Compte rendu des journées FRéDoc 2013

« Gestion et valorisation des données de la recherche »

Rédacteur principal du compte rendu : Julien Muller (Inist-CNRS)

Les journées FRéDoc 2013, sur le thème « Gestion et valorisation des données de la recherche », se sont déroulées du 7 au 10 octobre à Aussois. Elles n'ont pas eu pour but de rendre compte de ce sujet de façon exhaustive, mais surtout d'ouvrir des portes, des pistes de réflexion, par le partage d'une vision globale des enjeux et des évolutions à venir, et de retours d'expériences à partir de mises en œuvre concrètes dans différentes structures.

Un compte rendu collaboratif ainsi que l'ensemble des diaporamas des différentes interventions sont d'ores et déjà disponibles aux adresses suivantes :

https://docs.google.com/document/d/1H_bjt7O5ayMDU8FNmUac_3jI7VG-0fbmHkjB-T_aZEM/edit

http://renatis.cnrs.fr/spip.php?article266

Pour des informations générales sur l'actualité des données de la recherche, nous vous rappelons le site : http://www.donneesdelarecherche.fr/

L'objet de ce compte rendu n'est donc pas de reprendre le programme de ces journées à la lettre, mais de faire une synthèse des idées essentielles qui y ont été évoquées.

- **Qu'est-ce qu'une donnée ?**

En introduction à son propos, Francis André a proposé de définir ainsi ce que sont les données de recherche : ce sont des données numériques (nativement ou non) produites dans un processus de recherche, financées sur fonds publics ou réutilisées pour la recherche (ex : données publiques type enquêtes).

Selon la définition fournie par Aurélie Moriceau dans un cadre plus juridique, est considérée comme donnée la « représentation d'une information sous une forme conventionnelle », et comme recherche la « production de connaissances nouvelles ».

Une donnée n'est pas porteuse de façon aussi évidente qu'une publication de l'information qu'elle contient : c'est souvent le chercheur qui en a véritablement connaissance.

- **Vers une science des données**

Francis André l'a rappelé au début de ces journées : nous entrons depuis quelques années dans une science basée sur les données (quatrième paradigme), notamment avec la problématique des Big Data, qui concerne les très grandes masses de données qui peuvent se chiffrer en zettabytes, voire en yottabytes. Comment gérer ces flux continus de données ? L'enjeu dans la gestion des grandes masses de données est l'automatisation et le traitement par algorithme. Quelle place peut occuper le documentaliste dans ce cadre-là ? Le documentaliste détient une expérience qui peut aider à intégrer de la sémantique, mettre du sens sur une image, un objet, etc., ce qu'un algorithme ne pourra jamais faire. Il s'agit de capitaliser nos compétences en indexation et en documentation, de les transmettre et de les valoriser : cela doit constituer un *in-put* du travail des informaticiens.

Si le mouvement Big Data a le vent en poupe, Alain Collignon nous rappelle qu'il y a également une forte attente des chercheurs sur la façon de gérer les *small data*. Plusieurs réponses peuvent leur être apportées : sensibilisation sur l'open data, conseils et assistance sur les droits liés aux données (propriété intellectuelle, droit des bases de données, protection des données personnelles…), formation aux bonnes pratiques en matière de gestion des données, plan de gestion des données de la recherche.

- **Ouverture des données**

Marc Leobet nous a rappelé au cours de sa présentation de la directive INSPIRE (directive européenne visant à établir une infrastructure d'information géographique dans la communauté européenne pour favoriser la protection de l'environnement) que les données de recherche, étant financées sur fonds publics, entrent dans le cadre juridique de l'open data. Bien que les données élaborées soient protégées par la propriété intellectuelle (contrairement aux données brutes), les métadonnées doivent être diffusées et rendues disponibles. Le 3 décembre 2013, toutes les données devront en principe être cataloguées et visualisables en ligne (Géocatalogue). De même, selon la charte du G8+5 signée par François Hollande, en 2015, toutes les données publiques devront être accessibles et gratuites.

Le 24 janvier 2013, le Ministère de l'Enseignement supérieur et de la Recherche a renouvelé son soutien au libre accès des résultats de la recherche.

- **Une culture du partage / *data sharing***

Il faut développer une culture du partage. Or, il n'y a pas, à l'heure actuelle, de réponse ferme sur la reconnaissance du chercheur qui met à disposition ses données. Un certain nombre de questions se posent également comme celle du droit de propriété sur les données. On se retrouve ainsi dans une

situation où, comme nous l'a présenté Odile Hologne, 84 % des chercheurs aimeraient avoir accès aux données d'autres chercheurs, mais seulement 36 % sont prêts à mettre les leurs à disposition.

Selon Simon Hodson de CODATA, les politiques ont un rôle important à jouer pour le développement d'une culture de partage des données. Et il est essentiel d'engager les chercheurs dans l'élaboration de ces politiques dont l'un des objectifs est de protéger ceux qui font l'effort de partager leurs données. Changer de culture, cela signifie plusieurs choses :

- ne pas considérer les données comme une chose privée, mais comme un produit public de la recherche ;
- élargir les critères pour évaluer la recherche ;
- exiger l'ouverture intelligente des données appuyant les conclusions des articles scientifiques.

S. Hodson nous fournit l'exemple du Royaume-Uni où la mise en place d'une politique des données a joué un rôle important dans la valorisation des données et dans un changement de culture qui vise une meilleure reconnaissance des chercheurs qui investissent de leur temps dans les données de recherche.

- **Quelles données partager/conserver ?**

Il ressort des différentes interventions qu'il est difficile de répondre de manière univoque à cette question, et qu'elle doit être pensée au sein de chaque communauté scientifique, chaque champ disciplinaire.

Savoir quelles sont les données à conserver constitue un enjeu important. Selon quelle logique sélectionner les données ? Selon le nombre de consultations dont elles font l'objet ? Ne risque-t-on pas alors de perdre des données qui s'avèreront importantes à l'avenir ? Une des idées fortes retenues par Herbert Gruttemeier en synthèse de ces journées est que les données contiennent plus que ce qu'on en attendait au départ.

Comme l'a rappelé Mokrane Bouzeghoub, les données sont importantes non seulement pour elles-mêmes, mais également pour les relations qui existent entre elles (relations sémantiques).

- **Coopération**

La coopération est une des idées phares qui ressort de l'ensemble des interventions, et ce à plusieurs niveaux :

- Coopération entre les laboratoires au sein d'une même communauté de recherche ;
- Coopération entre les différentes disciplines (interdisciplinarité) ;
- Coopération aux plans national et international. De nombreuses initiatives pour coordonner l'action des communautés au niveau européen notamment ;
- Coopération entre les différents professionnels (scientifiques, documentalistes, informaticiens). Le professionnel de l'IST doit jouer un rôle de médiation entre informaticiens et scientifiques.

À ce titre, nous pouvons citer le projet MASTODONS. Démarré en 2010 et soutenu par les instituts du CNRS, il a pour vocation de faire émerger une communauté internationale autour de la science des données et, par la coopération entre différentes disciplines, produire de nouveaux concepts et des solutions originales.

Autre exemple : au CDS de Strasbourg, l'effectif (une trentaine de personnes) est réparti de façon équivalente entre chercheurs, documentalistes et informaticiens. En travaillant dans un environnement scientifique, les documentalistes développent des compétences disciplinaires pointues. Chaque documentaliste y travaille avec un astronome, ce qui permet d'assurer la fiabilité du traitement des données.

S. Pouyllau souligne que pour les SHS, les relations entre professionnels sont déjà largement engagées (avec bien entendu des disparités en fonction des communautés). Beaucoup d'efforts ont déjà été faits pour une transmission pédagogique de l'IST vers les chercheurs. Ce qu'il faut maintenant, c'est entretenir de la régularité dans ces relations.

- **Importance des réseaux**

Le *digital turn* (tournant numérique) implique un nécessaire changement de comportement. Stéphane Pouyllau a insisté sur ce point : la logique qui prévaut jusqu'à présent dans les laboratoires consiste à créer une multitude de plateformes de données, celles-ci se multipliant jusqu'à constituer une galaxie hétérogène qui rend le partage des données très problématique. Comme l'a rappelé Marie-Christine Jacquemot au cours de la table ronde, plutôt que de multiplier les initiatives éparses, il convient de partir de ce qui existe déjà à l'heure actuelle pour amorcer une réflexion sur la façon dont peuvent s'organiser les réseaux au sein des communautés afin de partager efficacement leurs données.

- **Qualité des données**

Un enjeu majeur de cette nouvelle science des données réside dans l'évaluation et le contrôle de la qualité des données partagées. Si comme l'a souligné Marie-Christine Jacquemot, la qualité des données relève de la responsabilité du chercheur, le professionnel de l'IST a un rôle majeur à jouer concernant la qualité des métadonnées (rôle de conseil sur les formats, la complétion des différents champs, utilisation des référentiels…) : elle est en effet indispensable non seulement à la visibilité, mais également à la bonne réutilisation des données.

« Les données ont-elles une objectivité en elles-mêmes ou sont-elles biaisées par des transformations subjectives ? ». Cette question posée par Mokrane Bouzeghoub évoque le problème de la qualité des données comme des métadonnées de façon tout à fait pertinente. Il est en effet tout à fait essentiel de s'interroger sur ce que les données produites doivent à leur contexte de production, aux théories qui les sous-tendent ainsi qu'aux outils utilisés pour les produire. La pérennité des données et leur réutilisation dans un avenir plus ou moins éloigné dépendent en partie du sérieux avec lequel nous répondons à cette question. Emmanuelle Morlock souligne l'importance de renseigner les méthodes et processus ayant produit ces données. Selon elle, il n'y a pas de réutilisation des données sans intelligibilité dans toutes leurs dimensions.

Garantir la qualité des données signifie également la mise en place de certifications et de moyens de traçabilité des données. Catherine Pequegnat, dans sa présentation du RESIF (infrastructure de recherche pour l'observation des déformations de la terre), a évoqué ce point : si les données transmises au sein de ce réseau n'ont encore jamais été corrompues à ce jour, le fait que cela soit techniquement possible pose la question suivante : « Quels moyens mettre en œuvre pour assurer la qualité des données sur le long terme ? »

Comme l'a justement souligné Marc Leobet au cours de son intervention, il n'est pas rédhibitoire pour une donnée de ne pas être totalement précise : ce qui est important, c'est que son degré de précision et de fiabilité soit renseigné afin qu'un futur réutilisateur puisse en apprécier la qualité.

- **Normalisation et standardisation**

Il est nécessaire d'harmoniser les outils et les pratiques afin de promouvoir une culture du partage. Ce n'est cependant pas toujours évident. Stéphane Pouyllau évoque ainsi la diversité des données en SHS et la difficulté qui en découle de créer une infrastructure unique pour répondre aux besoins des communautés.

La normalisation des pratiques et l'utilisation de standards partagés par les communautés restent cependant un objectif important, en cela qu'il permet un gain d'efficacité pour les chercheurs, mais également sur le plan économique.

- **Ingénierie de la connaissance**

La présentation de Claire Nedellec et Agnès Girard montre l'importance des technologies sémantiques dans l'analyse des données en étroite collaboration avec les chercheurs (construction d'une onto-terminologie pour l'analyse stratégique à des fins de gouvernance).

De même, Fabien Amarger nous a présenté un outil permettant d'agréger et d'interroger en langage naturel des informations hétérogènes à l'aide d'une base de connaissances dans le domaine agricole, et de contribuer ainsi au *linked open data*.

- **Gestion des données / Cycle de vie des données**

Le cycle de vie des données est un des points sur lesquels Stéphane Pouyllau attire notre attention. Au-delà des plateformes qui se multiplient, un enjeu majeur pour les professionnels de l'IST est la prise en charge du cycle de vie des données (stockage, conservation, pérennisation). Les documentalistes ont un rôle de sensibilisation à jouer auprès des chercheurs sur les formats de métadonnées et la gestion des données à long terme. Il faut penser à la gestion des données dès le début d'un projet.

L'attribution de DOI, identifiants pérennes pour le partage de données, attribués en France par l'Inist-CNRS dans le cadre du consortium DataCite (cf. présentation d'Herbert Gruttemeier), augmente la visibilité des données et renforce leur statut en tant que contribution scientifique. Le système des DOI, déjà bien reconnu dans le monde de l'édition académique, constitue un moyen efficace de citer les données et de les lier ainsi aux publications.

- **Top-down et bottom-up**

À plusieurs reprises, la nécessité d'une démarche combinant approche ascendante (*bottom-up*) et approche descendante (*top-down*) a été évoquée. Le programme de ces journées, composé à la fois d'interventions sur les grandes structures et les grandes orientations politiques et stratégiques, et de présentations des projets et initiatives engagés au sein de laboratoires et à un niveau plus local, entrait en résonance avec ces deux logiques qui doivent finalement se rencontrer pour tracer les contours d'une politique efficace en matière de données de recherche.

- **Évolution des métiers de l'IST**

Cela n'est pas une nouveauté, avec la gestion des données de recherche, les métiers de l'IST vont connaître des évolutions importantes.

En termes de compétences tout d'abord, différents profils commencent à émerger (data scientist, data librarian, data archivist, …) : tous requièrent, à des degrés divers, et au-delà des compétences traditionnelles des documentalistes, des connaissances informatiques et disciplinaires.

En termes de relations de travail, des évolutions semblent également voir le jour : il est nécessaire pour les documentalistes de travailler en étroite collaboration avec les chercheurs dont les données sont à traiter ainsi qu'avec les informaticiens. Selon Stéphane Pouyllau, les documentalistes doivent « pousser la porte » des laboratoires pour accompagner les chercheurs dans la structuration de leurs données. C'est une occasion unique pour les documentalistes de revenir dans le processus de la recherche.

Concernant la formation à ces nouveaux métiers, il n'y a pas, à l'heure actuelle, de formation initiale dans ce domaine. Quelques retours d'expériences permettent cependant de mettre en avant des pistes pour se préparer aux évolutions du métier. On peut citer notamment la démarche « constructiviste » présentée par Alain Collignon pour l'Inist. À l'INRA, les formations initiales étant trop « classiques » pour pouvoir répondre aux besoins, des formations continues ont été mises en place à la fois sur les dimensions management (gestion de projet, etc.) et sur des aspects plus techniques (manipulation des données, ontologies, etc.).

Dans sa présentation sur l'évolution des métiers, Odile Hologne a mis en avant un certain nombre de compétences qui doivent être maîtrisées par les professionnels de l'IST :

- La mise en œuvre de DOI
- Le conseil aux auteurs concernant la citation des données
- Le dépôt centralisé des données publiées
- Le repérage des bonnes et mauvaises pratiques des éditeurs
- L'évaluation des nouveaux entrepôts de données externes
- La certification des données
- La connaissance des normes et des standards
- L'élaboration d'une ontologie/terminologie
- L'évaluation de la qualité d'une source
- Etc.

Il y a nécessité pour les professionnels de l'IST à développer une double compétence (technique ou scientifique).

Selon Emmanuelle Morlock, le rôle du documentaliste n'est pas nécessairement de participer à chaque projet, mais de diffuser des méthodologies auprès des chercheurs.

Pour Odile Hologne, il doit s'insérer dans le processus de recherche en se positionnant dans un rôle de maintenance des référentiels pour décrire et maintenir l'interconnexion des données entre elles. Ces référentiels doivent être coconstruits avec les chercheurs.

Stéphane Pouyllau évoque le fait que la base des métiers de l'IST est de travailler sur de la donnée structurée (publication, livre, etc.). Il est donc naturel de travailler sur les données de recherche : ce qui change fondamentalement, c'est la masse de données et le fait que l'on n'a plus à faire à une donnée figée, mais à un flux qui ne s'arrête jamais.

Nicolas Limare a présenté l'initiative originale d'édition d'un journal de recherche en traitement d'images avec accès au code source et démonstration du logiciel en ligne. Selon lui, les publications scientifiques sont à réinventer. Il a soulevé la question suivante : est-ce un nouveau rôle pour les professionnels de l'IST ?

- **BSN 10**

Au cours de son intervention, Alain Colas (MESR) a rappelé que la politique des données de la recherche en France reste à construire. À court terme, sera mis en place BSN 10 (segment consacré aux données de la recherche), dont la première réunion est prévue avant fin novembre 2013. Plusieurs défis sont à soulever dans le cadre de BSN 10 :

- Définition du périmètre de l'objet « données de la recherche »
- Penser les données au plus près des usages
- Les moyens et financements
- Les infrastructures
- Signalement et valorisation des données de recherche
- Questions d'ordre juridique (propriété intellectuelle notamment)
- Quelle politique de conservation et de stockage des données
- Inscription dans le contexte international.

Annexe

Liste des intervenants :

- Fabien Amarger (*IRIT-IRSTEA*)
- Francis André (*Chargé de mission CNRS*)
- Mokrane Bouzeghoub *(DAS INS2I, Mission pour l'interdisciplinarité, CNRS, projet Mastodons)*
- Alain Colas (*MISTRD, MESR*)
- Alain Collignon (*INIST/CNRS*)
- Cristinel Diaconu (*Centre de Physique des Particules de Marseille CNRS/IN2P3 et Aix-Marseille Université*)
- Renaud Fabre *(DIST du CNRS)*
- Agnès Girard (*INRA*)
- Herbert Gruttemeier (*INIST/ CNRS*)
- Odile Hologne (*INRA*)
- Simon Hodson (*ISCU-CODATA*)
- Marie-Christine Jacquemot *(INIST/CNRS)*
- Marc Leobet (*Chargé de mission et PCE INSPIRE*)
- Soizick Lesteven (*CDS - Centre de Données astronomiques de Strasbourg*)
- Nicolas Limare (*CMLA, CNRS*)
- Aurélie Moriceau (*CECOJI*)
- Emmanuelle Morlock (*HiSoMa*)
- Claire Nedellec *(INRA)*
- Catherine Pequenat *(Institut des Sciences de la Terre, Observatoire des Sciences de l'Univers de Grenoble)*
- Stéphane Pouyllau (*TGIR Huma-Num*)
- Susan Reilly (*LIBER*)
- Roxane Silberman (*TGIR PROGEDO/Réseau Quételet*)