

Ar(abes)ques

JANVIER - FÉVRIER - MARS 2014

DOSSIER

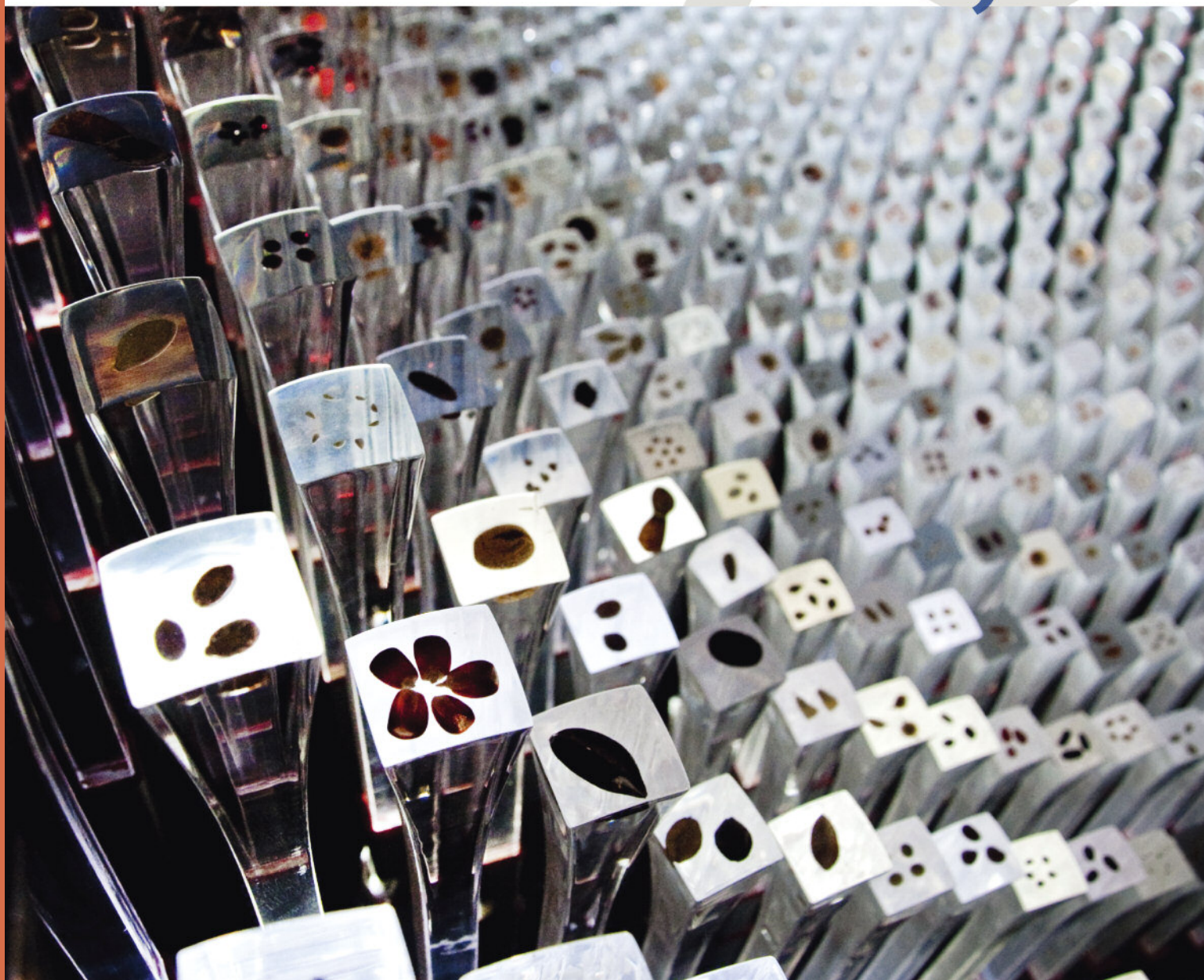
Semer, essaimer

La valorisation des données de la recherche

PLEINS FEUX SUR • La Bibliothèque Kandinsky : à la croisée des arts

ACTUALITÉS • Rencontre avec Jérôme Kalfon, directeur de l'Abes
Les chantiers de l'Abes

abes
agence bibliographique de l'enseignement supérieur



(Dossier) SEMER, ESSAIMER

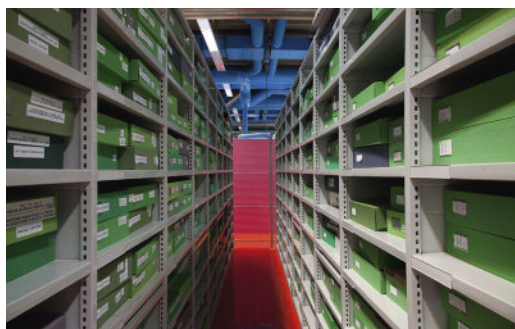
La valorisation des données de la recherche

Thème de nombreuses journées d'étude et de publications en 2013, les données de la recherche sont désormais un terreau en fertilisation tant pour les chercheurs, les informaticiens que pour les bibliothécaires coiffés d'une nouvelle casquette, celle de *data librarian*. En ouvrant ce dossier par ce qui fut le thème de la journée d'étude du dernier congrès de l'ADBU et en le refermant sur les journées Frédoc, organisées par le réseau Rénatis, nous avons voulu souligner les nombreuses réflexions et interrogations qu'elles suscitent. En donnant la parole à des acteurs politiques, des chercheurs, des bibliothécaires, des professionnels de l'information scientifique et technique et des spécialistes de l'archivage, nous avons souhaité témoigner des convergences nécessaires dans un domaine encore en pleine construction et inscrit dans la droite ligne de la *Déclaration de Berlin sur le libre accès à la connaissance* de 2003. Mutualiser pour mieux partager, tel est le défi lancé pour cette nouvelle année que nous vous souhaitons riche et fructueuse.

Merci à Thierry Fournier, coordonnateur de la commission Recherche et documentation de l'ADBU, pour sa contribution précieuse à ce dossier.

22 (Pleins feux sur...)

La Bibliothèque Kandinsky : à la croisée des arts



© Centre Pompidou-MNAM-Bibliothèque Kandinsky / Hervé Veronèse

24 (Actualités)

28 (Agenda)

04 Les données de la recherche : définition et enjeux THIERRY FOURNIER

06 Des archives aux données de la recherche MAGALIE MOYSAN

07 La *Déclaration de Berlin*, dix ans après ROGER GENET

09 Les données de la recherche des thèses JOACHIM SCHÖPFEL

10 Collectionner des données ou expérimenter : une querelle des Anciens et des Modernes ? SAMUEL GOËTA

12 Du nécessaire partage des données scientifiques : l'exemple de l'astronomie FRANÇOISE GENOVA

14 Sur le terrain des données de la recherche au SCD Lille 1 MARIE-MADELEINE GÉROUDET

16 L'Inist-CNRS face au défi des données de la recherche CATHERINE POUPON-CZYSZ

18 Archivage intermédiaire de données scientifiques au Cines MARION MASSOL / STÉPHANE COUTIN

20 Un déluge de données : retour sur les journées Frédoc 2013 JEAN-MARIE FEURTET / MARION GRAND-DÉMERY

Ar(abes)ques

REVUE TRIMESTRIELLE DE L'AGENCE BIBLIOGRAPHIQUE DE L'ENSEIGNEMENT SUPÉRIEUR
227 avenue Professeur Jean-Louis Viala CS 84308 - 34193 Montpellier cedex 5
T. 04 67 54 84 10 / F. 04 67 54 84 14 - www.abes.fr

Directeur de la publication : Jérôme Kalfon

Comité de rédaction : Dominique Esmenjaud, Christine Fleury, Marion Grand-Démery, Philippe Le Pape, Béatrice Pedot, Maryse Picard, Marie-Pierre Roux

Coordination éditoriale et secrétariat de rédaction :

Béatrice Pedot (beatricepedot@wanadoo.fr)

Conception graphique : Anne Ladevie (anne-ladevie@orange.fr)

Impression : Pure Impression

Couverture : Détail de la *Seed Cathedral* (Cathédrale des semences), pavillon britannique de l'exposition universelle de Shanghai (2010). Construite à partir du thème « Bâtir sur le passé, bâtir notre avenir », elle est composée de 60 000 capsules transparentes qui contiennent des graines du monde entier.

Phot. [Jim] / Flickr (CC BY-NC-CA 2.0)

http://www.flickr.com/photos/james_scott/5694498756/

Revue publiée sous licence Creative Commons CC BY-ND 2.0 (Paternité - Pas de modifications) sauf pour les images qui peuvent être soumises à des licences différentes ou à des copyrights.

ISSN (papier) 1269-0589 / ISSN (web) 2108-7016

D'aval en amont

L'exposition des données de la recherche va bouleverser, encore plus profondément que le basculement de la documentation du papier vers l'électronique, les métiers des bibliothécaires et des documentalistes. Plutôt que de traiter l'information une fois produite, ils seront amenés à intervenir en amont du processus de publication.

Ce processus est en marche. Ainsi pour les thèses, les bibliothécaires ne se contentent pas d'enregistrer, indexer et diffuser. Ils accompagnent le doctorant dans le processus de production de

tion de nouvelles connaissances. Il faut pour cela qu'ils soient acceptés, que leur faculté à intervenir dans ce nouveau champ soit considérée comme légitime.

Ils auront pour cela besoin de compétences de plus en plus pointues afin d'acquérir et enrichir les métadonnées, nettoyer les données pour les rendre réutilisables, participer à leur exploitation sémantique ou procéder à l'extraction de connaissances. *Database curator, data libra-*



main.

Il y a peu, et dans certains domaines encore, les données expérimentales étaient considérées comme la propriété de celui qui les avait produites. Que ce soit le G8, les financeurs de la recherche, la communauté scientifique internationale, tous s'accordent sur l'indispensable ouverture des données de la recherche. Elle s'inscrit dans un contexte plus large d'ouverture des données publiques en général. Cette volonté n'obéit pas à un principe généreux, à une volonté de gratuité telle que l'on peut la connaître dans la lecture publique. Elle répond à un besoin fonctionnel, pratique. Sans ouverture, pas de possibilité d'assemblage, de réutilisation, de construction de nouvelles connaissances.

L'un des avantages de l'exposition des données de la recherche sur les publications scientifiques, c'est qu'il n'y a pas de modèle économique préexistant.

sa thèse. Ils apportent leurs compétences en proposant des outils de structuration des documents afin que leurs métadonnées soient exploitables, conseillent les doctorants en matière de droits d'auteur et les assistent dans leurs recherches documentaires.

Avec les données de la recherche, un pas supplémentaire est franchi. La maintenance de référentiels, le respect de normes sont des éléments indispensables pour les gérer et les valoriser. Pour qu'elles soient réutilisables, il est aussi nécessaire de documenter le contexte de leur production. Le travail du professionnel de la documentation et celui du chercheur s'en trouveront intimement liés, imbriqués. Ainsi, au moment même où, avec le recul du papier, les centres de documentation tendent à disparaître des laboratoires, des spécialistes de l'information sont invités à y retourner pour participer à la construc-

tion – l'équivalent français pour ces termes n'est pas entré dans les usages –, il s'agit quasiment de nouveaux métiers que bibliothécaires et documentalistes sont bien placés pour assumer si, et seulement si, d'ambitieux plans de formation continue sont menés.

Le difficile et âpre combat pour l'accès ouvert aux publications connaît peut-être une issue favorable du fait même de l'ouverture des données. L'un des avantages de l'exposition des données de la recherche sur les publications scientifiques, c'est qu'il n'y a pas de modèle économique préexistant. Il n'y a pas à gérer de transition du papier vers l'électronique. Il n'y pas (ou peu) d'acteurs qui s'en soient (encore ?) appropriés les droits exclusifs pour en faire un centre de profit. Les principes de fonctionnement sont en cours de construction et la communauté scientifique internationale doit garder la

Les pratiques, en la matière, ont devancé le droit. Il faudra probablement légiférer, mais il s'agira là d'une tâche complexe car elle ne peut être menée que dans une perspective internationale.

L'exposition des données est l'occasion de remettre les choses à leur juste place : le principe de la libre circulation des idées et des informations est la règle ; la propriété intellectuelle, qu'il s'agisse du droit d'auteur, du copyright ou des brevets, l'exception. Une exception nécessaire, et circonscrite dans le temps, pour permettre à l'auteur une juste rémunération de son travail, mais une exception dont il ne faut pas abuser.

D'aval en amont, la pente est parfois raide. Mais que de riches et passionnantes perspectives ! Elles nous incitent à aborder avec optimisme l'année 2014, une année que l'ensemble des équipes de l'ABES vous souhaite belle et heureuse.

JÉRÔME KALFON
Directeur de l'Abes

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 le 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_wsi_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



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.

le minimum de barrières sur le web, ce qui implique aussi de s'interroger sur la collaboration/compétition avec des acteurs privés⁴, que ce soit des éditeurs (*publishers*) classiques ou des prestataires développant des outils spécifiques⁵.

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 établissements¹.

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 rationaliser 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.

La Déclaration de Berlin, 10 ans après

Dans le cadre de la conférence Berlin 11 (19 et 20 novembre 2013), Roger Genet, directeur général pour la recherche et l'innovation (ministère de l'Enseignement supérieur et de la Recherche) a rappelé l'immense chemin parcouru depuis la *Déclaration de Berlin sur le libre accès à la connaissance*. Nous publions ici de larges extraits de son intervention.

[...] Au début de notre siècle, à Budapest, à Bethesda, et surtout à Berlin, les chercheurs ont voulu dire qu'il fallait inventer de nouvelles formules pour que l'ouverture d'opportunités immenses de diffusion offertes par le numérique, et aussi par l'extension de la recherche à beaucoup de pays neufs, ne se transforme pas en son contraire : une difficulté à accéder aux résultats de la recherche pour des raisons économiques.

Quels sont ces messages venus des chercheurs ? On peut en discerner quatre majeurs :

- les fonds publics, issus des citoyens, doivent faire retour à la société ;
- la communication des résultats scientifiques fait partie des buts de la science ;
- une science plus ouverte aujourd'hui donnera une meilleure science demain ;
- une science plus accessible et plus avancée est un gage de développement culturel, économique et sociétal. [...]

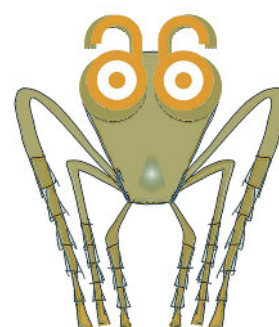
Depuis un peu plus d'un an, l'histoire s'accélère, et un nouveau seuil de visibilité et d'engagement politique a été franchi :

- par des déclarations marquantes : celle de la Commission, en juillet 2012, par le canal de sa direction générale Réseaux de communication, contenus et technologies ; celle du G8 en juillet 2013 ; de nombreuses prises de position et normes au niveau national ; la préparation d'Horizon 2020, qui prévoit d'élever au rang de principe général le libre accès aux publications et aux données de la recherche ;
- par l'arrivée au premier plan de préoccupations nouvelles, touchant à l'accès aux données de la recherche, qui élèvent le niveau d'exigence qui s'attache aux travaux de recherche, à leur publication, à leur contrôle par la communauté scientifique internationale. [...]

L'*open access* a enrichi le paysage de la communication scientifique, ce qui veut dire aussi qu'il l'a rendu plus complexe en s'ajoutant au mécanisme traditionnel de l'abonnement et en se présentant

sous diverses formes. Cette multiplicité des canaux de diffusion et des modes de financement de la communication scientifique est à l'image de la diversité des besoins de la science, parfois contradictoires : le besoin du partage le plus large entre chercheurs et, dans le même temps, celui de la reconnaissance par les revues les plus prestigieuses, qui offrent des repères nécessaires aux financeurs de la science ; la possibilité de la diffusion la plus rapide, mais aussi la nécessité de préserver l'édition scientifique. Ces choix sont difficiles ; ils expriment une tension que ressentent fortement les chercheurs eux-mêmes et les institutions. Mais ils suscitent une dynamique de coopération, de dépassement des fragmentations internationales comme nationales, dont les effets positifs s'accroissent. [...] Comme l'a dit la ministre, Geneviève Fioraso, dans sa déclaration du 24 janvier 2013, l'*open access* n'est pas une fin en soi. Il est devenu un moyen optimal de diffusion de la production scientifique qui se présente sous diverses formes : *gold* (financement des revues en amont par les auteurs), *green* (auto-archivage par les auteurs), *platinum* (financement public de plateformes permettant l'accès gratuit aux publications et l'achat de services supplémentaires). [...]

L'*open access*, ne le cachons pas, est un moyen d'inciter à plus de transparence des coûts ; il peut permettre de limiter les tentations d'abus de position dominante qui peuvent guetter certains éditeurs, lorsque leur capacité d'anticipation, à l'orée du numérique, et leur place dans les mécanismes d'évaluation de la recherche leur ont assuré une capacité unilatérale à imposer leur tarification aux chercheurs et aux institutions. Il faut rétablir la confiance, et pour cela, dissiper l'impression qu'ont parfois les chercheurs, de payer la marque d'une revue plus que son travail. Mais si le modèle de l'auteur-payeur venait à s'universaliser, il ferait peser un autre danger : un soupçon sur la neutralité de la sélection. À l'inverse, la communication scientifique directe, ou avec un filtrage minimum, si féconde soit-elle potentiellement, ne peut se substituer à l'évaluation longue par les pairs.



➔ Depuis 2003, l'*open access* a accompli un grand bond en avant.

Marina Noordegroot/Flickr (CC BY-NC-SA 2.0)

Nous essayons de tenir compte de tous ces enjeux : l'apport propre des éditeurs, l'indépendance de l'évaluation, le libre accès de toutes les équipes à la publication, et l'amélioration qu'apporte une circulation accélérée des résultats. En accord avec la Commission, la France préconise donc la mixité – autant qu'elle la constate ! –, sachant que personne ne peut aujourd'hui prédire avec certitude quel est le meilleur modèle – ni celui qui s'imposera dans une société et une économie mondialisées de la connaissance.

La coexistence peut représenter un idéal, qui évite d'avoir à choisir entre des modèles dont aucun ne s'impose avec évidence, mais, dans la pratique, elle peut révéler des incohérences ou des problèmes :

- comment éviter de payer deux fois dans un régime de mixité : comme auteur et comme lecteur ? C'est la question du *gold hybride* ;
- comment faire appliquer les règles que nous édictons ? Faut-il mettre en place une obligation de publication en archives ouvertes des résultats de projets financés, sous peine que les articles qui font défaut ne soient plus pris en compte dans les financements ou les promotions, à l'exemple de l'université de Liège ?
- comment maîtriser, dans des budgets contraints, la dépense publique, sinon par une préférence pour les voies qui ne demandent pas d'investissement public supplémentaire, notamment la voie *green*, dont on observe un gain remarquable ?

Face à ces différentes questions, la France s'efforce de mettre en œuvre une politique à la fois pragmatique et cohérente, guidée par le souci d'aboutir à des solutions adaptées, équilibrées par rapport aux principaux problèmes que rencontre la mise en œuvre de l'*open access* :

- un réalisme, qui nous a conduits à organiser le point de vue de la recherche publique sur l'évolution des dispositifs actuels de validation des publications, pour dialoguer ensuite avec les éditeurs privés, notamment sur la question des durées d'embargo ; pour les STM, un travail avec l'Agence nationale de la recherche, principale agence de financement, va fixer un plafond dans les fourchettes européennes et américaines (6 ou 12 mois). Pour les SHS, des décisions seront prises au printemps après communication des résultats d'une étude menée par un laboratoire spécialisé dans les impacts des politiques publiques ;
- une mise en œuvre à travers un plan d'action qui redynamise et met en cohérence différentes infrastructures nationales : Hal (plateforme d'archives ouvertes nationale, transdisciplinaire, qui reçoit près de 3 000 documents par mois et accueille plus de 80 archives d'institutions scientifiques) ; theses.fr (portail de consultation de thèses –

30 000 aujourd'hui –, adossé à des bases de signalement des thèses soutenues et des thèses en cours) ; OpenEdition (plateforme nationale de revues sur le mode *platinum*, qui édite plus de 380 revues en SHS, ainsi que des carnets de recherche) ; Persée (portail en libre accès de collections rétrospectives de revues de SHS, aujourd'hui plus de 140 et près de 3 millions de visites par mois) ; le Cines (entrepôt d'archivage pérenne pour toutes ces plateformes et, demain, pour les données de la recherche).

Ce plan est conduit dans le cadre de la Bibliothèque scientifique nationale (BSN), programme ministériel qui organise la coopération entre les principaux acteurs de la recherche et permet d'arbitrer les grandes questions, tout en respectant l'autonomie de ces acteurs. BSN va prochainement étendre son activité aux données de la recherche.

La nouvelle question des données de la recherche est celle du passage de l'ouverture des résultats à la « science ouverte », mais elle est aussi une partie de l'ensemble plus vaste des projets de large diffusion des données publiques.

De grands arbitrages nous attendent. Pour les réussir, il est important que nous partagions nos points de vue et nos travaux, notamment sur les questions suivantes, cruciales, parmi d'autres :

- la définition des périmètres de données à conserver, qui est variable selon les disciplines et lourde d'incidences financières ;
- la définition des conditions et standards pour des données de qualité ;
- la fixation du curseur entre les données qui devront faire l'objet d'une diffusion immédiate (notamment la mise à disposition des preuves sur lesquelles s'appuient les publications) et celles dont la propriété donne un avantage compétitif à des équipes, que l'on peut reconnaître, mais pour combien de temps ?

Au-delà des régulations et des structures, de tous les mécanismes que nos institutions et nos gouvernements pourront inventer, la seule question qui vaille et dont, *in fine*, tout dépend, c'est la volonté et la motivation des chercheurs. Ce sont eux qui, il y a dix ans, en ayant inspiré la *Déclaration de Berlin*, ont manifesté l'ambition que les scientifiques reprennent la maîtrise de la diffusion de la connaissance. Ce sont eux qui ont porté l'objectif du libre accès aux résultats de la recherche. Le nouvel horizon est celui des données de la recherche. De leurs données. La même conviction devrait nous permettre d'avancer dans ce domaine, encore plus proche du cœur de la science. Car celui qui cherche la vérité est aussi celui à qui il importe de la faire connaître.

ROGER GENET

*Directeur général pour la recherche
et l'innovation, MESR*

● ● ● LES DONNÉES DE LA RECHERCHE DES THÈSES

Avec le développement de l'eScience, l'accès aux données de la recherche devient un enjeu pour la veille scientifique. Tandis que les projets d'infrastructure portent avant tout sur de grands réservoirs de données (*big data*), d'autres résultats (*small* ou *smart data*) restent moins accessibles. Certaines données sont déposées avec les thèses électroniques. Leur intérêt est double : d'une part, leur représentativité, richesse et qualité et, d'autre part, la possibilité de les réutiliser avec d'autres résultats. Il s'agit d'annexes, d'enregistrements sonores, de matériel audiovisuel, tableaux, bases de données, résultats d'enquêtes bruts, etc. Stockées auparavant comme matériel complémentaire dans les bibliothèques, ces données sont, avec la mise en place des dispositifs comme Star, désormais déposées en format numérique. Comment ouvrir ces données, les signaler dans l'environnement des archives institutionnelles, les rendre réutilisables et exploitables, notamment pour la veille scientifique ? On peut identifier trois verrous majeurs pour la mise en œuvre de cette accessibilité.

Barrières technologiques

Souvent, les formats sont peu adaptés pour une réutilisation et/ou les métadonnées sont absentes. Les métadonnées sont pourtant essentielles pour l'interprétation, la préservation, le partage et la réutilisation des données. Un problème majeur réside dans l'absence d'un identifiant unique et pérenne comme le DOI (*Digital Object Identifier*). Quant aux formats, la Commission européenne (CE) a proposé de mettre à jour la directive de 2003 sur la réutilisation des données publiques, afin de rendre obligatoire la mise à disposition de ces données dans des formats courants, lisibles par des machines. Quelles sont les options identifiées dans le domaine des thèses électroniques ?

Obstacles juridiques

Par rapport à la diffusion et à la réutilisation des résultats de la recherche, le cadre légal des thèses paraît inadapté. Étendre la protection du droit d'auteur de la thèse électronique aux résultats (données) est en conflit avec la politique du libre accès à l'information scientifique et des données ouvertes. La CE et le gouvernement français font la promotion d'une diffusion des données publiques avec une licence ouverte minimaliste, là où les auteurs et les établissements adoptent souvent une stratégie plus restrictive (pas de modification, pas d'exploitation à

but lucratif, accès en intranet, etc.) qui entrave ou interdit toute réutilisation.

Barrières organisationnelles

Avant toute exploitation, le traitement des données nécessite un *workflow* centré sur les besoins de l'auteur-producteur et adapté aux spécificités des données (acquisition, suivi, sécurité, mise à disposition, etc.). De même, il faudra de nouveaux services et fonctionnalités pour ajouter de la valeur aux *small data* et pour faciliter leur découverte et réutilisation. À ce jour, peu d'archives institutionnelles ont commencé à développer ce genre d'outils, à l'instar des archives de données.

Vers un programme de recherche

Les verrous évoqués ci-dessus sont souvent liés. Ensemble, ils représentent des obstacles à la réutilisation des données scientifiques déposées avec les thèses. Nous avons donc mis en place un projet de recherche européen (@pic-etd : *Academic Publishing in Change-Electronic Theses and Dissertations*) pour étudier ces questions et pour instruire à la fois de bonnes pratiques (recommandations) et un cahier des charges pour faciliter l'accès aux données en fonction des disciplines, des solutions techniques et procédures existantes. Cette recherche fait partie d'un programme plus large qui inclut deux autres projets, l'un sur la publication des thèses électroniques (avec l'université de Ljubljana, Slovénie), l'autre sur leurs restrictions d'accès (avec l'Institute for Science Networking de l'université d'Oldenburg, Allemagne).

Le projet @pic-etd, retenu sur la liste complémentaire du programme blanc 2013 de l'Agence nationale de la recherche (ANR), a été redéposé dans le cadre du programme 2014. Les résultats sont attendus pour début 2014. L'Abes sera associée à la réalisation de ce projet programmé dans le cadre d'un consortium composé de :

- laboratoires de recherche : Geriico (Lille), Elico (Lyon), Dicen-IDF (Paris) ;
- services, réseaux : Abes, CCSD, ANRT, British Library, TIB Hannover, DataCite, Dini, ProQuest ;
- experts : M. Foulonneau (H. Tudor, Luxembourg), C. Leduc (ISCID-CO), L. Maurel (BDIC, Paris).

JOACHIM SCHÖPFEL

Directeur de l'Atelier national de reproduction des thèses
joachim.schopfel@univ-lille3.fr

Samuel Goëta, doctorant à Télécom ParisTech, tient, dans le cadre de son projet de thèse sur « Les coulisses de l'open data : sociologie de la production et de la libération de données publiques », un carnet de recherche en ligne^[1]. Nous reprenons ici un extrait d'un billet publié le 11 janvier 2013 à la suite d'une intervention de Bruno J. Strasser, professeur de biologie à l'université de Genève, venu présenter ses travaux sur l'histoire des données dans les sciences dans le cadre du projet Sacred^[2].

Collectionner des données ou expérimenter : une querelle des Anciens et des Modernes ?

L'essor des sciences du vivant a été accompagné par l'apparition à la Renaissance des cabinets de curiosité où étaient entreposées, classifiées et exposées des espèces hétéroclites. Amasser des plantes et des espèces naturelles était alors un divertissement commun pour la haute société de l'époque. Collectionner et montrer sa collection était un marqueur social d'érudition. Cette tradition de collection relevait principalement d'une science amateur et d'une tradition naturaliste qui aboutit au XIX^e siècle aux muséums de sciences naturelles et leurs immenses collections d'espèces. Au début du XX^e siècle, cette longue tradition déclina sous l'assaut de la science expérimentaliste qui consacre le laboratoire comme le seul lieu de la recherche scientifique.

Deux méthodes scientifiques et deux traditions épistémologiques divisent au milieu du XX^e siècle les sciences du vivant :

- les méthodes comparatives : collectionner, classifier, comparer, corréler ;
- les méthodes expérimentales : observer, analyser, généraliser à partir du cas particulier.

Pour Bruno J. Strasser, la *datadriven science* trouve ses sources dans la tradition comparative dont les pratiques et les métiers sont similaires à celle de cette « nouvelle » manière de faire de la science.

CODIFIER LE GÉNOME : UNE HISTOIRE DE BASE DE DONNÉES ET D'INDIVIDUS

Dans une période où l'expérimentation triomphe comme la seule manière de faire de la « vraie » science, le projet de codifier et de numériser l'ADN dans les années 60 marque le retour à la tradition comparative. Enregistrer une base de données, classifier et comparer des séquences de protéines ne diffère pas des pratiques de collection et de comparaison des espèces dans la science comparative. Pour Strasser, le musée et le serveur sont deux objets standardisés qui servent à produire du savoir.

Le premier projet de constitution d'une base de don-

nées massive en génétique, l'*Atlas of Protein Sequence* (1965), dirigé par Margaret Dayhoff, fut un échec du fait de la difficulté à collecter les données venant de chaque laboratoire. Dayhoff ne parvenait pas à convaincre ses collègues de diffuser les données du génome dans sa base de données en raison d'un régime de propriété intellectuelle qui, malgré un système d'accès par modem, ne permet pas la redistribution des données. Les données expérimentales sont alors un objet privé qui appartient à celui qui les a produites. Appliquant des techniques de cristallographie issues de la chimie, une discipline proche de l'industrie qui n'a pas pour habitude de diffuser ses données, le projet *Protein DataBank* lancé en 1969 ne parvient pas non plus à obtenir suffisamment de données et menace de fermer. Ce n'est finalement qu'à la fin des années 70 dans le Nouveau Mexique à l'université de Los Alamos qu'un projet de base de données génétiques parvient à décoller. Il s'agit du projet *GenBank* conduit par Walter Goad, un scientifique au parcours tumultueux qui a participé aux recherches sur la bombe H avant de concevoir ce projet, qui comporte aujourd'hui les séquences de nucléotides de près de 300 000 espèces. Quelles ont été les raisons de son succès ?

LA RECETTE DE L'OPEN SCIENCE : ÉCHANGE DE CAPITAUX SYMBOLIQUES ET APPARENCE D'OUVERTURE

Dès son lancement, *GenBank* est présenté comme un projet dans lequel l'utilisateur est aussi contributeur. Dans les années 80, ce projet réussit le tour de force de l'*open access* à une époque où le partage des données des recherches n'a rien d'une évidence. Walter Goad met en place un système vertueux dans lequel il est indispensable de partager des données pour accéder aux publications. Selon Bruno J. Strasser, le succès de *GenBank* vient de son inspiration de la philosophie des économies morales, un système dans lequel les contributions s'équilibrent pour éviter le problème du passager clandestin (*free rider*)^[3].

[1] <http://coulisses-opendata.com>

[2] www.iscc.cnrs.fr/spip.php?article1708

[3] En théorie économique, celui qui profite d'un système sans y contribuer et le mettant ainsi en péril.



↖ **Cabinet de curiosité de Ferrante Imperato, 1672.**
Entreposer et classer :
des cabinets de curiosité
aux entrepôts de données
du XXI^e siècle...

L'autre aspect du succès de *GenBank* sur lequel insiste Strasser, c'est l'apparence d'ouverture du système. « *Une force importante de votre projet est son ouverture* », écrit un ami de Goad dans une lettre. Pour obtenir le contrat qui a financé le lancement du projet en 1982, son concepteur ne cesse de donner des signes d'ouverture y compris en insistant sur la connexion du service au réseau Arpanet qui commence à relier les universités américaines. Pour Strasser, le succès de *GenBank* réside finalement dans le registre symbolique et la communication plutôt que dans la technologie du service.

NOUVELLES PRATIQUES, NOUVEAUX MÉTIERS

Avec la disponibilité de données génétiques de plus en plus importantes, de nouveaux métiers émergent, certains parlent même d'une « nouvelle espèce de scientifiques » (*a new bride of scientists*). Les *computational scientists* font partie de cette nouvelle manière de faire de la science, ni vraiment expérimentale ni vraiment comparative, qui s'emploie à analyser les données que produisent d'autres. Ils revendiquent rapidement leur statut d'auteur scientifique en proposant des publications aux revues scientifiques, qui voient d'un mauvais œil ces scientifiques qui abandonnent le microscope pour l'ordinateur en réutilisant les données mises à disposition. En 1987, le journal *American Statistics* réduit leur travail à cette expression « *Have computer, give me data* », signe d'un malaise de la communauté scientifique devant ces chercheurs qui publient en leur nom avec les données des autres.

Autre métier déconsidéré : celui de « *database curator* », en charge d'enrichir les métadonnées et de nettoyer les données pour les rendre réutilisables. Strasser raconte le témoignage d'un *database curator* qui se plaignait que personne ne comprenait son travail à un cocktail lors d'une conférence et laissait entendre qu'il n'était pas perçu comme un collègue par ses pairs. On retrouve là une réaction commune devant le travail souvent déconsidéré des « petites mains de la société de l'information », souvent jugées comme des gratte-papier ainsi que l'expliquent Jérôme Denis et David Pontille dans leur article « *Travailleurs de l'écrit, matières de l'information* »^[4].

Aujourd'hui, l'*open access* est la norme pour les publications scientifiques bien que les régimes de licence et les coûts de publication dans les principales revues forment un méli-mélo incompréhensible. Les pratiques de réutilisation de données scientifiques sont désormais courantes dans la recherche ; selon Strasser, un des prochains prix Nobel de médecine pourrait même ne « jamais avoir tenu une pipette de sa vie ». Enfin, l'*open science* questionne le rôle du chercheur : son monopole remis en cause, le modèle qui émerge rappelle celui des cabinets de curiosité à la Renaissance. En rompant avec l'emprise de la science expérimentale, il est possible d'envisager des formes de science ouvertes à tous. Par exemple, le projet Foldit^[5] se présente sous la forme d'un jeu qui permet à chacun de contribuer à l'étude de la structure des protéines en résolvant des puzzles.

SAMUEL GOËTA

Doctorant à Télécom ParisTech
samuel.goeta@telecom-paristech.fr

[4] www.cairn.info/resume.php?ID_ARTICLE=RAC_015_0001

[5] Foldit est un jeu vidéo expérimental développé en collaboration entre le département d'informatique et de biochimie de l'université de Washington : <http://fold.it/portal>

Du nécessaire partage des données scientifiques

L'exemple de l'astronomie

Les ministres de la Recherche du G8, réunis avec les présidents des académies des sciences à Londres le 12 juin 2013, ont eu des mots très forts pour soutenir le partage et la réutilisation des données de la recherche^[1]. Cette prise de position est un nouveau signe de l'attention croissante portée à ces données – et au fait que celles-ci doivent être « ouvertes » – par les gouvernements et les agences de financement de la recherche.

Certaines disciplines n'ont pas attendu que le sujet soit à la mode pour mettre en place des politiques de partage de leurs données et se mettre en ordre de marche pour prendre en charge et distribuer celles-ci. Les astronomes ont commencé dès 1976 à réfléchir à un format qui leur permette de réutiliser les images conservées sous forme électronique. Dans ce format d'échange, les métadonnées qui décrivent l'image et les conditions dans lesquelles celle-ci a été obtenue sont incluses en en-tête du fichier de données.

LES ENJEUX DU PARTAGE DES DONNÉES

Il y a des raisons scientifiques au cœur de la recherche en astronomie pour que les chercheurs souhaitent réutiliser les données. Pour comprendre les phénomènes physiques à l'œuvre, il faut le plus souvent faire appel à la combinaison d'observations de différents instruments. Les phénomènes observés sont aussi très souvent variables, d'où des comparaisons entre observations faites à différentes époques. De plus, la discipline utilise de grands instruments au sol et dans l'espace et la réutilisation des données permet d'optimiser le retour scientifique sur ces investissements.

Le format FITS (*Flexible Image Transport System*), dont discutaient les radioastronomes dans les années 70, est toujours l'un des outils de base de l'astronomie. C'est un bien commun de la discipline, régulièrement mis à jour et maintenu sous l'égide de l'Union astronomique internationale : le partage des données a été dès l'origine conçu pour lever les barrières entre les instruments d'observation, mais il se joue aussi des frontières. Grâce à FITS, tout astronome peut visualiser et utiliser les données de n'importe quel télescope. L'existence d'un format commun a aussi permis de développer des outils partagés, qui peuvent être utilisés par l'ensemble de la communauté pour manipuler les données. Les astronomes ne se sont pas arrêtés au partage des données d'observation. Ils se sont aussi inté-

ressés très tôt au partage des résultats de la recherche. Le Centre de données astronomiques de Strasbourg (CDS) a été créé dès 1972 par l'Institut national d'astronomie et de géophysique, prédécesseur de l'Institut national des sciences de l'univers du CNRS, en collaboration avec l'université Louis-Pasteur, devenue l'université de Strasbourg. Sa mission est de collecter l'information « utile » sur les données astronomiques, de les améliorer par des évaluations critiques et des comparaisons, de les distribuer à la communauté astronomique internationale et de conduire des recherches utilisant ces données. Une partie de la collecte d'information repose sur le dépouillement des articles publiés dans les journaux académiques de la discipline, d'où l'on extrait entre autres la liste des objets cités dans l'article, et des informations sur ces objets : par exemple, le nom qui leur est donné dans l'article (les objets astronomiques peuvent avoir de nombreux noms différents) ou leur position dans le ciel. Le CDS collecte aussi les grands catalogues des objets astronomiques et des images de référence du ciel et, en collaboration avec les journaux et les auteurs des articles, les tables publiées dans les journaux, ainsi que, de plus en plus, d'autres données (images, séries temporelles...) attachées aux articles.

UNE CONSTELLATION DE COMPÉTENCES ET DE PROFILS

Il est important de noter que le CDS a été créé au sein d'une structure de recherche, l'Observatoire astronomique de Strasbourg (l'actuelle UMR 7550 Université de Strasbourg/CNRS). Il inclut des scientifiques qui mènent leurs propres programmes de recherche en plus de leurs tâches au service de la communauté. Ce contact constant avec la recherche vivante est certainement l'une des clés du succès du CDS sur le long terme, puisqu'il donne un regard direct sur les besoins des chercheurs qui utilisent les services, qui sont donc correctement pris en compte, voire anticipés. Un autre aspect des com-

[1] www.gov.uk/government/news/g8-science-ministers-statement

[2] Christian Huitema, *Et Dieu créa l'Internet*, Eyrolles, 1996.

[3] *Riding the wave : How Europe can gain from the rising tide of scientific data*, European Union, 2010, <http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf>

pétences indispensables porte sur la construction du contenu : identifier et valider des données extraites des publications, construire les métadonnées permettant de décrire le contenu d'une table. C'est le travail des documentalistes du Centre de données, qui ont des compétences de pointe sur, par exemple, les objets astronomiques ou les quantités présentes dans les tables des articles. Enfin, les informaticiens de l'équipe développent les bases de données et les interfaces utilisateurs et ont une activité soutenue de veille technologique et méthodologique, ce qui permet de détecter rapidement les nouveautés susceptibles d'améliorer le service rendu aux utilisateurs.

Sur cette longue durée, le CDS a vécu plusieurs révolutions technologiques majeures : l'arrivée du *World Wide Web* vers 1993 en est évidemment une, mais on peut aussi citer le fait que la première liaison Internet entre la France et les États-Unis, en 1988, a été mise en place pour une démonstration d'une base de données du CDS aux États-Unis². Pour la distribution des catalogues, on sera passé de l'envoi par la poste de cartes perforées, de disquettes ou de bandes magnétiques, sur commande reçue également par courrier postal, à la mise à disposition des fichiers par ftp, ainsi que dans une base de données pour permettre la recherche des éléments individuels de chaque table. Ce qui a été constant depuis l'origine, c'est la quête inlassable de la meilleure qualité possible pour le contenu et de la meilleure réponse possible aux besoins des utilisateurs. Les documentalistes jouent évidemment un rôle majeur sur le long terme pour assurer la qualité du contenu. L'équipe fonctionne en équipe intégrée et ses membres, avec leurs différents profils, collaborent étroitement dans les activités journalières et dans la définition de la stratégie de développement. Le succès, sur le long terme, est au rendez-vous puisque le CDS est labellisé « infrastructure de recherche » par le ministère de l'Enseignement supérieur et de la Recherche, et que ses services reçoivent plus d'un million de requêtes par jour en moyenne.

VERS UNE INFRASTRUCTURE INTERNATIONALE ET INTERDISCIPLINAIRE

Les astronomes utilisent quotidiennement les données disponibles dans les archives des observa-

toires, les services du CDS, la base de données bibliographique maintenue au Smithsonian Astronomical Observatory avec l'aide de la Nasa, etc. Depuis plus de 10 ans, la communauté construit aussi l'Observatoire virtuel astronomique, qui vise à donner accès de façon transparente à l'ensemble des données en ligne : l'utilisation de protocoles et de vocabulaires standardisés permet de trouver les données disponibles et d'y accéder avec des outils communs, sans devoir passer par les interfaces propres à chaque service. Et les outils sont eux-mêmes interopérables entre eux.

De nombreuses autres disciplines scientifiques suivent des chemins similaires et mettent en place des politiques de conservation et de partage des données. Les agences de financement demandent aussi de plus en plus que les résultats de la recherche soient disponibles. Le dépôt des données dans un « référitoire » institutionnel permet de répondre à cette obligation, mais il ne permet pas forcément que les données soient trouvables et réutilisables. Pour définir des standards d'échange et des métadonnées, il faut un travail au niveau des disciplines, où se trouve la connaissance des données et celle du fonctionnement de la communauté. Ces « piliers disciplinaires » serviront de base à l'« infrastructure collaborative des données » internationale et interdisciplinaire préconisée par le *High Level Expert Group on Scientific Data* mis en place par la Commission européenne en 2010³.

FRANÇOISE GENOVA

Directrice du Centre de données astronomiques de
Strasbourg
francoise.genova@astro.unistra.fr

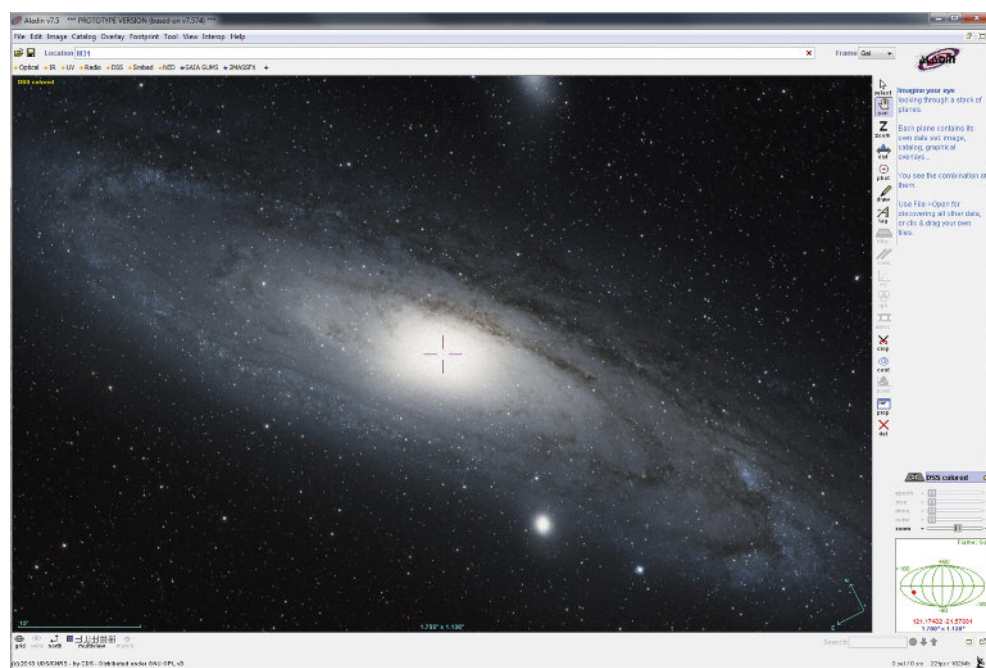


Image du Digitized Sky Survey, STScI/Nasa

La galaxie d'Andromède, visualisée avec le logiciel Aladin du Centre de données astronomiques de Strasbourg.



POUR EN SAVOIR PLUS

SITE DU CENTRE
DE DONNÉES
ASTRONOMIQUES
DE STRASBOURG :
<http://cdsweb.unistra.fr>

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 diffé-

- 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

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.

rentes 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.

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-ci¹ ;

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 TAIR² 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



Claudia Vojta, MPZ 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.

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 affûté 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.

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

En favorisant la génération de volumes considérables de données, à la fois manipulables et exploitables, l'évolution actuelle des technologies a permis l'émergence d'un quatrième paradigme : celui de la science guidée par les données. S'y ajoute l'effacement des barrières géographiques et disciplinaires au profit d'une science mondialisée et transdisciplinaire. Tant technologiques que scientifiques, ces évolutions s'inscrivent enfin dans la mouvance *open data*, incitation à l'ouverture des données publiques, renforcée à l'échelle européenne par le programme cadre de recherche « Horizon 2020 ».

L'Inist-CNRS face au défi des données de la recherche

Les pratiques et usages des chercheurs ont évolué en conséquence, favorisant l'émergence de nouveaux besoins, qui témoignent du passage d'une logique de services centrés sur les publications à une nouvelle gamme d'objets numériques, parmi lesquels les données de la recherche figurent en bonne place.

Ce contexte global a conduit l'Institut de l'information scientifique et technique (Inist), unité propre de service du CNRS – reconnue pour ses activités traditionnelles de fourniture de documents et de production de bases de données bibliographiques – à entamer une vaste réflexion prospective sur l'extension de son périmètre opérationnel dans le but d'y inclure des services de gestion et de valorisation des données de la recherche.

MISE EN PLACE D'UNE ORGANISATION DÉDIÉE

Créée en septembre 2012, une cellule de coordination « Données de la recherche » composée de deux ingénieurs, vise à assurer la gestion et l'harmonisation des activités de l'Inist relatives aux données de la recherche. Ces deux coordinateurs animent un groupe de huit correspondants-documentalistes issus des services scientifiques en sciences, techniques et médecine (STM) et sciences humaines et sociales (SHS) dont le rôle est de diffuser l'état d'avancement des travaux de la cellule au sein de leurs services respectifs.

Ils gèrent parallèlement les relations entre les différents pôles métiers de l'Institut et interviennent sur le plan national et international avec les partenaires externes dans les activités et projets liés aux données de la recherche.

RENFORCEMENT DES COMPÉTENCES : UNE APPROCHE CONSTRUCTIVISTE

À l'instar de nombreux pays, la recherche française souffre de fortes réductions budgétaires qui contraignent au redéploiement du personnel en place plutôt qu'à des recrutements supplémentaires. Bien qu'en

France aucune offre de formation continue ne soit encore disponible, l'expérience d'autres pays plus avancés en la matière comme le Royaume-Uni ou les États-Unis s'avère en revanche riche d'enseignement, exploitable notamment dans des méthodes d'apprentissage constructiviste ou de type projet. Une étude de faisabilité a déterminé les nouveaux besoins des chercheurs en matière de gestion de données de la recherche, ainsi que les connaissances et compétences nécessaires pour y répondre. Elle a permis aux coordinateurs, qui bénéficiaient d'une expérience acquise sur de précédents projets, d'initier une approche de formation continue au sein de la cellule.

Les correspondants continuent pour leur part à être impliqués dans l'indexation des bases de données bibliographiques Pascal et Francis, tout comme dans les activités liées à la terminologie, en consacrant désormais près d'un tiers de leur temps de travail à l'apprentissage de nouvelles compétences. Ces spécialistes de l'information développent progressivement une culture des données dans leurs domaines d'expertise scientifique, destinée à court terme à guider les chercheurs en matière de bonnes pratiques (entrepôts de données, standards de métadonnées, outils et terminologie...).

En parallèle, certains d'entre eux participent actuellement à des projets pilotes de gestion de données de recherche au cours desquels ils bénéficient d'une période d'immersion dans les laboratoires et acquièrent par là même une meilleure compréhension du processus de recherche.

Cette approche méthodologique visant à accompagner l'évolution du métier de documentaliste a été présentée à Munich lors de la 42^e conférence annuelle¹ de la Ligue des bibliothèques européennes de recherche (Liber).

PARTENARIATS NATIONAUX ET INTERNATIONAUX

L'Inist a récemment établi des coopérations sur le plan local avec deux unités mixtes de recherche

[1] www.liber2013.de

[2] www.sim4rdm.eu

[3] <http://rbdd.cnrs.fr>

[4]

www.donneesdelarecherche.fr

[5] www.ortolang.fr
(ANR-11-EQPX-0032)

[6] www.datacite.org

[7] Contact : datasets@inist.fr



© Phot. Georges Fessy

CNRS-Université de Lorraine :

- une application de gestion des matériels d'étude propres à un laboratoire de biologie moléculaire, qui permettra de faciliter les échanges et la réutilisation des données ;
- une étude de cas réalisée dans le cadre d'un projet européen *Support Infrastructure Models for Research Data Management (SIM4RDM)*² amenant le laboratoire à adopter les bonnes pratiques de gestion de données.

Au niveau national, l'Inist collabore avec le Réseau Bases de données (R-BDD)³, dont la mission consiste à valoriser les bases de données produites par les communautés de recherche en les rendant interopérables sur les plans technique et sémantique. L'Institut héberge et alimente également une plateforme d'information et de veille⁴ sur les données de la recherche, créée en 2012 à l'initiative du ministère de l'Enseignement supérieur et de la Recherche. Par ailleurs, depuis 2011, l'Inist est membre de l'Équipex Ortolang (*Open Resources and TOols for LANGUAGE*)⁵, validé dans le cadre des Investissements d'avenir. Son but est de proposer une infrastructure en réseau offrant un réservoir de données (corpus, lexiques, dictionnaires, etc.) et d'outils sur la langue et son traitement disponibles et documentés. Cette infrastructure poursuit un double objectif :

- la mutualisation d'acquis et le partage de connaissances sur le français et les langues de France pour les communautés de linguistes et de traitement automatique du langage ;
- l'usage et le transfert des ressources et outils mis en place au sein des laboratoires publics vers les partenaires industriels.

L'Inist assure le rôle d'opérateur technique (infrastructure) pour l'hébergement et l'exploitation de la plateforme. De plus, il participe à la définition de la structure des métadonnées attendues, préparant ainsi sa fonction de maintenance des corpus déposés dans le temps.

Au niveau international, l'Inist est membre du consortium DataCite⁶, qui a comme objectif de valoriser les données de la recherche et les informations non textuelles dans la communication scientifique. L'Institut est à ce titre habilité à attribuer des DOI (*Digital Object Identifier*)⁷, identifiant pérenne qui facilite la citabilité contribuant ainsi à une meilleure visibilité et accessibilité des données.

DE NOUVELLES ORIENTATIONS STRATÉGIQUES

Soucieux de relever le défi des évolutions technologiques et scientifiques, l'Inist a mis en place une organisation fonctionnelle spécifiquement dédiée. Accompagner l'évolution du métier d'ingénieur-documentaliste vers la notion anglo-saxonne de « *data librarian* », développer de nouvelles compétences destinées à apporter de nouveaux services : ainsi s'affiche sa volonté de concevoir et développer une véritable infrastructure sociale de soutien aux chercheurs. D'ores et déjà impliqué dans de nombreux projets nationaux et internationaux, l'Institut fait montre de réactivité afin de relever les nouveaux enjeux de l'*Open Science*.

CATHERINE POUPON-CZYSZ

Responsable du Pôle Structuration & valorisation des connaissances en IST, Inist-CNRS
Catherine.Poupon-Czys@inist.fr



Le bâtiment de l'Inist à Vandœuvre-lès-Nancy.



POUR EN SAVOIR PLUS

CONTACT :

DonneesRecherche@inist.fr

Les données scientifiques, qu'elles soient issues de différents capteurs ou le résultat de simulations, voient leurs volumes et leur diversité croître de manière exponentielle (*big data*) et requièrent des infrastructures et des solutions permettant leur traitement (acquisition, analyse, visualisation, préservation) et leur partage. Les deux missions stratégiques nationales du Cines, calcul intensif et archivage pérenne, le positionnent naturellement pour fournir ce type de services.

Archivage intermédiaire de données scientifiques au Cines

Le Cines participe au projet Eudat (*EUropean DATA*)¹ qui a pour objectif la mise en place d'une infrastructure collaborative destinée à la conservation et au partage des données provenant des communautés européennes de chercheurs. Il est le seul centre de calcul français constituant un nœud de l'infrastructure Eudat. Le service de base offert en termes de pérennisation est le stockage sécurisé des fichiers ou « préservation du train de bits » (qui consiste à s'assurer que la succession de 0 et de 1 « incrustée » dans le média informatique n'est pas altérée au fil du temps) avec l'ajout d'un identifiant pérenne. Son intégrité et son accessibilité au travers d'un tel identifiant sont ainsi assurées contrairement à sa lisibilité (i.e. capacité à ouvrir un fichier « ancien » sur un système informatique « récent » et à en lire le contenu) et son intelligibilité au cours des années (i.e. capacité à comprendre le contenu d'un document archivé malgré le temps qui passe – sans la pierre de Rosette, les hiéroglyphes seraient lisibles mais très probablement inintelligibles), que seul un service d'archivage saura garantir.

ARCHIVER QUOI ?

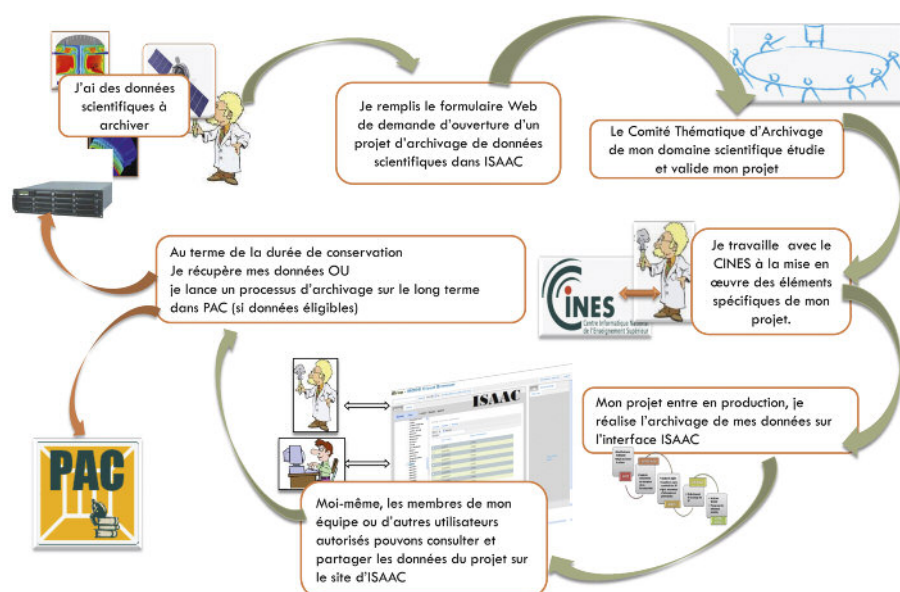
Lors des phases initiales du cycle de vie des données, les scientifiques qui les produisent ont des difficultés pour déterminer celles qui nécessiteront un archivage définitif. La mise en place d'un projet d'archivage pérenne sur la plateforme d'archivage au Cines (PAC) nécessitant un travail conséquent, l'incertitude sur les besoins constitue un frein à l'archivage des données scientifiques.

Afin de mieux comprendre leurs besoins, les équipes du Cines ont mené une enquête auprès d'environ 150 laboratoires de recherche français qui a fait apparaître les points suivants :

- il existe une volonté de conservation et d'accès aux données pendant au moins 3 à 5 ans ;
- le type de données correspond souvent à des résultats de calcul et d'observation, des codes sources ;
- habituellement les données sont conservées dans les laboratoires dans des formats extrêmement diversifiés ;
- l'archivage des fichiers explicatifs, et des métadonnées embarquées dans les fichiers, doit être pris en compte ;
- il n'existe pas de jeux ou de standard de métadonnées par thématique de recherche ;
- un partage des données dans un cercle restreint et la gestion des droits sur ce partage ont été décrits comme des nécessités ;
- les besoins en termes de volume de stockage sont significatifs : on parle de un à plusieurs dizaines de téraoctets par projet.

UNE OFFRE COMPLÉMENTAIRE D'ARCHIVAGE INTERMÉDIAIRE

Pour offrir une solution en adéquation avec les besoins, le Cines a mis en place un service « d'archives intermédiaires de données scientifiques » pour des communautés d'utilisateurs structurées. Ce service, nommé Isaac (Information scientifique archivée au Cines)², correspond à un stockage sécurisé des données, comportant obligatoirement un jeu minimal de métadonnées descriptives associées, afin d'en faciliter la recherche et la compré-



➔ Différentes étapes de la vie d'un projet d'archivage dans Isaac.

hension, pour une période déterminée (3 à 5 ans), et offre une possibilité d'accès partagé à ces données pour la communauté. Au terme de ce laps de temps, selon la décision de leur producteur, les données sélectionnées seront soit versées dans un système d'archivage définitif, soit restituées à l'utilisateur pour destruction ou simple stockage dans son laboratoire.

Bien évidemment, ce service ne peut être fourni que pour une communauté d'utilisateurs « structurée » partageant les mêmes formats de données (formats d'échange standards qui sont utilisés pour l'archive) et désireuse de mettre en œuvre une démarche combinant diffusion et conservation à long terme. Ce n'est pas une solution proposée pour du stockage prolongé des données, mais bien une approche pour faciliter la préparation d'une archive pérenne.

DES FACTEURS DE SUCCÈS

Lors de la mise en place de la plateforme Isaac et des différents services offerts au travers du nœud eudat@cines, le Cines a identifié plusieurs facteurs qui peuvent assurer le succès des projets d'archivage.

- **Le besoin d'informer sur l'archivage** : il est nécessaire de communiquer auprès des utilisateurs potentiels sur la plus-value de l'archivage électronique par rapport à un service de « sauvegarde sécurisée », couramment offert par la plupart des directions des systèmes d'information. Ceci doit permettre d'identifier les données nécessitant un archivage et de justifier le travail nécessaire à sa mise en œuvre.

- **Une relation de confiance** : la plupart des enquêtes effectuées auprès des scientifiques indique que confier ses données à un tiers nécessite la mise en place d'une relation de confiance. Pour ceci, le Cines met en place des conventions précisant clairement les notions de propriété et de responsabilité. Il continue sa démarche qualité s'appuyant sur les certifications adéquates, par exemple le *Data Seal of Approval* pour la plateforme Isaac.

- **L'intégration de l'archivage dans la chaîne de gestion des données** : les processus liés à l'archivage des données sont souvent considérés comme un surcroît de travail. Un des objectifs du Cines est de faciliter l'intégration de cette préoccupation et des actions induites dans les processus de gestion des données afin, par exemple, d'automatiser certains traitements ou de capturer certaines méta-données dès la création des données.

- **La migration vers l'archivage pérenne** : un des risques dont les utilisateurs doivent être conscients est le manque de caractérisation des données ou du format des fichiers lors de leur versement dans l'archive intermédiaire et, donc, une perte possible de la connaissance du contexte de production de l'information lorsqu'on se décide ultérieurement à

● ● ● QU'ES AQUO ?

Big data

Consiste à intégrer, synchroniser, traiter et valoriser de très grands volumes de données informatiques, extrêmement variées et de différentes natures. Ce terme s'est largement répandu en 2012. À l'opposé, les *small data* concernent des données qui peuvent être gérées directement par les individus.

Data mining

Traduit par « exploration de données », « fouille de données » ou encore « extraction de connaissances à partir de données » a pour objet l'extraction d'un savoir ou d'une connaissance à partir de grandes quantités de données par des méthodes automatiques ou semi-automatiques.

Data warehouse

Traduit par « entrepôt de données », désigne une base de données utilisée pour collecter, ordonner, et stocker des informations provenant de base de données opérationnelles et fournir ainsi une aide à la décision pour un établissement ou une entreprise. L'Amue, par exemple, a mis en place un entrepôt de données pour faciliter le développement d'outils de pilotage pour les établissements d'enseignement supérieur.

Retour sur quelques termes croisés dans le dossier de ce numéro d'Arabesques.

Digital Object Identifier (DOI)

Littéralement « identifiant d'objet numérique », mécanisme d'identification de ressources numériques, comme un film, un rapport, des articles scientifiques, etc. Le DOI d'un document permet notamment une identification pérenne de celui-ci et, par exemple, de retrouver l'emplacement d'un document en ligne si son URL a changé. Les DOI facilitent l'utilisation des bases de données bibliographiques, des logiciels de gestion bibliographique et permettent de produire des citations plus fiables et plus pérennes.

Repository

Traduit par « dépôt », « référentiel » (ou même « référentiel »), désigne un stockage centralisé et organisé de données. Ce peut être une ou plusieurs bases de données où les fichiers sont localisés en vue de leur distribution sur le réseau ou bien un endroit directement accessible aux utilisateurs.

Source : Définitions principalement extraites de Wikipédia.

les verser dans l'archive définitive. Cette recherche d'équilibre entre la simplicité de versement initial et la capacité à la pérennisation doit être évaluée avec les utilisateurs.

- **La volumétrie des données** : la volumétrie des données scientifiques à archiver pose plusieurs problèmes techniques qu'il faut résoudre pour garantir une qualité de service lors des dépôts ou des accès.

MARION MASSOL

Responsable du département Archivage et diffusion, Cines
massol@cines.fr

STÉPHANE COUTIN

Chef de projet, Cines
coutin@cines.fr

[1] www.eudat.eu

[2] www.cines.fr/spip.php?rubrique369

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 recherche¹. 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

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

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.

(*big data*). Au cours d'une présentation du dispositif Mastodons², 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

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

[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 Research 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 (Codata⁵), synthétise en une notion-clé, celle d'*intelligent openness*, l'exigence de données scientifiques accessibles, évaluables, intelligibles, et (ré-)utilisables⁶. 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 récom-



Photo : Jean-Marie Feurtet

➔ Cascade d'Aussois (73).

pense 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 Inspire⁸ favorisant l'interopérabilité des données cartographiques publiques (*via* le Géocatalogue, point d'accès national).

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/uploads/Files/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>



© Centre Pompidou-Mnam-Bibliothèque Kandinsky / Hervé Veronèse

La Bibliothèque Kandinsky : **à la croisée des arts**

Récemment déployée dans le Sudoc, la Bibliothèque Kandinsky confirme son rôle d'appui à la recherche en histoire de l'art. Collections et enjeux.

Installée au troisième étage du Centre Pompidou, la Bibliothèque Kandinsky, Centre de documentation et de recherche du Musée national d'art moderne-Centre de création industrielle (Mnam-Cci), est une bibliothèque patrimoniale et une bibliothèque de recherche spécialisée dans le domaine de l'art moderne et contemporain. Elle est l'héritière des fonds constitués au sein de la bibliothèque du Musée national d'art moderne depuis ses origines, de la documentation rassemblée à l'Hôtel Salomon de Rothschild par le Centre national d'art contemporain et de la documentation du Centre de création industrielle avant sa fusion avec le musée. Dotée de collections exceptionnelles, elle a vocation à servir un public composé de chercheurs, d'étudiants, de professionnels du monde de l'art, de l'architecture et du design et d'artistes. Structurée en deux secteurs, celui des imprimés et celui des archives et de la documentation, elle se conçoit comme un lieu de

conservation patrimoniale fréquemment sollicité pour enrichir les expositions de musées français et étrangers et comme une source essentielle pour la recherche et l'écriture de l'histoire de l'art des XX^e et XXI^e siècles.

UNE COLLECTION FONDAMENTALE EN ART, ARCHITECTURE ET DESIGN

La bibliothèque est riche de 200 000 imprimés dans les domaines des arts plastiques, de la photographie, du cinéma expérimental, de la vidéo, des nouveaux médias, de l'architecture et du design. Les axes de développement des collections s'appuient sur la programmation des expositions du musée, la documentation des artistes, architectes et designers de la collection et intègrent les scènes artistiques émergentes dans un contexte de mondialisation. L'accent est mis sur l'histoire des grands mouvements artistiques et de la pensée esthétique, la théorie de l'art et la muséographie de l'art contemporain. Sur le plan patrimonial, la tonalité particu-

lière des collections d'imprimés de la Bibliothèque Kandinsky procède des bibliothèques privées d'artistes, d'architectes, de critiques, de conservateurs, de marchands d'art qui lui ont été données ou léguées (Kandinsky, Brancusi, Delaunay, Surville, Larionov, Kahnweiler, Malraux, Sottsass, pour n'en citer que quelques-uns). Elle conserve également une collection exceptionnelle de livres d'artiste et livres de photographie qui témoigne des grandes tendances artistiques internationales et du travail de jeunes artistes dont les créations constituent le patrimoine de demain.

Le fonds de revues de 7 000 titres s'ouvre à des domaines autres, tels que la littérature, la poésie, le théâtre, le cinéma, la photographie, la philosophie et la politique. Il comprend des bulletins et revues publiés par les musées et galeries et des revues d'artiste. Ce fonds offre des ensembles majeurs pour les avant-gardes artistiques : constructivisme, futurisme, dadaïsme, surréalisme en Europe

et en Amérique du Sud. Les revues des grands mouvements d'après-guerre (*CoBra*, *Lettrisme*, *Internationale situationniste*) sont également présentes. En 2006, la bibliothèque a fait l'acquisition du fonds Destribats, collection de revues des avant-gardes internationales, grâce au mécénat du groupe Lagardère. Ces revues font actuellement l'objet d'une mise en valeur importante dans les salles consacrées au nouvel accrochage « Modernités plurielles 1905-1970 ».

PLUS D'UNE CENTAINE DE FONDS D'ARCHIVES

Depuis sa création, la Bibliothèque Kandinsky développe une politique de constitution et de conservation de fonds d'archives privées consacrés à des artistes dont les œuvres figurent dans les collections du Mnam, des photographes, des architectes ou des acteurs du monde de l'art (galeristes, critiques et historiens). La bibliothèque est riche aujourd'hui de plus d'une centaine de fonds acquis par achats ou grâce à la générosité de donateurs ou encore par des legs ou des donations. Les fonds d'archives comprennent généralement, en partie ou en totalité, des manuscrits (brouillons ou textes inédits, correspondances reçues et doubles de correspondances envoyées), des *ephemeras* (cartons d'invitation, archives de presse d'un artiste), des photographies et des documents audiovisuels, ainsi que des livres et des périodiques constituant la bibliothèque d'un artiste, d'un critique. Ces acquisitions prestigieuses, volumineuses ou modestes, ces documents rares, quelquefois inattendus et trop souvent ignorés, constituent un outil de recherche précieux sur l'histoire de l'art moderne et contemporain.

À ces fonds d'archives viennent s'ajouter quelques 12 000 dossiers individuels d'artistes et 3 000 dossiers collectifs consacrés à différents mouvements ou techniques artistiques, au milieu de l'art et à des personnalités remarquables...

DES PROGRAMMES DE NUMÉRISATION VARIÉS

Depuis 2002, des campagnes de numérisation ont été menées et, plus récemment, grâce aux crédits du grand emprunt. Des corpus divers ont pu être valorisés publiquement même si, en raison du caractère contemporain des collections, une part importante de ces documents est en accès restreint.

À ce jour, sont disponibles des reportages documentaires consacrés à des expositions

temporaires et à des lieux de création d'artistes contemporains de la collection permanente du musée, des reportages réalisés par Véra Cardot, Pierre Joly ou encore Jean Prouvé, ce dernier constituant l'ensemble documentaire le plus important consacré à l'architecture en France des années 1940 à 80.

Numérisé également, le fonds Brauner comprend notamment une sélection de cahiers-manuscrits et de carnets illustrés, les cours du CNAM de Jean Prouvé constitués de manuscrits et dessins...

En 2013, des revues d'Europe de l'Est et une sélection de revues françaises ayant fait l'objet d'une convention de partenariat avec la Bibliothèque nationale de France ont été numérisées intégralement.

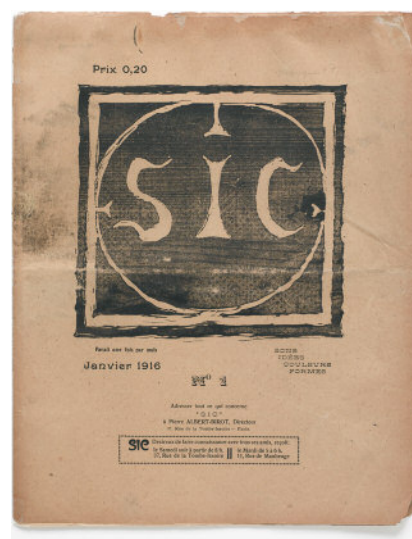
La partie audiovisuelle n'est pas en reste : collections de films, de vidéos, d'enregistrements sonores sur les artistes et le monde de l'art, entretiens avec des collectionneurs, critiques ou commissaires d'exposition, œuvres sonores ou musicales, conférences ou tables rondes autour des expositions temporaires.

UNE IMPLICATION FORTE DANS LES PROGRAMMES DE RECHERCHE

La bibliothèque Kandinsky participe à plusieurs programmes de recherche et accueille régulièrement, dans ses locaux, des étudiants bénéficiant des bourses doctorales du Centre Pompidou et des étudiants des universités et écoles d'art. C'est dans ce contexte qu'un cercle de réflexion sur l'histoire des expositions du Centre Pompidou a été mis en place en 2010. Son objet est d'interroger le format et la nature des expositions, la pratique de commissariat et la place de l'histoire des expositions dans l'histoire de l'art. Dans cette dynamique a été lancée l'élaboration d'un catalogue raisonné en ligne des expositions du centre Pompidou, dont Stéphanie Rivoire, responsable du secteur des archives et de la documentation de la bibliothèque, est chef de projet. Il servira autant d'outil pour les chercheurs que de modèle prescriptif pour une normalisation du processus d'archivage des expositions.

... ET DANS LES PROJETS D'EXPOSITION

En 2012, la Bibliothèque Kandinsky a participé à la quatrième édition du Nouveau festival du Centre Pompidou, rendez-vous pluridisciplinaire qui a fait la part belle à Guy de Cointet, artiste français aux frontières des arts plastiques, du théâtre et de la perfor-



➔ La revue *SIC*, issue du fonds Destribats, est une revue d'avant-garde française, publiée de janvier 1916 à décembre 1919, sous la direction de Pierre Albert-Birot. Elle a notamment compté pour collaborateurs Guillaume Apollinaire, Max Jacob, Louis Aragon, Raymond Radiguet...

mance dont elle conserve les archives (dessins, carnets de notes, travaux préparatoires). Durant l'été 2014, la bibliothèque proposera une exposition documentaire dans une des galeries du musée afin de célébrer les 25 ans de l'exposition « Magiciens de la terre », qui s'est tenue simultanément au Centre Pompidou et à la Grande halle de La Villette en 1989 et dont la caractéristique fut de placer sur la scène internationale les arts non occidentaux.

L'INSERTION AU SEIN DU RÉSEAU SUDOC

Début 2013, la Bibliothèque Kandinsky a rejoint le réseau Sudoc. Il s'agit là de répondre à la fois à la nécessité d'être davantage visible dans le milieu universitaire, mais également de bénéficier du catalogage partagé et de tisser de nouveaux liens avec le réseau des bibliothèques spécialisées en histoire de l'art. Notre volonté est de nous ouvrir encore plus largement à d'autres plateformes collaboratives et d'explorer de nouvelles possibilités de partenariats avec des bibliothèques et des communautés de chercheurs en histoire de l'art. Nous y travaillons au quotidien.

SYLVIA BOZAN

Responsable du secteur des imprimés
Sylvia.Bozan@centrepompidou.fr



POUR EN SAVOIR PLUS

SITE DE LA BIBLIOTHÈQUE KANDINSKY
<http://bibliothequekandinsky.centrepompidou.fr>

Mutualiser et décroisonner !

Rencontre avec Jérôme Kalfon, directeur de l'Abes

Jérôme Kalfon a été nommé directeur de l'Abes, le 1^{er} octobre 2013, après avoir dirigé le SCD de l'université Paris Descartes. Responsable du département Études et prospectives de Couperin de 2006 à 2012, il a animé de nombreux groupes de travail et de réflexion (Archives ouvertes, Ebooks, Archivage et accès pérenne...). Sa forte implication au sein des associations et groupements professionnels (ADBU, GFII...) lui a permis d'échanger avec de nombreux acteurs de l'information scientifique et technique (chercheurs, enseignants, fournisseurs...). Membre du conseil scientifique de l'Abes depuis sa création en 2010, il a suivi de près les derniers développements de l'agence.

Arabesques lui a posé quelques questions sur sa perception de l'Agence après ses premières semaines d'immersion.

Vous arrivez à l'Abes à un moment où celle-ci se trouve confrontée à des dossiers majeurs comme le projet de système de gestion de bibliothèque mutualisé (SGBM) ou l'acquisition de documentation électronique sous forme de licences nationales (Istex). Comment abordez-vous ces questions ?

Elles illustrent les mutations des rôles de l'Abes. Le Sudoc, les applications liées au dépôt des thèses au format numérique, le financement d'opérations de rétro-conversion démontrent que l'Abes est montée en charge dans son rôle d'opérateur de l'État. Grâce à mes prédécesseurs et à des équipes compétentes et motivées, l'Abes a conquis la confiance des établissements. De ce fait, l'Agence a évolué, endossant aussi un rôle actif dans les actions de mutualisation. Un projet comme le SGBM en est l'illustration. Tous ces dossiers réclament une évolution dans le mode de gouvernance. Les décisions seront une émanation de la volonté des établissements. Mon expérience au sein de Couperin m'a convaincu

de la force d'un réseau, surtout si les membres de ce réseau ont la profonde conviction qu'ils participent aux décisions. Beaucoup d'établissements ont répondu positivement à l'appel à participation au projet SGBM, dès sa phase pilote. De nombreux autres nous ont encouragés et se sont dits prêts à rejoindre cette initiative dès la première phase de déploiement. C'est très motivant. Il n'en reste pas moins que, dans un contexte en rapide mutation et avec des solutions qui ne sont pas encore arrivées à maturité, il s'agit d'un lourd défi. Un défi qu'il est indispensable de relever. L'alliance de l'Abes et d'un certain nombre d'établissements me semble être la bonne échelle pour un tel projet, notamment dans sa phase pilote. Quoi qu'il en soit, proposer ne veut pas dire imposer, les établissements doivent choisir librement leur système.

Pour en venir à Istex, il convient de rappeler qu'il ne s'agit pas simplement d'un projet d'acquisitions – négociées par Couperin – sous forme de licences nationales. Certes, c'est le principal rôle qui est confié à l'Abes dans ce dispositif, mais il s'agit surtout de participer à une offre de services totalement nouveaux. En ce sens,

il s'agit d'un projet de recherche-développement associant des professionnels de l'information scientifique et technique (IST), notamment à l'Inist et à l'Abes, des laboratoires de recherche et, bien entendu, des utilisateurs. Grâce à des outils de fouille dans le texte intégral, il s'agira de faire remonter des informations pertinentes pour un chercheur ou un domaine. Istex vient illustrer le fait qu'aucune agence, aucun réseau ne peut agir seul. Il est, par excellence, l'illustration d'un décroisonnement, de l'engagement d'acteurs différents dans un projet commun. Chercheurs, bibliothécaires, ingénieurs apprennent à mieux se connaître et à conduire des actions coordonnées. Quand organismes, opérateurs, universités, grandes écoles, agences rejoignent un cadre structurant tel que la Bibliothèque scientifique numérique (BSN) pour mieux travailler ensemble, on peut espérer des résultats... d'excellence.

Quel devenir pour les missions traditionnelles de l'Abes ?

Sur le fond, les missions traditionnelles de l'Abes demeurent. Je dirais même qu'elles sont le socle de confiance des établissements. Sur la forme, elles se doivent d'évoluer. Les missions traditionnelles sont celles-là mêmes qui permettent aux nouvelles missions de trouver à l'Abes un terreau propice à leur développement. Les nouvelles missions, ou plus exactement les nouvelles applications, les nouveaux développements, permettent aux services traditionnels d'être mieux assurés. De nouveaux services, comme Périscope, sont justement le résultat de l'addition des services dits traditionnels et de nouvelles applications qui peuvent être développées grâce à une base miroir. Dans cette même logique, nous allons rechercher les améliorations les plus utiles aux établissements sans exiger pour autant de lourds développements. Car l'engagement dans



d'ambitieux projets ne doit pas pour autant nous faire oublier les besoins d'amélioration des services au quotidien. Bien au contraire, si l'Abes s'y montre attentive, on peut espérer, en retour, une meilleure implication dans les projets à venir.

Quel rôle spécifique pourrait jouer l'Abes pour renforcer la coopération entre les différents acteurs de l'IST ?

Parler de spécificités, c'est souvent chercher à se retrancher, à se distinguer, à expliquer que l'on est différent. S'il est une spécificité de l'Abes, c'est d'être un outil pour le travail en commun. Si vous me permettez de jouer sur les mots, je

dirais que sa spécificité, c'est de n'être pas spécifique. L'Abes dispose d'un statut (établissement public administratif) qui permet aux établissements de conduire des actions communes.

Tout comme l'Amue, le Cines, l'Inist, les organismes de recherche, l'Abes a des caractéristiques qui découlent de ses missions. La spécificité, c'est l'île ; à l'heure où les médias et les métiers convergent, il convient surtout de jeter des ponts. Enfin, pour ne pas laisser totalement votre question sans réponse, je dirais que l'une des caractéristiques des métiers de l'Abes est de structurer des référentiels, des normes, des autorités, des langages communs. Il s'agit là d'autant de passerelles.

L'Abes souffre parfois encore d'un manque de visibilité auprès des instances universitaires et de la recherche. Par quels moyens accroître celle-ci ?

Oui, peut-être justement parce que nous avons trop souvent songé à nos spécificités. Par quels moyens ? La BSN me semble être un cadre structurant pour la coopération. Mais la coopération peut s'étendre au-delà, je songe par exemple à la Bibliothèque nationale de France, mais aussi et surtout à l'Europe.

Pensez-vous qu'en dehors du Sudoc PS qui intègre des bibliothèques du réseau de lecture publique pour la signalisation des périodiques, d'autres coopérations pourraient être nouées avec celles-ci ?

Bien entendu, l'heure est au décloisonnement. C'est déjà le cas dans d'autres domaines, je songe à l'accès aux acquisitions effectuées dans le cadre des licences nationales, accessibles via le réseau des bibliothèques publiques. Enfin, la perspective des nécessaires adaptations au nouveau schéma EAD serait l'occasion d'envisager une mutualisation des moyens en vue de développer un outil national de production EAD.

Comment dépasser la perception essentiellement technique des missions de l'Abes pour les inscrire plus largement dans la construction d'une société de la connaissance ?

L'Abes est un opérateur essentiellement technique. Il ne faut pas le nier, mais au contraire l'affirmer. Mais derrière des choix en apparence techniques se cachent des enjeux stratégiques. Ce n'est pas aux techniciens, pour aussi compétents qu'ils soient, de faire des choix stratégiques. Il faut être en mesure d'expliquer aux décideurs et à l'ensemble des acteurs concernés les enjeux de ces choix. On peut aussi imaginer qu'en faisant participer l'ensemble des acteurs à des consultations sur des choix techniques, ils seront conduits à s'interroger sur leurs enjeux. Je pense qu'il devient nécessaire d'emprunter ces nouvelles voies.

*Propos recueillis par
le comité de rédaction*



**LA BDIC ET
LA GRANDE GUERRE**

A lors que fin novembre 35 000 personnes étaient déjà inscrites à la première vague de Moocs (*Massive Online Open Course*), lancée à travers la plateforme France Université Numérique, la Bibliothèque de documentation internationale contemporaine (BDIC) figure parmi les pré-curseurs avec un programme construit à partir des sources sur la Grande Guerre conservées dans ses collections.

Conçu comme un laboratoire, « *La Première Guerre mondiale expliquée à travers ses archives* » (13 janvier-28 avril 2014) s'adresse à un large public dans une perspective de formation continue.

INSCRIPTION GRATUITE EN LIGNE :

www.france-universite-numerique-mooc.fr/courses/Paris10/10001/Trimestre_1_2014/about

Les chantiers de l'Abes

« SGBM », « Bacon », « Hub », les couloirs de l'Abes – et sans doute ceux de nombreuses bibliothèques des réseaux – résonnent ces temps-ci de mots mystérieux, curseurs d'une actualité de rentrée chargée en projets. Mais qu'ont-ils donc en commun ces mots qui nécessitent l'usage d'un glossaire ou d'un manuel technique ?

Points de convergence : la documentation électronique

Les équipes de l'Abes, en collaboration avec le consortium Couperin et de nombreux experts des réseaux documentaires, sont largement mobilisées autour de trois projets, ambitieux, convergents et interdépendants, à savoir : la mise en œuvre d'un **système de gestion de bibliothèque mutualisé** (SGBM), la conception d'un **hub de métadonnées** et celle d'une **base de connaissance nationale** (Bacon). Leur objectif commun est inscrit au cœur des engagements du projet d'établissement 2012-2015 : apporter aux bibliothèques une réponse globale, concrète et complète à la problématique lancinante du signalement, de l'accès et de la gestion de leur documentation numérique. Facettes d'une même stratégie qui s'appuie sur les recommandations de l'étude **Découvertes** (*Discovery*), commanditée par l'Abes à Maurits van der Graaf du cabinet Pleiade, leur aboutissement profile, à moyen terme, la refonte d'un système universitaire de documentation, en voie d'obsolescence à l'ère de la transition numérique.

Tentons de comprendre en quoi ces trois projets sont interdépendants. Si le projet de SGBM répond au besoin exprimé par les bibliothèques du réseau de renouveler – voire de repenser – leur système d'information, plus précisément leur système de gestion intégré de bibliothèque, devenu aujourd'hui bien malhabile face à l'explosion des ressources numériques, les projets « **Hub de métadonnées** » et « **Bacon** » se concentrent pour leur part sur l'automatisation des *process* autour des métadonnées. Avec le SGBM, il s'agit de fluidifier les circuits, de moderniser les interfaces de travail, d'alléger les équipements informatiques grâce à un système hébergé dans le *cloud*. En revanche, la maîtrise collective des (méta)données est considérée comme une nécessité

absolue si l'on veut en garantir la qualité, l'indépendance et la libre réutilisation, notamment au sein du web de données. En associant ces trois projets, l'Abes est porteuse d'une ambition nationale pour les métadonnées de l'information scientifique et technique (IST), pierres angulaires de la dynamique internationale qui vise à construire la société (du partage) des connaissances.

Points d'actus en points de suspension...

Le projet SGBM est entré dans sa seconde phase avec comme échéance, au printemps 2014, la rédaction du cahier des charges en vue de l'appel d'offres pour un marché à dialogue compétitif. Pour ce faire, l'Abes a lancé durant l'été un appel à candidature auquel 44 établissements ont répondu (qu'ils en soient ici remerciés), 15 d'entre eux s'étant portés candidats et 12 ayant été retenus afin de constituer, en concertation avec l'ADBU, deux groupes de travail, dont les réunions de lancement ont eu lieu en octobre. D'un côté, les travaux du groupe de travail « **Catalogage et production de métadonnées** », piloté par Maurits van der Graaf, aboutiront en mars à la rédaction d'une étude sur « les objectifs d'évolution de service de catalogage partagé et de la production des métadonnées communes dans le réseau Sudoc et Sudoc-PS ». De l'autre, le groupe de travail « **Sites pilotes** », coordonné par Jean Bernon et composé pour une large part des chefs de projets des établissements pilotes engagés dans la mise en place d'un SGBM, a pour mission de recueillir et mettre en forme l'ensemble des informations indispensables à la migration entre les systèmes.

De son côté, l'équipe projet « **Bacon** », composée de deux chefs de projets fonctionnels (Benjamin Bober de l'Abes et Thomas Porquet



POUR EN SAVOIR PLUS

POUR SUIVRE
L'ACTUALITÉ DE
CES PROJETS

• **Système de gestion de bibliothèque mutualisé (SGBM) :**

<http://sgbm.Abes.fr>

• **Base de connaissance nationale (Bacon) :**

<http://bacon.Abes.fr>

• **Qualinca :**

www.lirmm.fr/qualinca

de Couperin), d'un chef de projet technique (Olivier Martinez), d'un développeur (Julien Couzinié) et de plusieurs collègues des SCD, a pour mission de concevoir le prototype d'une base de connaissance nationale. Lors de la réunion de lancement du 24 septembre, il a été décidé que ce prototype se focaliserait sur la récupération automatique ou semi-automatique de données en provenance des éditeurs, sur leur mise en conformité avec la recommandation KBART (v.1), afin de les exposer de façon adéquate pour répondre aux différents besoins et systèmes des établissements. Pour ce faire, l'équipe informatique s'appuie sur les solutions logicielles *open source* de KB+ (projet britannique porté par le JISC) et de CUFTS (projet porté par la Simon Fraser University). Côté fonctionnel, les problématiques des exports, des mises à jour et des résolveurs de liens seront au cœur de l'étude. Résultats attendus pour la fin janvier 2014.

Pour sa part, depuis septembre, l'équipe projet « **Hub de métadonnées** », pilotée par Yann Nicolas et Christophe Bonnefond, s'est fixé pour objectif transversal de doter l'Abes de nouveaux outils et méthodes d'analyse et traitement des métadonnées. La première phase du projet consiste à traiter les métadonnées des documents acquis sous licence nationale dans le cadre du programme Istex. Au programme des opérations qui viendront conforter et transformer le prototype réalisé en début d'année en outil de production : vérifier, préciser et corriger les métadonnées livrées par les éditeurs afin de permettre l'exemplarisation automatique des documents concernés dans le Sudoc et, parallèlement, enrichir ces métadonnées en les liant aux différents référentiels (IdRef, Vial, Rameau, Dewey, etc.). Il s'agit également de multiplier les canaux de sortie afin d'encourager à terme la récupération des métadonnées par les professionnels et de garantir leur exposition sur le web de données. En cela, le hub de métadonnées se situe d'emblée comme l'une des sources d'alimentation de la future base de connaissance nationale.

CHRISTINE FLEURY

Chargée de la cellule Communication, ABES
fleury@abes.fr



Et Qualinca ?

Depuis 2012, l'Abes est engagée aux côtés du Laboratoire d'informatique, robotique et microélectronique de Montpellier (LIRMM) dans un projet de recherche financé par l'Agence nationale de la recherche « Qualinca » (Qualité et interopérabilité de grands catalogues documentaires). Selon le parti pris du projet, la qualité des catalogues tient essentiellement à celle des liens : liens externes tels que les préconisent les principes du *linked data*, mais aussi liens internes, notamment entre notices bibliographiques et notices d'autorité. S'appuyant sur les technologies du web sémantique, il vise à faire progresser l'approche théorique des laboratoires partenaires tout en garantissant la dimension opérationnelle de celle-ci. Ainsi, l'Abes ne perd pas de vue ses propres objectifs : se doter d'outils et de modes opératoires adaptés à la transition en cours. Pour constituer

l'équipe, pilotée par Yann Nicolas et Christophe Bonnefond, l'Abes a recruté en juin dernier sur financements du projet de recherche, Aline Le Provost, analyste de données spécialiste de la modélisation. Renforcée côté informatique par Julien Gibert, développeur fin connaisseur des données bibliographiques, l'équipe projet s'appuie sur les enseignements apportés par le développement du prototype SudocAD¹, par les chantiers de modélisation et de conversion des données du Sudoc en RDF et selon le modèle FRBRoo², modèle conceptuel au cœur du projet. Les travaux en cours consistent à définir méthodes et outils capables d'automatiser la correction de liens erronés et le liage entre notices³. Des travaux de haute technicité qui viennent enrichir les perspectives du hub de métadonnées... et celles du Sudoc.

CHRISTINE FLEURY

[1] SudocAD : A Knowledge-Based System for the Author Linkage Problem (KSE 2013), www.lirmm.fr/qualinca/sites/default/files/sudocAd.pdf

[2] FRBRoo : *Functional Requirements for Bibliographic Record - Object Oriented*.

[3] Preuve de reconnaissance, l'Abes était invitée à présenter un point d'étape du projet Qualinca lors de la 22^e réunion du groupe d'harmonisation FRBR-CIDOC CRM (Héraklion, 21-25 octobre 2013).

(Agenda)

Janvier

LE 23

JOURNÉE FULBI

Paris, Médiathèque Marguerite-Duras

« Numérique et handicap : nos données sont-elles accessibles ? », tel est le thème retenu pour la journée d'étude annuelle de la Fédération des utilisateurs de logiciels pour bibliothèques, documentation et information.

Entrée gratuite sur inscription préalable obligatoire. www.fulbi.fr

Février

DU 27 AU 28

FSR 2014

Rome (Italie)

The Italian Library Association (AIB) et la Bibliothèque Vaticane programment 2 jours de conférences autour du thème « *Faster, Smarter, and Richer : Reshaping the Library Catalogue* ». Seront abordés la coopération entre bibliothèques dans le traitement et la gestion de contenus, les partenariats avec les institutions culturelles, la gestion spécifique des fonds patrimoniaux, les questions liées au stockage de l'information...

www.aib.it/attivita/congressi/2013/38702-fsr-2014-outline

Mars

DU 20 AU 21

LA PÉRENNISATION DE L'INFORMATION

Toulouse, Département Archives et médiathèque (DAM)

Les journées d'étude annuelles du DAM de l'université de Toulouse 2-Le Mirail proposent aux professionnels de la documentation et aux chercheurs de s'interroger sur la pérennisation de l'information actuellement produite et diffusée dans un contexte numérique.

<http://dam.univ-tlse2.fr/accueil-dam/la-recherche-et-les-br-ressources-documentaires/les-journees-d-etudes-du-dam/>

DU 21 AU 24

SALON DU LIVRE DE PARIS

Paris, Porte de Versailles

Alors que Shangai sera la ville invitée de cette 34^e édition du Salon du livre de Paris, les lettres argentines seront également à l'honneur : une occasion de célébrer le centenaire de la naissance de Julio Cortázar, figure emblématique de la littérature argentine de la seconde moitié du XX^e siècle qui a longtemps vécu à Paris. www.salondulivreparis.com

DU 24 AU 27

CODE4LIB

Raleigh (USA)

Le congrès Code4lib rassemble, depuis 2006, les acteurs de l'informatique documentaire à l'échelon international. Il facilite l'échange sur les différentes pratiques et la mise en place de collaborations ou de partenariats.

<http://code4lib.org>

DU 26 AU 27

DOCUMATION 2014

Paris, CNIT Paris La Défense

Pour célébrer sa vingtième édition, ce salon consacré à la gestion de l'information et du document numérique accueillera plus de 160 exposants du secteur de la gestion de contenus, documents et données et proposera des rencontres interprofessionnelles. www.documation.fr

En route pour...

IFLA
2014
LYON

Congrès Mondial des Bibliothèques
et de l'Information
80^e Conférence et Assemblée
générale de l'IFLA
16-22 août 2014 Lyon

DU 16 AU 22 AOÛT

Palais des Congrès de Lyon

Le 80^e congrès de l'Ifla se prépare activement autour du thème « *Bibliothèques, citoyenneté, société : une confluence vers la connaissance* ».

De nombreux volontaires, qui seront chargés de l'accueil des congressistes, ont déjà répondu présents à l'appel lancé par le Comité national pour l'organisation du congrès.

Diverses autres formes de participation vous sont proposées, selon vos disponibilités et votre budget, notamment :

- PARTICIPER AU CONGRÈS (pour sa totalité ou à la journée) : les inscriptions sont ouvertes depuis le 1^{er} octobre 2013 (tarif réduit jusqu'au 15 mai 2014) ;
- RÉPONDRE AUX APPELS À COMMUNICATION lancés par les différentes sections de l'Ifla (pour les connaître, s'inscrire à la liste de diffusion Ifla-L) ;
- PROPOSER UN POSTER pour présenter un projet ou une action professionnelle (jusqu'au 3 février 2014) ;
- TRADUIRE par écrit les différentes interventions (contact : CFIBD) ;
- RÉSERVER un stand avant fin janvier 2014 pour bénéficier d'un tarif privilégié.

Les organisateurs comptent sur la présence en nombre des professionnels français de l'enseignement supérieur et de la recherche.

→ Retrouvez toutes les informations utiles concernant la préparation du congrès sur le site de l'Ifla (www.ifla.org) et du CFIBD (www.cfibd.fr)



À NOTER DÈS À PRÉSENT

Les prochaines Journées ABES se tiendront au Corum de Montpellier le mardi 20 et le mercredi 21 mai. Le programme complet et le formulaire d'inscription seront accessibles à partir de la mi-mars sur le site web de l'agence.