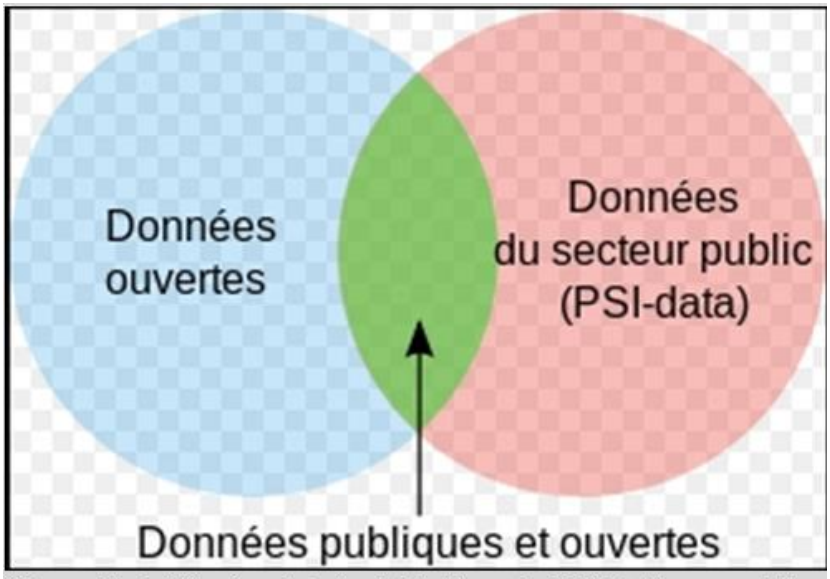
Quelles stratégies de recherche face à la nouvelle massification des données ? Colloque AEF, ADBU, 2014

Cécile Arènes

Les billets d'EnssibLab
14 avril 2015

Avec la mise à disposition d'énormes masses de données à 98% numérique, les enjeux des données sont devenus cruciaux pour l'économie numérique et les stratégies scientifiques de la recherche. Comment gérer, indexer, conserver, actionner, recouper toutes ces données ?
Comment mobiliser les professionnels de l'information et les chercheurs sur les enjeux de l'indexation à l'heure où les législations sur le TDM[1] (Text and Data Mining) suscitent des inquiétudes dans le monde de la recherche ? Comment les politiques de recherche et de l'IST peuvent-elles se repositionner dans ce contexte ? Cécile Arènes, dont le mémoire de fin d'étude du diplôme de conservateur des bibliothèques portait sur les modes de communication de la recherche[2], revient pour EnssibLab sur ces questions débattues avec les acteurs de la recherche et de l'IST, décideurs d'organismes de recherche, chercheurs, professionnels des bibliothèques et éditeurs au colloque organisé par l'AEF et l'ADBU en décembre dernier. Elle nous livre une lecture des débats orientée sur les enjeux politiques et stratégiques soulevés par l'exploitation des données de la recherche, qui impactent tout autant le travail des chercheurs et des bibliothécaires au quotidien que l'avenir des politiques publiques.

L'ouverture des données publiques : une chance pour la recherche ?



©Oppen-data-definition.fr par Peter Krantzderivative work - CC BY 3.0 - Diagramme de Venn montrant la relation entre les termes données publiques et données ouvertes.

Le mouvement en faveur de **l'ouverture des données**, porté par l'Union européenne[3], oblige le monde universitaire à s'interroger sur ses pratiques comme sur ses orientations. Lors de la journée sur les données de la recherche organisée par l'ADBU et l'AEF, Renaud Fabre, directeur de l'IST du CNRS (DIST) évoquait une nécessité d'y travailler pour ne pas rester en retrait dans ce domaine, qui va conditionner la recherche à l'avenir. La volonté politique de nombreux intervenants, acteurs de l'enseignement supérieur, s'est clairement dessinée en faveur de l'open science4. Alain Beretz, président de l'université de Strasbourg (UNISTRA) et de la League of European Research Universities (LERU) affirmait sans ambages que les résultats de la recherche financée sur fonds publics doivent être publics. La possibilité du mandat de dépôt a été rappelée plusieurs fois et Jean-Pierre Finance, délégué permanent pour la Conférence des présidents d'université à Bruxelles (CPU) n'a pas manqué de rappeler le modèle exemplaire que constitue le lancement d'Orbi par l'université de Liège. La question est vaste, elle est parfois épineuse — le débat autour de la licence Elsevier/Couperin a laissé des traces de mécontentement encore vives. Ce contexte est cependant propice, car tous les acteurs sont sensibilisés au coût de la documentation électronique et à ses conséquences sur l'accès aux résultats de la recherche publique. De fait, les professionnels de l'IST doivent être à même d'appréhender ces questions pour proposer des pistes de travail innovantes, face à un sujet complexe.

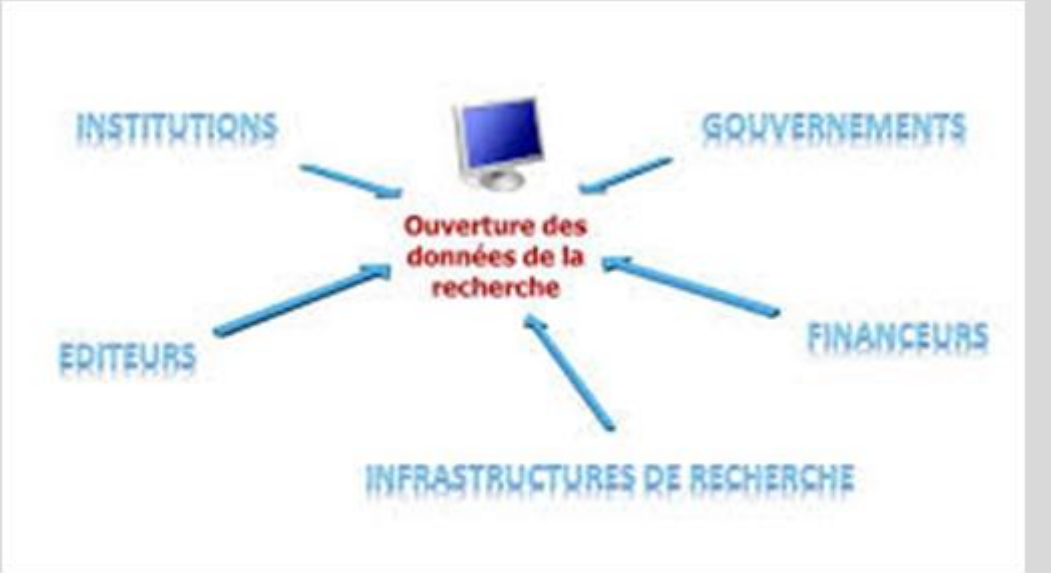
Rapprocher les enseignants-chercheurs et les professionnels de l'information autour des données de la recherche

Ils vont devoir travailler ensemble, expliquaient de concert Christophe Perales, président de l'ADBU, et Renaud Fabre, afin de mutualiser les ressources comme les compétences. Jean-François Balaudé (CPU) appelait de ses vœux un **rapprochement des enseignants-chercheurs et des professionnels de l'information**. Les données sont des objets complexes à traiter, toutefois, ainsi que le rappellait Gildas Illien, directeur de l'information bibliographique et numérique à la BNF, les enjeux autour de la question, que sont leur conservation, leur description et leur partage, ne sont pas inconnus des bibliothécaires. Avec le numérique, c'est le cycle de vie des données qui est bouleversé, car il est indispensable d'**anticiper leur préservation au moment même de leur production**, en raison de l'obsolescence rapide des supports. Mettre à disposition des référentiels dans des **formats interopérables** pour favoriser les requêtes constitue un autre axe fort, les données devant pouvoir être liées entre elles. Dans ce contexte, **il ne s'agit plus d'être un point central, il suffit d'être nœud**, car pour diffuser aujourd'hui, on ne duplique plus, on réalise des graphes.
Ces évolutions techniques nécessitent de nouveaux savoir-faire chez les professionnels de l'IST. Jean Chambaz, président de l'université Pierre-et-Marie-Curie Paris 6 (UPMC), n'a pas manqué de relever que la **formation au management de données** doit être mise en place dès maintenant, dans les cursus. Pour les chercheurs comme Françoise Genova, directrice du centre de données astronomiques de Strasbourg (CDA), les compétences disciplinaires s'acquièrent au contact des chercheurs — d'où la nécessité d'avoir des ***embedded librarians***[5], comme le soulignait Christophe Perales — tandis que la technique pure nécessite des profils spécialisés.
Plusieurs disciplines travaillent déjà à la **mutualisation des réservoirs de données**. C'est le cas en astronomie, comme l'a décrit Françoise Genova. Les astronomes utilisent de façon massive une infrastructure de données, sans point central et les requêtes proviennent de partout dans ce système complètement distribué. Même chose à l'INSERM où, à l'heure où un hôpital de taille moyenne produit plus de 600 To de données par an, on prend le problème à bras le corps, comme l'expliquait Jérôme Weinbach, directeur scientifique et opérationnel du programme national cohortes maladies rares RaDiCo.

Clarifier le cadre juridique

Toutefois, la **maîtrise technique et la mise en place de bonnes pratiques** ne suffiront pas tant que ne seront pas clarifiés un certain nombre de points d'ordre juridique. Lionel Maurel, juriste, bibliothécaire et co-fondateur du collectif SavoirsCom1, rappelait un contexte français encore très incertain sur l'exploration de données, marqué par un cadre juridique en évolution, au niveau de la fouille de données par exemple, et soulignait l'importance de se positionner pour la défense du domaine public informationnel. Sarah Jones du JISC[6] expliquait pour sa part que la législation au Royaume Uni a considéré le TDM comme une **exception au droit d'auteur dans le cadre d'une recherche à but non lucratif** et les licences recommandées par le DCC sont des Creative commons. En France, beaucoup d'éléments restent encore à clarifier, comme la question de la **propriété des données de la recherche** et la **protection des données personnelles**.

Harmoniser les pratiques institutionnelles



©Source Inist – E-learning « introduction à la gestion et au partage des données de la recherche »

Quel rôle pour les professionnels dans ce contexte ?
Sarah Jones a évoqué la nécessité de faire du lobbying contre la réforme du copyright afin d'**autoriser le TDM**, de développer la sensibilisation des chercheurs à la question des droits et de la propriété des données, enfin de **décrire les données** pour favoriser les découvertes. Les institutions doivent aussi travailler à une **harmonisation des pratiques**, ce que n'ont pas manqué de faire remarquer les éditeurs invités, Mondane Marchand (Thomson-Reuters) et Valérie Thiel-Mba (Elsevier). En effet, le signalement des chercheurs français, via leur signature dans les articles scientifiques, n'est toujours pas normalisé et donne lieu à un impact moindre en terme de **bibliométrie**. La question des données est encore largement **délaissée par les chercheurs, car elle n'entre pas dans leur évaluation et ne constitue pas un enjeu réel**.
Les débats n'ont pas manqué de porter sur le périmètre des réservoirs de données, notamment sur **l'articulation entre un dépôt d'archive national et une solution institutionnelle**. L'université de Strasbourg a d'ores et déjà préparé l'avenir des résultats de la recherche qu'elle produit. Avec le projet Archives ouvertes de la connaissance, ce sont les publications scientifiques et leurs jeux de données qui seront collectées et partagées. L'incitation de l'institution sera forte, même si le mandat de dépôt n'est pas à l'ordre du jour.
Ainsi, on peut conclure avec Francis Jutand[7] (CNN), que **l'*open data***, c'est pousser plus loin la coopération déjà amorcée avec **l'*open access***. Alain Abecassis (MENESR) rappelait d'ailleurs la détermination du ministère à consolider le mouvement, en incitant fortement à ouvrir les données, voire à éventuellement y contraindre.

Notes

[1] Le Text and Data Mining (TDM) est la technologie d'exploration de textes et de données qui permet d'extraire de manière automatique ou semi-automatique des connaissances, à partir d'énormes masses de données. Le site de veille de Lalist recence les actualités sur le TDM.
[2] ARÈNES, Cécile, 2015. *Les modes de communication de la recherche aujourd'hui : quel rôle pour les bibliothécaires ?*

[3] La législation sur l'ouverture et la réutilisation des données publiques est encadrée au niveau européen par la Directive 2003/98/CE (« Informations du secteur public » (PSI, Public Sector Information)) du Conseil de l'Union européenne du 17 novembre 2003 sur la réutilisation des données publiques, et modifiée par la Directive 2013/37/UE du Parlement européen et du Conseil du 26 juin 2013. En France, la plateforme française d'ouverture des données publiques, Data.gouv.fr, a été ouverte en 2011 par la mission interministérielle Etalab chargée de créer et d'alimenter le portail de données publiques ouvertes. Depuis 2013, l'association OpenDataFrance réunit les collectivités engagées dans l'ouverture des données publiques.
[4] Voir ce billet de blog *Big data : le boom des données numériques* pour une définition plus extensive de l'*open science*. De fait, le concept regroupe plusieurs initiatives, conduites à la fois par des chercheurs et des non-chercheurs, sur l'idée d'une recherche plus ouverte, plus transparente et plus collaborative. Cette logique vise également la multiplication des interactions entre le monde de la recherche et les sphères de la société civile, de l'entreprise ou même des arts. L'*open science* se situe entre le monde de la recherche, les technologies numériques de l'information et de la communication et les pratiques et cultures « open » émergentes. Il en découle par exemple de nouveaux modes d'organisation, pair à pair (P2P) et contributifs.
[5] Pour bien appréhender la polyvalence des compétences requises par ce nouveau métier des bibliothèques, il est intéressant de consulter la définition et l'approche de *the American Library Association* dans les cours de e-learning qu'elle propose ou encore d'un bibliothécaire américain dans son blog *The Embedded Librarian*.
[6] *Joint Information Systems Committee*, est l'organisme public britannique chargé de soutenir l'enseignement supérieur et la recherche dans l'utilisation des TIC.
[7] Directeur scientifique de l'Institut Mines Télécom et membre du Conseil national du numérique, Francis Jutand est le coordinateur de l'ouvrage collectif *La métamorphose numérique* paru en 2013 aux éditions Alternatives et dont la singularité est d'analyser du point de vue de l'homme et non celui du pure progrès technologique « les contours d'une civilisation en devenir en insistant sur les conditions requises pour en faire une civilisation fondée sur la connaissance et la coopération. »