La huitième édition des journées de formation du Réseau national de l'information scientifique et technique (Rénatis) du CNRS a réuni à Aussois (Savoie) les acteurs du monde scientifique (chercheurs, documentalistes, administrateurs de systèmes d'information, etc.) autour de la question des données de la recherche[1]. Un partage d'expériences et de projets sur les pratiques et les compétences à développer pour assurer une mise à disposition efficace de ces données.

Un déluge de données : retour sur les journées Frédoc 2013

La définition même du périmètre d'étude auquel ces quatre journées étaient consacrées ne va pas de soi. Si les données de la recherche sont toujours (ré)utilisées dans un processus de recherche, elles peuvent ne pas être produites dans ce cadre. L'attention des congressistes s'est essentiellement portée sur les données financées sur fonds publics, ayant vocation à être partagées au bénéfice des chercheurs, mais aussi de la société tout entière (santé, environnement...). Dès octobre 2003, la *Déclaration de Berlin sur le libre accès* associe publications et données : ces dernières doivent également être « librement accessibles et compatibles », ainsi que l'a rappelé Francis André, chargé de mission sur les données de la recherche au CNRS.

Les données de la recherche ont pour autre caractéristique d'être validées, leur appartenance à un processus de recherche supposant une certification en amont. Le passage au numérique a rendu cruciales la maîtrise de la masse et la gestion des flux de données, désormais inscrites dans la problématique plus générale des données massives (*big data*). Au cours d'une présentation du dispositif Mastodons[2], M. Bouzeghoub (CNRS) relève que les enjeux actuels ne se concentrent plus sur l'aspect quantitatif, mais plutôt sur l'hétérogénéité des formats de métadonnées utilisés, sur la multiplicité de leurs origines et contextes disciplinaires, sur la vélocité des traitements et sur les possibilités d'exploitations sémantiques, défis autrement plus délicats à relever.

EScience, data-based science, data-intensive science : les qualificatifs ne manquent pas pour désigner le « quatrième paradigme de la science »[3]. La terminologie n'est pas davantage stabilisée dès lors qu'il s'agit d'en identifier les spécialistes, tour à tour nommés *data curator, data manager, data scientist* ou *data archivist* au fil des interventions.

L'INCONTOURNABLE TRIPLET : CHERCHEUR-INFORMATICIEN-PROFESSIONNEL DE L'IST

Aujourd'hui, la gestion des données de la recherche met en jeu des compétences issues de trois univers professionnels : l'informatique, l'information-documentation, la recherche.

En effet, la compétence des données de la recherche ne vient pas uniquement des professionnels de l'information scientifique et technique (IST) qui, bien qu'experts de la donnée par nature, sont plus que jamais amenés à s'associer avec des chercheurs et des informaticiens. La mise en place d'un moteur de recherche sémantique bibliographique sur les systèmes d'élevage dans le cadre du projet TriPhase (Inra) suppose ainsi une forte interaction entre ingénieurs documentaires et chercheurs, ces derniers étant les plus à même de valider les concepts qualifiant ce qu'ils produisent. Les professionnels de l'IST travaillent également avec des équipes informatiques qui mettent en œuvre les technologies nécessaires à la réalisation du projet. Les documentalistes, par leur attention particulière aux normes et aux standards et par leur rôle dans la maintenance de référentiels indispensables à l'interopérabilité, ont une vocation naturelle à guider les choix de solutions techniques en partenariat avec les équipes informatiques. Ces mêmes choix impliquent de connaître parfaitement les besoins des chercheurs, qui varient d'une discipline à l'autre : la communication avec les scientifiques est plus que jamais essentielle dans le quotidien des professionnels de l'IST. Comme le souligne H. Gruttemeier (Inist), la définition d'un véritable plan de gestion de données (ou DMP, *Data Management Plan*) suppose des échanges approfondis entre équipe documentaire et laboratoire en amont de tout projet de

La définition d'un véritable plan de gestion de données suppose des échanges approfondis entre équipe documentaire et laboratoire en amont de tout projet de recherche.

[1] Programme et supports des interventions sont disponibles à l'adresse http://renatis.cnrs.fr/spip.php?article266. Voir également le compte @Fredoc2013 sur Twitter.

[2] Mastodons, appel à projets lancé en 2012 portant sur les grandes masses de données scientifiques, est un « défi » porté par la Mission pour l'interdisciplinarité du CNRS : www.cnrs.fr/mi/spip.php?article53

[3] Dans la nouvelle *eScience* définie par Jim Gray, données et logiciels reconfigurent la manière de faire science ; il n'est plus toujours nécessaire de partir d'une théorie pour faire une découverte (cf. Tony Hey (éd.), *The fourth paradigm : Data-intensive scientific discovery*). Sur le renouvellement induit des processus éditoriaux et l'élargissement de la notion de données aux services de type logiciel, voir en particulier l'édifiante présentation par N. Limare du journal de recherche Ipol (*Image Processing On Line*) : http://renatis.cnrs.fr/IMG/pdf/LIMARE_IPOL.pdf

recherche. La gestion des données de la recherche nécessite également des compétences juridiques, lesquelles font encore bien souvent défaut : il revient au monde de l'IST non seulement de répandre les bonnes pratiques de gestion des données, mais aussi de promouvoir les règles d'usage en matière de propriété intellectuelle, ainsi que l'a rappelé S. Reilly, chargée de projets européens pour Liber (*Association of European Research Libraries*), lors de sa présentation des *Dix recommandations* du groupe de travail *eScience Reseach Data Management*[4].

À la question de la nécessité de définir un nouveau métier à part entière, les avis ont sensiblement divergé. Pour l'heure, le traitement et la diffusion des données de la recherche constituent un enjeu de veille et de formation continue aussi bien dans les domaines technologiques que juridiques et passent par un travail d'équipe impliquant chercheurs, informaticiens et ingénieurs documentalistes. Selon O. Hologne (Inra), la coordination de ce triplet doit nécessairement faire intervenir les compétences managériales et les capacités de médiation du professionnel de l'IST.

RECOMMANDATIONS, POLITIQUES DE GESTION ET ACTIONS MUTUALISÉES

Le caractère relativement pionnier des sujets abordés explique un certain manque de recul sur des problématiques aussi centrales que celles des coûts de la gestion de données, indubitablement plus élevés que le coût moyen des publications d'articles. Cependant, C. Diaconu estime l'investissement rentable à terme, le coût d'un projet de DMP équivalant à environ 1 % du coût total de fonctionnement d'un organisme de recherche, tout en entraînant une augmentation de 10 % et une amélioration de la production scientifique.

Aussi les recommandations et les actions de mutualisation sont-elles souvent neuves et en devenir. Du G8 à l'association Science Europe, en passant par RDA (*Research Data Alliance*), des groupes de travail se sont saisis dernièrement de la question stratégique des données, en diffusant des recommandations, généralement convergentes, à l'attention des communautés scientifiques. Le rapport de la Royal Society, *Science as an Open Enterprise*, cité par S. Hodson (Codata[5]), synthétise en une notion-clé, celle d'*intelligent openness,* l'exigence de données scientifiques accessibles, évaluables, intelligibles, et (ré-)utilisables[6]. Plus éclairant encore sur les tendances à l'œuvre pour stimuler les « politiques de données » (*data policies*), le rapport *Riding the Wave*[7] préconise la mise en place, à l'horizon des années 2020, d'une e-infrastructure collaborative internationale : un tel cadre devant rendre plus effectifs les efforts actuels de protection, d'accompagnement et de récompense des scientifiques ayant de bonnes pratiques. Ainsi le réservoir Dryad, en gérant la découverte, la réutilisation et la citabilité (attribution de DOI) d'une grande variété de données associées à « leurs » publications se développe-t-il déjà avec l'ambition « d'étendre le contrat social » aux données de la recherche. À l'échelle hexagonale, Alain Colas a présenté le futur segment BSN 10 comme un fer de lance du décloisonnement entre enseignement supérieur et recherche ; des cadres réglementaires plus contraignants s'appliquent déjà, tels que la directive Inspire[8] favorisant l'interopérabilité des données cartographiques publiques (*via* le Géocatalogue, point d'accès national).

Photo : Jean-Marie Feurtet

Cascade d'Aussois (73).

Le congrès d'Aussois s'est achevé sur un appel à accélérer la constitution d'une communauté interdisciplinaire de la science des données, en renforçant et en coordonnant des initiatives de mutualisation.

Jean-Marie Feurtet, *Abes*
feurtet@abes.fr
Marion Grand-Démery, *Abes*
grand-demery@abes.fr

[4] *« Offer [...] intellectual property rights advice », Ten recommendations for libraries to get started with research data management,* rapport publié en 2012 : www.libereurope.eu/sites/default/files/The%20research%20data%20group%202012%20v7%20final.pdf

[5] Codata ou Comité pour les données scientifiques et technologiques (*Committee on Data for Science and Technology*) est un acteur historique du partage interdisciplinaire des données : www.codata.org

[6] *« Open Data is data that meets the criteria of intelligent openness. Data must be accessible, useable, assessable and intelligible »,* rapport publié en juin 2012 et disponible à l'adresse : http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

[7] Rapport présenté en octobre 2010 à la Commission européenne par le groupe d'experts Données de la science : http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf

[8] http://inspire.ign.fr