laboratoires. Aussi le périmètre de l'enquête s'est limité à sept laboratoires pour lesquels nous contactons d'abord le directeur et, dans certains cas, un ou plusieurs chercheurs de l'équipe. Le choix des laboratoires repose sur plusieurs critères : l'échantillon doit être représentatif des différentes disciplines de Lille 1 et intégrer des équipes de recherche de taille variée. L'intérêt potentiel des données du laboratoire et l'existence d'initiatives préalables en matière de données de la recherche sont également pris en considération. Enfin, en nous fondant sur le résultat de l'enquête précédente, nous choisissons des directeurs de laboratoires qui se sont montrés sensibles à l'enjeu de l'*open access*.

Tous les chercheurs le disent : pour qu'un jeu de données soit réutilisable, il faut que le contexte de sa production soit documenté de manière précise.

Cibler les directeurs de laboratoire permet de les interroger à la fois en tant que chercheurs et en tant que responsables d'une structure de recherche. L'entretien est conduit autour de quatre groupes de questions sur :

- le laboratoire et ses thématiques de recherche ;
- les données produites par celui-ci[1] ;
- le recours à des sources extérieures de données ;
- les pratiques et les besoins en matière d'échange, de conservation et de diffusion des données.

PREMIERS CONTACTS

Chaque entretien commence par une définition des données de la recherche. Malgré ce préalable, mes interlocuteurs commencent souvent par évoquer les publications produites par le laboratoire, à la fois parce qu'ils relient cet entretien à l'enquête précédente sur le dépôt en archive ouverte et parce qu'ils perçoivent mieux le rôle que peut jouer la bibliothèque dans le champ des publications. Il faut parfois changer de vocabulaire : l'expression « *données brutes* », même si elle est plus restrictive, facilite en général le début de l'entretien. Viennent alors d'autres interrogations – auxquelles il faut se préparer – sur la légitimité de la bibliothèque sur le terrain des données de la recherche. Dans un laboratoire, l'évocation du lien à créer entre publications et données suscite même cette exclamation : « *ah oui, sur les publications, vous êtes tout à fait à votre place!* ». « *Je peux vous envoyer le fichier, mais vous ne parviendrez pas à l'ouvrir* ». Certains chercheurs éprouvent des difficultés à décrire leurs jeux de données à un non-spécialiste. Cette étape de description implique en effet de leur part une démarche pédagogique : il leur faut revenir sur le contexte de la recherche, décrire les étapes de production et de traitement des données, caractériser le format final du jeu de données. Tous les chercheurs ne jugent pas nécessaire d'atteindre un tel degré de précision. Il faut parfois insister en demandant à visualiser des exemples. Un directeur de laboratoire finit par me conduire auprès de chercheurs de son équipe : ceux-ci manifestent un peu d'étonnement en ouvrant devant moi leurs fichiers de données. D'autres chercheurs se prennent davantage au jeu et décrivent de manière très fine leurs jeux de données.

CONSERVER, DIFFUSER, ÉCHANGER

Interrogés sur ces trois points, les directeurs de laboratoire se montrent surtout sensibles à l'enjeu de la conservation : un chercheur en chimie évoque des données produites il y a 15 ans et qui auraient pu

être réutilisées dans un contexte différent aujourd'hui. Un autre a effectivement perdu un nombre considérable de données au format numérique et envisage de numériser la sortie papier. Certains laboratoires expriment ainsi le besoin d'un soutien dans la gestion de la conservation de leurs données.

Données conservées, données utiles ? Tous les chercheurs le disent : pour qu'un jeu de données soit réutilisable, il faut que le contexte de sa production soit documenté de manière précise. Or ce travail de documentation demande un temps qui fait défaut aux chercheurs. Le directeur d'un laboratoire de biologie évoque pourtant la mise en œuvre d'une action de conservation sur une thématique ciblée, qui concernait un grand nombre de chercheurs du laboratoire. Dans ce contexte seulement, les données ont été documentées, ce qui a permis de constituer une base d'une grande richesse. Est-elle diffusée à l'extérieur ? Dans l'absolu, ce directeur ne s'y opposerait pas, mais – et il insiste fortement sur ce point – les modalités actuelles d'évaluation d'un laboratoire de recherche ne l'incitent en rien à consacrer du temps à la diffusion de ses données.

La diffusion est le sujet où s'exprime le mieux la différence entre disciplines : en biologie et en économie, les chercheurs ont fréquemment recours à des sources extérieures de données et identifient mieux la plus-value apportée par leur diffusion. Cela ne signifie pas pour autant que le pas est facile à franchir : avec un chercheur en biologie, la discussion s'engage autour de la base TAIR[2] qu'il utilise fréquemment. Quand je lui demande s'il soumet des données dans cette base, il semble découvrir l'existence d'une procédure de dépôt. Cette possibilité l'interroge : est-ce que toutes les données déposées dans TAIR ont été validées au préalable par la publication d'un article ? Une partie des directeurs rencontrés se montre de fait intéressée par la connaissance qu'une bibliothèque pourrait apporter sur les entrepôts thématiques de données.

Dans d'autres disciplines, le recours à des sources extérieures se révèle beaucoup moins fréquent : un chercheur en chimie évoque les données expérimentales qu'il produit grâce aux instruments accessibles au sein du laboratoire. Lorsque la concurrence est forte sur une thématique donnée, l'objectif est de développer une approche originale, ce qui implique d'effectuer de nouvelles expérimentations qui aboutiront à la production de nouvelles données. Pour ce directeur de laboratoire, la publication demeure l'unité d'échange pertinente dans sa discipline : lorsqu'il lit un article, il n'éprouve pas nécessairement le besoin d'avoir accès aux données sous-jacentes.

Lorsqu'ils rencontrent ce besoin, les chercheurs utilisent souvent leur réseau pour récupérer les données de manière informelle. Même lorsqu'ils ne connaissent pas l'auteur, un échange par courriel leur suffit parfois pour accéder aux données qui les intéressent. De la même manière, ils fourniront facilement à un chercheur « de confiance » des données qu'ils ont produites. Ils savent que celui-ci connaît le contexte de leur production et qu'il aura recours au même logiciel qu'eux pour les exploiter. La crainte d'une utilisation erronée d'un jeu de données qu'ils auraient produit est commune à plusieurs des chercheurs rencontrés : « *je ne suis pas certain d'avoir envie que n'importe qui utilise mes données* », explique l'un d'eux. Là encore, le bibliothécaire doit préparer ses arguments, expliquer, par exemple, que les publications peuvent elles aussi engendrer des interprétations erronées, sans pour autant que cela remette en cause la qualité du travail initial du chercheur.

Mener des entretiens sur les données de la recherche exige ainsi d'avoir au préalable affuté ses arguments. La légitimité de la bibliothèque dans ce domaine se joue à bien des égards lors de ces premières rencontres : venu pour poser des questions, le bibliothécaire se retrouve questionné à son tour. Les réponses qu'il peut offrir dès lors sont autant de jalons posés pour un dossier sur lequel tout reste à faire.

Claudia Vojta, MPIZ Köln / Wikimedia Commons (CC BY-SA 2.0)

Culture en serres de différentes plantes du modèle **Arabidopsis Thaliana. La base de données TAIR donne accès à l'ensemble des informations génomiques et moléculaires concernant cette plante utilisée comme organisme de référence pour la recherche végétale ou fondamentale, l'évolution et la génétique.**

Marie-Madeleine Géroudet
Responsable du Service Bibliothèque numérique, SCD Lille 1
marie-madeleine.geroudet@univ-lille1.fr

[1] Pour cette partie de l'entretien, je me suis inspirée de la méthode proposée par la Purdue University Library dans le *Data Curation Profile*, disponible sur : **http://datacurationprofiles.org**

[2] *The Arabidopsis Information Resource* (TAIR) est une base de données génétiques et moléculaires sur la plante modèle *Arabidopsis Thaliana*, disponible sur : **www.arabidopsis.org**