Arlabesques

OCTOBRE - NOVEMBRE - DÉCEMBRE 2014

76

DOSSIER

Enrichir pour partager

Les bibliothèques dans l'économie de la connaissance

PLEINS FEUX SUR • La bibliothèque de Paris 8, labellisée Marianne

ACTUALITÉS • Sudoc II / Journées CTLES / Déclaration IABD...

abes
agence bibliographique de l'enseignement supérieur

(Dossier) ENRICHIR POUR PARTAGER

Les bibliothèques dans l'économie de la connaissance

Les bibliothèques et les institutions nationales ou internationales qui œuvrent à leurs côtés sont des acteurs incontournables pour la construction d'une économie de la connaissance ouverte et partagée. Ce dossier pointe la plus-value apportée à différents niveaux pour enrichir les données (hub de métadonnées de l'Abes), pour simplifier les échanges d'informations entre leurs producteurs (identifiant ISNI), pour faciliter l'accès aux métadonnées des publications en série (ISSN)... Ce mouvement n'en est qu'à ces débuts et le traitement automatique de masse dans le cadre du *text* et *data mining* devra lever encore quelques obstacles juridiques et techniques pour répondre aux besoins des étudiants et des chercheurs. En amont, les bibliothèques universitaires auront sans doute aussi à accompagner les enseignants-chercheurs dans leur activité de publication en *open access* en termes de procédures et de mise à disposition d'outils (cf. l'expérience conduite à Clermont Université). Permettre à chacun de trouver son miel dans la ruche gigantesque d'informations et de données, tel est défi majeur à relever pour ces prochaines années.

22 (Pleins feux sur...)

La bibliothèque de Paris 8, labellisée Marianne

© Thierry Kiefer/SCD Paris 8

24 (Actualités)

28 (Agenda)

04 **Bibliothèques et économie(s) de la connaissance :** *quelles valeurs ?*
Lionel Maurel

06 Modèles économiques et économie de la connaissance : approches du GFII
Olivier Delteil / Ruth Martinez

07 *Acquisitions Istex :* **traiter et enrichir les données** Marion Grand-Démery

09 **L'ISNI** *prend son essor* Pauline Chougnet

11 Articulation ISNI-IdRef : un enjeu pour l'identification pérenne François Mistral

12 L'ISSN en mutation Gaëlle Béquet

13 **La fouille de textes** *au service de la documentation* Mathieu Roche / Sophie Fortuno

15 **Metadata mining :** *fouiller les données des catalogues ?* Gildas Illien

17 Des données ouvertes sur l'enseignement supérieur et la recherche
Emmanuel Weisenburger

18 **Des publiothèques** *universitaires ?*
Olivier Legendre

20 **La veille informationnelle :** *enjeux et évolutions* Frédéric Martinet

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
Coordination éditoriale et secrétariat de rédaction :
Béatrice Pedot (beatricepedot@wanadoo.fr)
Comité de rédaction : Dominique Esmenjaud, Christine Fleury, Marion Grand-Démery, Philippe Le Pape, Béatrice Pedot, Maryse Picard, Marie-Pierre Roux.
Conception graphique : Anne Ladevie (anne-ladevie@orange.fr)
Impression : Pure Impression

Couverture : *Wall Street bubbles – Always the same.*
Caricature de John Pierpont Morgan par Udo J. Keppler, publiée dans *Puck*, n° 1264 (22 mai 1901). Library of Congress Prints and Photographs Division Washington (LC-USZC4-7880) www.loc.gov/pictures/resource/cph.3g07880

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

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

Valeur, valeurs
Modèle, modèles

Nous devrions nous réjouir, la bibliothèque est partout, dans toutes les poches, jamais l'information n'a été aussi accessible à tous qu'aujourd'hui. Paradoxe, jamais nous ne nous sommes autant plaints du coût, sans cesse croissant, de l'accès à l'information scientifique et technique. Alors même que la masse de contenus disponible explose, que quantitativement l'offre des bibliothèques s'accroît considérablement, jamais nous n'avons assisté à autant d'indignation face à des restrictions d'accès à l'information. Car dans un monde où l'abondance est naturelle et la rareté artificielle, nous n'avons pas trouvé le moyen de faire sans cet artifice, ni même nous assurer qu'il est toujours et partout souhaitable.

Dans cet univers mouvant et incertain, nous rencontrons des droits qui n'obéissent pas à certaines lois, des lois qui n'obéissent pas à certains droits légitimes.

Parce que la technique bouleverse les pratiques, on interroge l'économie de la connaissance. La technique a ses lois, l'économie les siennes, il serait vain de les ignorer, de vouloir s'en abstraire. Seule la compréhension de la technique et de l'économie permet d'agir sur les terrains où elles s'appliquent. Agir aussi en fonction d'autres lois, dont celles qui président à la construction des savoirs. Dans cet univers mouvant et incertain, nous rencontrons des droits qui n'obéissent pas à certaines lois, des lois qui n'obéissent pas à certains droits légitimes, et un droit à construire : le droit de la science ouverte.

Le droit – donc les lois – est fondé sur des valeurs, des fondamentaux. Parmi ces fondamentaux, celui que les idées et les connaissances n'appartiennent à personne, qu'elles sont un bien commun de la société dans son ensemble qui ne doit connaître de frontières ni géographiques ni économiques ni techniques. Il en va ainsi notamment du droit de lire, par tous les moyens, y compris assisté par des machines, pour construire de nouvelles connaissances à partir des contenus préexistants, en tirer le meilleur parti. Ces valeurs débouchent sur une ligne de conduite : éviter la captation, qu'elle soit privée, car une économie basée sur la rente conduit à l'immobilisme, ou publique, car la vocation d'un service public dans la société de la connaissance est d'offrir des perspectives aux nouveaux entrants et créer les conditions pour leur émergence, à l'opposé donc de l'enclosure.

Les acteurs devront agir sur des tensions antagonistes. Ainsi les piliers du métier du bibliothécaire sont ébranlés. Acquérir pour une communauté ? Si l'on va jusqu'au bout de la logique du libre accès, l'acquisition, au sens de transaction économique, disparaît et les crédits y afférents n'ont plus lieu d'être. Cataloguer ? Tendanciellement, le catalogage se fera à la source, dans divers formats et divers modèles. Le traitement de masses considérables d'informations – dont les publications ne seront qu'une infime mais indispensable partie – sera confié à des acteurs agissant à des échelles sans commune mesure avec la bibliothèque traditionnelle.

L'identité du métier évolue, le nombre de fonctions et les attentes vis-à-vis des spécialistes de l'information s'accroissent plus vite que ne disparaissent certaines tâches. Lier publications et données, les désambiguïser, les organiser, les retraiter, accompagner le publiant est un nouveau rôle des bibliothécaires et s'inscrit dans une longue tradition, celle de classer pour penser. Il y a tout lieu d'être optimiste, à condition de savoir s'adapter.

Dans cet univers incertain, nous pouvons avoir une certitude : la transition sera longue. Divers modèles vont coexister durablement : acquisitions à titre payant ou à titre gratuit ; utilisation de contenus du domaine public ou, inversement, mise dans le « domaine public » de contenus en principe sous droits ; inscription des acteurs à différents niveaux de la chaîne de valeur. Nous n'allons pas vivre avec un seul modèle mais avec plusieurs et nous devrons assumer les tensions qui en résultent.

Ce numéro d'*Arabesques* ne prétend pas clore le sujet, mais donner quelques repères sur les enjeux et évolutions en cours, étayés par des cas concrets.

Jérôme Kalfon
Directeur de l'Abes

Bibliothèques et économie(s) de la connaissance : *quelles valeurs ?*

Si, de toute évidence, les bibliothèques sont un des acteurs de l'économie de la connaissance, reste à définir de laquelle il s'agit et pourquoi. Point de vue engagé de Lionel Maurel, cofondateur du collectif SavoirsCom1 et membre du Conseil d'orientation stratégique de l'association La Quadrature du Net.

C'est devenu un lieu commun de dire que nous avons basculé dans une économie de la connaissance, mais cette mutation s'est fortement accélérée avec le développement des technologies numériques. Signe des temps, la capitalisation boursière de Google s'était hissée en février dernier au deuxième rang mondial[1], en passant devant celle du groupe pétrolier Exxon (Apple occupant le premier rang). Un vrai symbole de l'essor des entreprises ayant misé sur le « capitalisme cognitif », selon l'expression de Yann Moulier Boutang[2], par rapport aux firmes de l'âge industriel centrées sur l'exploitation des ressources rares.

L'activité des bibliothèques universitaires est directement affectée par cette évolution et cela n'est pas sans incidence sur les valeurs et principes à défendre collectivement par les professionnels du secteur. À première vue, les bibliothèques semblent naturellement placées du côté de l'accès à la connaissance et de la diffusion des savoirs, mais cette vision des choses est trop réductrice pour cerner tous les enjeux de la question telle qu'elle se pose aujourd'hui.

EXPLOITATION PRIVATIVE *VS* PARTAGE DU SAVOIR

Car il existe en réalité deux économies de la connaissance et de la culture, très différentes l'une de l'autre. La première – la plus visible – est centrée sur l'exploitation privative du savoir et de la culture, incarnée par des firmes comme Apple, Google, Facebook, Amazon, de grands groupes d'édition scientifique ou des laboratoires, utilisant des titres de propriété intellectuelle (droit d'auteur, marques, brevets) pour capter et concentrer la valeur. L'autre économie de la connaissance est organisée au contraire autour de l'ouverture et du partage du savoir, avec des exemples comme ceux du logiciel libre, de Wikipédia et des initiatives diverses œuvrant pour le développement de biens communs de la connaissance[3]. Ces deux économies s'appuient parfois l'une sur l'autre, mais il existe une tension et même un conflit latent entre ces deux logiques, qui sera sans doute un des traits dominants du XXIe siècle.

Ce qui se joue dans cet affrontement est un nouvel acte du « second mouvement des enclosures », selon l'expression du juriste américain James Boyle, qui après avoir frappé les ressources naturelles, a touché à partir du XIXe siècle la connaissance[4]. L'information scientifique et technique a été progressivement touchée par cette tendance, jusqu'à aboutir à la situation actuelle d'un marché dominé par un oligopole d'éditeurs scientifiques. L'augmentation exponentielle du coût des abonnements aux revues électroniques en résultant place les bibliothèques universitaires et de recherche dans une situation complexe qui brouille les valeurs défendues jusqu'ici par les professionnels de l'information.

VERROUILLEURS *VS* DONNEURS D'ACCÈS

En effet, les bibliothèques peuvent constituer un rouage à part entière de ce système d'enclosures sur le savoir et l'objectif consensuel de « donner accès à la connaissance » a profondément changé de nature avec le basculement dans l'environnement numérique. En contrôlant l'accès par les communautés qu'elles desservent aux ressources qu'elles acquièrent, les bibliothécaires se transforment de « donneurs d'accès » en « verrouilleurs d'accès ». Desservir un public donné revient à exclure les autres de l'accès à la ressource et une partie des compétences professionnelles mises en œuvre par les bibliothécaires se sont reconfigurées autour de ces techniques de contrôle visant à garantir que les ressources cognitives restent rares dans un environnement d'abondance.

Un geste accompli par l'activiste américain Aaron Swartz en 2011 a violemment mis en lumière les contradictions dans lesquelles cette situation plonge les bibliothèques et l'ensemble des professionnels de l'information. Profitant des codes d'accès qui lui étaient fournis par le Massachusetts Institute of Technology (MIT) à Harvard, Aaron Swartz téléchargea plus de 4,5 millions d'articles scientifiques figurant dans la base de données JSTOR avec l'intention de les libérer sur Internet, dans le but de dénoncer les restrictions d'accès à la connaissance.

[1] Nil Sanyas, « Google, deuxième capitalisation boursière entre Apple et Exxon », *Next INpact*, 11/02/2014 : www.nextinpact.com/news/85867-google-deuxieme-capitalisation-boursiere-au-monde-entre-apple-et-exxon.htm

[2] Cf. « Le capitalisme cognitif », *Multitudes* n° 32, éd. Amsterdam, printemps 2008.

[3] Sur cette notion, voir *Libres savoirs : les biens communs de la connaissance*, C&F Éditions, 2011 : http://cfeditions.com/libresSavoirs

[4] James Boyle, *The Public Domain : enclosing the commons of the mind*, Yale University Press, 2008.

Dans son *Guerilla Open Access Manifesto*[5], il justifie cette action en faisant appel à des considérations d'ordre moral :
« Vous qui avez accès à ces ressources, étudiants, bibliothécaires, scientifiques, on vous a donné un privilège. Vous pouvez vous nourrir au banquet de la connaissance pendant que le reste du monde en est exclu. Mais vous n'êtes pas obligés – moralement, vous n'en avez même pas le droit – de conserver ce privilège pour vous seuls. Il est de votre devoir de le partager avec le monde […] *Nous avons besoin de récolter l'information où qu'elle soit stockée, d'en faire des copies et de la partager avec le monde. Nous devons nous emparer du domaine public et l'ajouter aux archives. Nous devons acheter des bases de données secrètes et les mettre sur le web. Nous devons télécharger des revues scientifiques et les poster sur des réseaux de partage de fichiers. Nous devons mener le combat de la guérilla pour le libre accès ».*
Poursuivi pour ces faits par la justice américaine, A. Swartz risquait plus de 35 ans de prison et un million de dollars d'amende, ce qui le conduisit à mettre fin à ses jours. Des centaines de bibliothèques universitaires et de recherche dans le monde étaient et sont pourtant toujours abonnées à la base de données JSTOR, ainsi qu'à de nombreuses autres ressources similaires. Cet épisode tragique a montré que de très lourdes questions éthiques portent sur l'activité des bibliothèques, qui doivent être regardées en face.

Le professeur Grasset et sa « balance aux responsabilités ».
Image de B. Moloch, parue dans la revue *Chanteclair*, 1908. L'ensemble de la revue *Chanteclair* vient d'être numérisée par la BIU Santé et sera mise en ligne sous licence ouverte d'ici la fin 2014.

CONTRIBUER AUX BIENS COMMUNS DE LA CONNAISSANCE

Bien entendu, les bibliothèques peuvent aussi apporter une contribution décisive à la préservation et à la formation de biens communs de la connaissance. Elles le font, par exemple, en soutenant le mouvement de l'*open access*, par lequel les chercheurs essaient de reprendre le contrôle sur leurs publications. Elles le font également dans le cadre de l'activité des consortiums d'achat de ressources électroniques, afin de peser face aux grands groupes d'édition pour élargir les conditions de mises à disposition. Elles peuvent encore le faire en numérisant leurs propres collections appartenant au domaine public, à condition de ne pas tomber dans le « copyfraud » en entravant la réutilisation des fichiers par le rajout de nouvelles couches de droits[6]. Les bibliothèques ont aussi un rôle décisif à jouer en matière d'*open data*, en ouvrant à la réutilisation les données bibliographiques qu'elles produisent.

Au cours de la période qui s'ouvre, les débats seront sans doute nombreux à propos des valeurs à défendre par les bibliothécaires au sein des deux économies de la connaissance qui s'opposent, sachant qu'elles sont souvent inextricablement entremêlées. Les discussions soulevées cette année par la conclusion d'un accord entre Couperin et Elsevier en vue d'aboutir à une licence nationale pour la mise à disposition du corpus *Science Direct* ont été révélatrices à cet égard de ces tensions. Les licences nationales participent incontestablement au processus de reprise de contrôle sur l'information scientifique et technique. Mais fallait-il accepter le principe des clauses de confidentialité ou les dispositions portant sur le *text & data mining*, qui constituent pour cet éditeur une manière de maintenir une enclosure sur les données scientifiques[7] ?
On ne peut apporter de réponses à de telles questions en restant simplement sur le plan technique et il importe que de vrais espaces d'échanges s'ouvrent pour que les bibliothécaires, entre eux et avec d'autres communautés concernées, interrogent leur éthique professionnelle et la réinventent pour faire face à ces défis.

Lionel Maurel
Conservateur à la BDIC
Auteur du blog S.I.Lex
calimaq@gmail.com

[5] *Guerilla Open Access Manifesto* / Manifeste de la guerre pour le libre accès : **http://openaccessmanifesto.org/manifeste-de-la-guerilla-pour-le-libre-acces**

[6] Voir, par exemple, le choix en faveur de l'ouverture fait notamment par la BNU ou par la BIU Santé en utilisant la licence ouverte pour la diffusion de leurs documents numérisés : **www.bnu.fr/collections/la-bibliotheque-numerique/les-images-de-la-bnu-et-la-licence-ouverte** et **www.bium.univ-paris5.fr/histmed/medica_pres4.htm**

[7] Cf. Pier-Carl Langlais, « Faut-il signer l'accord avec Elsevier ? », *Sciences communes*, 12/02/2014 : **http://scoms.hypotheses.org/119**

MODÈLES ÉCONOMIQUES ET ÉCONOMIE DE LA CONNAISSANCE : APPROCHES DU GFII

Le Groupement français de l'industrie de l'information (GFII) rassemble les acteurs de la chaîne de l'information spécialisée. Les mutations récentes des modèles économiques dans le champ de l'information l'ont conduit à réactiver son groupe de travail consacré à l'économie de la connaissance, aujourd'hui animé par Olivier Delteil du groupe Les Échos.

Face à une forte demande de contenus ouverts, dans tous les secteurs, les producteurs et éditeurs de contenus et de services professionnels, publics ou privés, sont amenés à proposer, à côté des abonnements payants, de nouveaux modèles économiques, pour permettre une meilleure diffusion des connaissances, tout en garantissant la qualité et la pérennité des informations produites.
On se cantonnera ici à lister quelques modèles récents, sans reprendre ceux bien connus autour de l'abonnement et ses différentes formes actuelles, de l'*open access* ou des services à forte valeur ajoutée.

Le développement de stratégies de plateforme

Les acteurs/utilisateurs ont accès librement à des plateformes proposant des contenus réutilisables ou des services. Espaces ouverts d'échanges collaboratifs, elles permettent de créer des écosystèmes innovants, basés sur la puissance de création de la multitude, dont bénéficie directement le créateur de la plateforme (cf. la nouvelle version de data.gouv.fr). Les plateformes dominantes du marché (Google, Amazon, Facebook, Apple) rappellent que des modèles sont déjà proposés et cherchent à imposer leurs stratégies.

Un modèle bien connu, le freemium

Association d'une offre gratuite, en libre accès et d'une offre premium, haut de gamme, en accès payant, il offre des déclinaisons diverses. (cf. *Open Edition, Le Monde, Les Échos...*). Il est souvent développé en complément d'un modèle publicitaire, jugé aujourd'hui trop fragile pour assurer la pérennité des services, associé à des analyses marketing *« data driven »* très poussées : les données sont au centre de la construction de l'offre et des services pour enrichir l'expérience utilisateur, analyser les performances et surtout optimiser les parcours de transformation afin de susciter l'instant propice où l'utilisateur de contenus gratuits bascule vers le payant.

La mise à disposition des données via des API

C'est le modèle qui se développe le plus rapidement aujourd'hui. Il permet aux producteurs de construire des offres de services différenciées, avec des tarifications qui peuvent être graduées selon la valeur ajoutée apportée aux données, la complexité de l'API (interface de programmation) et de son intégration dans le *workflow* des utilisateurs, la spécialisation des usages... (cf. API de l'IGN). Si certaines API sont gratuites, d'autres sont payantes, suivant des modèles variés : volume des données ou des transactions, nombre de sollicitations, affiliation, syndication de contenus... Les API offrent souplesse et qualité de service pour la fourniture des données, mais les modes de facturation complexes qu'elles permettent ne sont pas toujours adaptés au système de facturation actuel des organisations. Le développement des API, notamment gratuites, a pour conséquence l'émergence de nouveaux acteurs, spécialisés dans leur agrégation (*mash up*).

L'apparition du financement participatif (crowdfounding)

Surtout observé dans le domaine professionnel, le *crowdfounding* a été évoqué par M. Trojette dans son rapport sur l'ouverture des données publiques comme une piste pour le financement des investissements nécessaires à la création de plateformes de mise à disposition d'informations publiques. Des exemples concrets sont fournis par l'initiative de l'éditeur De Gruyter avec Unglue.it[1] ou de Reveal Digital avec la création d'un site de *crowdfunding* pour mettre à disposition des bibliothèques des collections en *open access*[2].

Au-delà de ces différents modèles, les premières réunions du groupe de travail du GFII ont mis en lumière les préoccupations des producteurs d'information concernant le financement à long terme des politiques « o*pen* » avec le risque de la disparition de la production de certaines données, mais également le bouleversement de la chronologie des médias, la question du partage entre les différents acteurs de la chaîne de valeur à l'heure des *mash up*, le risque du piratage...
Pour continuer à répondre aux attentes des utilisateurs, les producteurs d'information doivent, plus que jamais, innover dans les contenus, mais aussi dans les services associés, le design et l'ergonomie sans oublier les modèles économiques afin de tenir compte des contraintes financières de leurs clients, dans leur diversité.

Olivier Delteil
Responsable du développement des activités numériques du groupe Les Échos

Ruth Martinez
Déléguée générale du GFII
www.gfii.fr

[1] https://unglue.it/pid/popular/4311
[2] http://voices.revealdigital.com

Acquisitions Istex : *traiter et enrichir les données*

L'enrichissement et la valorisation des données fournies par les éditeurs est une des nombreuses tâches gérées par le département Adéle et l'équipe du hub de métadonnées de l'Abes. Méthodes et enjeux.

Dans le cadre du projet Istex (Initiative d'excellence de l'information scientifique et technique) initié par le ministère de l'Enseignement supérieur et de la Recherche, l'Abes, *via* son département achat de documentation électronique (Adéle), passe des marchés avec des éditeurs scientifiques afin d'acquérir des données pour la communauté scientifique française.

DONNEZ, DONNEZ LES MÉTADONNÉES

Par « données », il faut bien sûr entendre à la fois les contenus – bases de données, collections de revues numériques, e-books, etc. – et les métadonnées relatives à ces mêmes contenus. Effectivement, l'acquisition de métadonnées est un élément particulièrement discuté. Et pas seulement leur acquisition d'ailleurs. Car ces métadonnées, il faut pouvoir les manipuler, les traiter, les enrichir, les redistribuer. Au-delà de l'acquisition, ce sont donc surtout les droits d'utilisation et d'exploitation de ces données qui sont négociés. Le contrat de licence, signé par l'éditeur lors de la passation du marché, définit d'ailleurs précisément ces règles d'utilisation. Il permet ainsi de « *fournir les métadonnées à l'ensemble des bénéficiaires afin qu'ils puissent les intégrer dans leur catalogue local ou leur outil de découverte* ». Le contrat prévoit également la modification du format des métadonnées et leur enrichissement. Concernant leur format, la licence n'impose pas de standard particulier. S'il est préférable que les métadonnées soient livrées en XML, l'éditeur garde néanmoins la possibilité de fournir des données en format Marc. En revanche, le contrat insiste sur la nécessité d'inclure « *l'intégralité des informations bibliographiques disponibles* » (l'expression apparaît plusieurs fois dans le document, c'est dire !). Conclusion : le format importe peu, finalement, tant que l'éditeur nous procure l'ensemble des données qu'il détient sur ses propres contenus.

LE RÔLE DU HUB DE MÉTADONNÉES

• *Avant la signature du marché*

En collaboration avec Adéle, le hub de métadonnées joue un rôle important dans les acquisitions Istex, en intervenant dès la phase de négociations. En effet, le hub est impliqué dans la constitution des listes de titres à acquérir qui sont proposées aux éditeurs, en réponse à leurs premières offres de contenus. Bien souvent effectivement, les premières listes de titres transmises par les éditeurs sont erronées ; bibliographiquement parlant, s'entend. Pour les listes de périodiques, par exemple, dont nous connaissons bien la complexité – une revue, au contraire d'un e-book, a en effet une vie bien mouvementée. Le chemin parcouru, depuis sa naissance, n'est jamais simple : fusion avec un titre cousin, scissions en un, deux, trois titres, voire plus, etc. – l'éditeur ne fournit généralement que le dernier titre paru d'un périodique, alors que les dates transmises correspondent bien souvent à un ensemble de titres. Le travail du hub consiste donc à reconstituer l'historique des revues proposées. Pour ce faire, un outil a été développé : il s'agit de *Métarevues*, un programme bien pratique, basé sur les données Sudoc, qui permet de lister dans l'ordre chronologique, tous les titres liés à la revue proposée, à partir d'un ISSN donné. Cet outil est d'ailleurs utilisé par l'application *Périscope* de l'Abes – en plus d'afficher les états de collection des périodiques signalés dans le Sudoc, le service permet désormais de visualiser les historiques complets liés à ces revues[1].

Bien souvent, les premières listes de titres transmises par les éditeurs sont erronées ; bibliographiquement parlant, s'entend.

• *Après la signature du marché*

Après la signature du marché, lorsque les listes de contenus ont été fixées et que l'ensemble des données a été livré, le hub procède au travail de signalement des ressources, et pas seulement dans le Sudoc, même si cela reste l'une des priorités. Des données de qualité, standardisées, sont ainsi transmises aux établissements bénéficiaires et aux éditeurs de bases de connaissance, afin d'assurer le signalement de ces ressources dans d'autres environnements que le Sudoc.

En redistribuant ces données vers d'autres outils de recherche documentaire, le hub permet le développement de l'accessibilité et de la visibilité des contenus – et par là même des auteurs et des édi-

[1] Pour plus d'informations, consulter le billet sur le blog technique de l'Abes, Punktokomo : **http://punktokomo.abes.fr/2014/06/26/metarevues-un-outil-dedie-au-traitement-des-periodiques**

teurs de ces contenus – acquis en licence nationale. C'est également un moyen de sensibiliser les éditeurs au travail bibliographique.
Après l'étape de signalement, le hub de métadonnées travaille aux enrichissements de ces données. D'autres informations viendront s'y ajouter. Ces informations complémentaires peuvent provenir de différentes sources. Il s'agira, par exemple, d'ajouter des liens à des autorités comme IdRef, Viaf, Rameau, ou encore d'insérer des codes de langues…

L'HOMME ET LA MACHINE

Si l'ensemble de ces traitements est grandement facilité par l'exploitation d'outils et de programmes informatiques, l'analyse humaine reste indispensable pour détecter certaines anomalies et penser les corrections nécessaires. En outre, chaque type de ressources comprend son lot de difficultés. Le traitement des données de périodiques, par exemple, impose plutôt un travail d'expertise en amont, lors de la constitution des listes contractuelles, tandis que l'effort majeur livré sur le traitement des données d'e-books se concentre un peu plus sur l'aval de la chaîne, lors de l'étape de signalement. Le hub doit générer lui-même des notices Marc à partir des métadonnées transmises par les fournisseurs, au contraire des périodiques. En effet, grâce aux notices du centre ISSN déjà présentes dans le Sudoc, le travail de signalement des revues ne consiste généralement qu'à ajouter des données d'exemplaires et compléter certaines données bibliographiques, comme celles de la zone 207 (informations données sur les dates et les numéros de volumes).
Au-delà des difficultés liées à la nature de la ressource traitée, il faut également bien noter que chaque nouveau corpus fraîchement débarqué peut également (se) présenter à chaque fois (avec) ses propres complexités, (avec) ses propres cas particuliers. À ce véritable travail sur mesure s'ajoute une autre complication : puisque plusieurs négociations sont menées de front, l'équipe du hub de métadonnées – du reste relativement réduite – doit faire face à plusieurs corpus simultanément. Dans ce contexte, il est difficile de dégager des priorités, ce qui rend l'organisation du travail assez complexe. Car en établissant des priorités, on sacrifie forcément certaines tâches : faut-il privilégier le travail en amont de la signature ou le travail en aval ? Faut-il traiter le maximum de corpus le plus tôt possible ou viser un travail de qualité pour chacun d'entre eux, quitte à en retarder le signalement ?

Jean-Raphaël Guillaumin / Flickr (CC BY-SA 2.0)

Du pollen au miel : une orgie de données pour les abeilles du hub !

UNE COOPÉRATION AVEC L'INIST RENFORCÉE

En guise de conclusion, il convient de noter que la collaboration Abes-Inist doit s'intensifier dès la rentrée de septembre 2014, laquelle permettra d'optimiser le travail sur le traitement des données. L'Inist est effectivement l'un des acteurs majeurs du projet Istex. Il travaille notamment à la validation des données livrées et à la réalisation de la plateforme qui accueillera l'ensemble des contenus et des données acquis en licence nationale et dont la mise en œuvre est prévue pour fin 2015.

MARION GRAND-DÉMERY
Abes
grand-demery@abes.fr

LE HUB, QU'ES AQUÒ ?

Pour rappel, le hub de métadonnées est un projet porté par l'Abes qui consiste à mettre en place des méthodes et des outils d'aide au traitement des métadonnées issues des éditeurs, et notamment de celles provenant des achats Istex. Ces procédures et ces outils sont sans cesse améliorés et développés au fur et à mesure de leur utilisation au sein même du hub, puisque ses équipes continuent effectivement de travailler sur le traitement des corpus dans un même temps. C'est d'ailleurs l'une des particularités du hub, qui se fait donc, au-delà du projet, véritable service en activité.

L'ISNI (*International Standard Name Identifier*/Code international normalisé des noms, ISO 27729:2012) est un identifiant international ISO attribué aux identités publiques[1] des personnes et des collectivités. Il permet de simplifier les échanges d'informations entre les producteurs de données de tous secteurs (musique, littérature, cinéma, spectacle vivant, recherche scientifique, etc.).

L'ISNI prend son essor

Impliquée dans la gestion de cet identifiant depuis ses débuts, la Bibliothèque nationale de France (BNF) l'a intégré dans ses propres données et en est devenue une agence d'enregistrement en janvier 2014.

FIABILITÉ DE L'ATTRIBUTION

L'identification des identités publiques se fonde sur des données fournies par des contributeurs de multiples domaines, rapprochées automatiquement par algorithmes, en une base centrale unique. L'attribution de l'identifiant ISNI est conditionnée par la présence de métadonnées discriminantes – noms et variantes de noms, dates, titres, identifiants associés, etc. – et n'est effectuée qu'à partir d'un certain seuil de confiance.

Dans l'ISNI, les informations provenant de bases de données variées sont rapprochées sur le principe de complémentarité. S'y confrontent des fichiers d'autorité, des bases bibliographiques, des registres de sociétés de gestion de droits ou d'organismes de recherche et professionnels. Les données des partenaires de VIAF (*Virtual International Authority File*/Fichier d'autorité international virtuel) – 40 institutions en juillet 2014 dont l'Abes et la BNF – sont présentes dans la base ISNI depuis son lancement et sont mises à jour régulièrement. Elles sont reliées à celles de 30 autres contributeurs[2].

La qualité de la base internationale ISNI est contrôlée par une équipe d'experts de la British Library et de la BNF, qui effectue des campagnes de tests et répond aux demandes de corrections faites par les utilisateurs. La base ISNI est en effet publique[3] et il est possible, depuis chaque notice, d'envoyer des suggestions d'amélioration à l'équipe qualité. Les techniques de rapprochement de données, élaborées par OCLC, l'agence d'attribution qui gère la base centrale, sont également en constante amélioration.

En fondant l'identification sur des données d'un haut niveau de complétude, contrôlées et mises à jour, l'ISNI contribue ainsi à la fiabilité des informations présentes sur le web. Plus encore, la gestion de l'identifiant ISNI, garantissant son unicité et sa pérennité, en fait l'un des outils indispensables du web sémantique.

ENGAGEMENT DE LA BNF

Les garanties de stabilité technique et administrative de l'ISNI sont assurées par un réseau d'acteurs agréés : une autorité d'enregistrement – l'Agence internationale ISNI (ISNI-IA) –, une agence d'attribution (OCLC Leiden) et des agences d'enregistrement[4]. La BNF a participé aux travaux de normalisation de l'ISNI à partir de 2006 et s'est depuis impliquée à divers titres dans ce réseau. Elle représente la Conference of European National Librarians (CENL), de concert avec la British Library, au comité des directeurs de l'ISNI-IA.

Plus récemment, la BNF a introduit l'ISNI dans ses données et a mis en place une agence d'enregistrement[5] pour son attribution aux contributeurs qu'elle identifie dans ses bases de données (les personnes ou collectivités pour lesquelles elle demande l'attribution doivent faire l'objet de notices d'autorité dans son catalogue général[6]). La volumétrie des ISNI attribués aux données de la BNF dans la base centrale ISNI s'est accrue depuis 2011 : plus d'un million d'ISNI lui étaient attribués au début de l'année 2013 auxquels se sont ajoutés 50 000 nouveaux ISNI depuis. Cette volumétrie annuelle devrait être dépassée en 2014, grâce à la mise en place de l'agence d'enregistrement. Tous les identifiants ISNI intégrés aux données de la BNF sont à la disposition de ses réutilisateurs au titre de la politique d'ouverture et de gratuité des données[7] de l'établissement. Ils sont diffusés *via* les produits courants et rétrospectifs et le transfert de notices en ligne. Ils peuvent être également récupérés par l'intermédiaire de data.bnf.fr, en RDF[8] et interrogés par requêtes SPARQL[9].

BÉNÉFICES ATTENDUS

L'identification des contributeurs (auteurs, participants, éditeurs, producteurs…) est une des grandes problématiques actuelles des acteurs du numérique. Dans ce contexte, l'ISNI facilite la navigation et les échanges de données. Au-delà, il peut aboutir à une coproduction de données et permet de rationaliser les opérations de gestion. Chaque identifiant ISNI est conçu comme une passerelle entre divers jeux de données et permet leur interconnexion. En juillet 2014, pour 8 millions d'identifiants attribués dans la base internationale, 16 millions

[1] Selon la norme ISO 27729, une identité publique est « *une identité utilisée publiquement par des individus ou des collectivités impliquées dans le processus de création, de production, de gestion et de distribution d'un contenu* » : **www.iso.org/iso/catalogue_detail?csnumber=44292**

[2] **www.isni.org/content/data-contributors**

[3] **www.isni.org/search**

[4] Actuellement la BNF, Bowker et Ringgold : **www.isni.org/content/isni-registration-agencies**

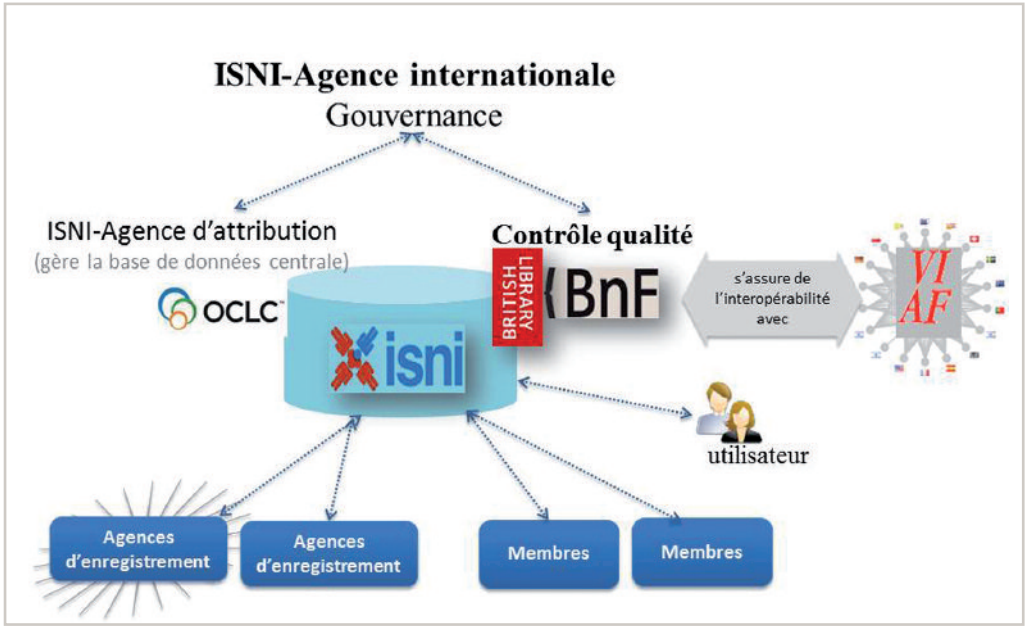
[5] L'agence d'enregistrement fait partie du département de l'Information bibliographique et numérique (Direction des services et des réseaux).

[6] **www.bnf.fr/fr/professionnels/isni_informer/s.isni_agence_bnf.html?first_Art=non**

[7] **www.bnf.fr/fr/professionnels/anx_recuperation_donnees/a.ouverture_donnees_bnf.html**

[8] **http://data.bnf.fr/semanticweb**

[9] **http://data.bnf.fr/sparql**



Le réseau d'acteurs de l'ISNI.

de liens sont ainsi déjà créés entre les contributeurs de l'ISNI, assurant entre eux une navigation fiable et puissante au profit des utilisateurs. L'offre documentaire de la BNF s'en trouve enrichie puisqu'il est possible d'élargir la recherche bio-bibliographique à partir du catalogue général ou de data.bnf.fr vers d'autres réservoirs de données qui utilisent l'ISNI.

Des liens sont faits tant avec des bases publiques, gérées par des institutions patrimoniales, qu'avec des bases privées et propriétaires auxquelles l'ISNI offre une solution d'interopérabilité. Il permet en effet de respecter la politique de chaque institution en matière d'exposition de données, en masquant certaines informations à la vue du public. Les fonctionnalités de l'ISNI répondent ainsi au besoin d'interconnexion entre les bases de données publiques et privées, considéré comme un enjeu national en France par le rapport Lescure[10] (proposition 79) en 2013.

En rapprochant les données provenant de différentes sources, l'ISNI aide à leur consolidation et à leur amélioration : des rapports d'erreurs provenant de l'ISNI sont utilisés pour rectifier des données fautives ou manquantes dans le catalogue de la BNF (dates biographiques, confusion d'homonymes, doublons).

En outre, son utilisation courante par les partenaires de la BNF permettra à terme d'accélérer et de fluidifier les processus de création de données : son usage dans les échanges relatifs au dépôt légal (traditionnel et numérique) entre les éditeurs et la BNF se mettra progressivement en place à partir de 2015 et accélèrera certaines étapes du catalogage. À mesure de la diffusion de l'ISNI chez les partenaires de la BNF, d'autres cas d'utilisation pourront être expérimentés dans les circuits de numérisation et pour le calcul et la distribution des droits de propriété intellectuelle.

DIFFUSION EN FRANCE

L'agence ISNI de la BNF partage son expertise avec les différents organismes français intéressés par l'adoption de l'ISNI, pour construire des circuits de données envisageables techniquement et financièrement. Son adoption par les producteurs de données en France s'accélère dans les secteurs de l'interprofession du livre (éditeurs, distributeurs, etc.), des institutions culturelles et de l'information scientifique et technique.

La diffusion de l'ISNI est soutenue par le département des Programmes numériques du ministère de la Culture. Elle permettra d'améliorer l'interopérabilité des bases de données publiques entre elles[11], de conférer une visibilité aux auteurs de ressources culturelles identifiés dans ces bases de référence, qu'ils soient déjà reconnus ou non à l'échelon international et, par ce biais, d'accroître la présence des données culturelles françaises sur le web.

Les mêmes atouts d'interopérabilité et de visibilité pourront bénéficier aux bases françaises de l'information scientifique et technique pour les personnes (chercheurs, auteurs de thèses, etc.) et collectivités (universités, unités de recherche, laboratoires, etc.) qu'elles identifient. L'ISNI permet aussi de rapprocher ces mêmes données avec celles des bases culturelles[12], comme les catalogues de bibliothèques publiques ou celui de la BNF, par exemple.

Pauline Chougnet
Responsable de l'agence d'enregistrement ISNI, BNF
pauline.chougnet@bnf.fr

[10] Pierre Lescure, Mission « Acte II de l'exception culturelle ». *Contribution aux politiques culturelles à l'ère numérique*, 2 volumes, mai 2013. En ligne : www.culturecommunication.gouv.fr/var/culture/storage/culture_mag/rapport_lescure/index.htm#/1

[11] http://fr.slideshare.net/culturefr/feuille-de-route-mtadonnes-et-30-camille-domange

[12] Jérôme Kalfon, « La place des réseaux de l'Abes à l'heure du décloisonnement », *Arabesques*, n° 74, avril-mai-juin 2014. En ligne : www.abes.fr/Arabesques/Arabesques-n-74

ARTICULATION ISNI-IDREF : UN ENJEU POUR L'IDENTIFICATION PÉRENNE

L'identification pérenne dont il est question ici concerne les auteurs, qu'ils soient des personnes physiques ou morales. IdRef y est restreinte à ses deux référentiels « Personnes » et « Collectivités ». Cette perspective ne doit pas faire oublier qu'existent dans IdRef, reflet des données d'autorités du Sudoc, d'autres types de référentiels pour les sujets, les toponymes, les ressources documentaires, etc., et leur caractère tout aussi indispensable.

Au sein de l'enseignement supérieur et de la recherche, la pression à évaluer et à valoriser la production bibliographique de la recherche apparaît en forte croissance depuis la fin de la décennie 2000. La visibilité, l'attractivité et la vitalité d'un établissement comptent parmi les enjeux majeurs liés à cette pression. Ceci met face à un défi de taille : comment les institutions peuvent-elles mesurer la totalité de la production scientifique de leurs chercheurs avec la plus haute précision ? En effet, les obstacles à un recensement fiable et exhaustif sont nombreux : la même publication peut être diffusée sur plusieurs bases de données ; des formes différentes d'une même publication peuvent apparaître sur une même base de données ; la même information sur un chercheur spécifique peut être présente dans plusieurs bases de données ; un même chercheur peut apparaître sur des bases distinctes avec des informations partiellement différentes ; les bases de données peuvent ne pas interagir ou n'interagir que partiellement les unes avec les autres. Semblables situations ne sont pas rares et forcent invariablement à dresser le constat de la perte de temps, d'énergie et d'information.

Des liens consolidés entre les ressources et leurs auteurs

Au cours de ces dernières années, l'Abes, comme bien d'autres acteurs de l'enseignement supérieur et de la recherche, a observé cette évolution et a progressivement été conduite à étudier les moyens de faciliter pour les chercheurs et les institutions la mesure fidèle et précise de leur production scientifique. À l'instar d'un certain nombre d'initiatives internationales actuelles et avec déjà quelques années d'expérience, l'Abes a placé sa réponse sous un éclairage particulier tant du point de vue technologique que stratégique : la recherche de fiabilité et d'exhaustivité passe par la construction de référentiels d'auteurs et par l'attribution d'identifiants uniques et pérennes, afin de générer des liens consolidés entre les ressources et leurs auteurs. L'application IdRef est l'une des réalisations tangibles de ce choix. Que l'on soit un internaute lambda ou un professionnel, que l'on soit un moteur de recherche ou une application documentaire (Calames ou Star), il est possible d'identifier et de lier un contributeur et sa production dans les catalogues de l'Abes, à partir des identifiants uniques et pérennes qu'elle propose (voir par exemple : Théodore Monod, **www.idref.fr/027035166**).
En fait, on retrouve ici une pratique traditionnelle des bibliothèques : la constitution du fichier d'autorités assurant le contrôle des variantes de nom d'un même auteur et la désambiguïsation des auteurs homonymes ; laquelle est soutenue par les standards les plus récents du web sémantique visant à décloisonner les silos bibliographiques pour produire recoupements et associations d'informations relatives aux producteurs et aux productions scientifiques.

Des passerelles pour garantir l'interopérabilité

D'autres acteurs et d'autres initiatives semblables apparaissent par ailleurs, dont l'ISNI[1]. Loin d'être antagonistes, les différents référentiels construisent un écosystème qui donne chair au web de données. Tandis qu'un identifiant unique au sein d'un référentiel unique est voué à l'incomplétude, à l'isolement et/ou à l'obsolescence, l'interopérabilité et l'alignement des différents référentiels ouvrent des perspectives nouvelles pour la qualité des catalogues et pour l'offre de services aux usagers.
De façon très concrète, un chercheur français référencé dans les systèmes Sudoc ou thèses.fr et disposant d'un identifiant IdRef n'aura pas à se soucier de sa visibilité internationale puisque l'existence d'une passerelle à double sens entre les deux référentiels lui permettra de se voir attribuer un identifiant ISNI utilisé à son tour comme passerelle vers d'autres systèmes bibliographiques qu'ils soient publics ou privés. Notons que cette solution est déjà à l'œuvre : l'Abes et la Bibliothèque nationale de France, en tant qu'agence d'enregistrement, ont travaillé ensemble au déploiement des identifiants ISNI dans IdRef et ont commencé à tester de prometteurs alignements ISNI-IdRef.

François Mistral
Abes
mistral@abes.fr

[1] Cf. article précédent.

L'ISSN EN MUTATION

Le Centre international ISSN fêtera ses quarante années d'existence en 2015. Projets et stratégie d'ouverture.

Agence d'enregistrement du numéro international normalisé des publications en série (ISSN)[1], le Centre international ISSN s'appuie sur un réseau de 88 centres nationaux et administre le registre international ISSN rassemblant 1 587 422 notices de ressources imprimées et 132 806 notices de ressources en ligne (chiffres 2013). Il commercialise des produits et services en lien avec la base ISSN : outre l'accès aux notices sur abonnement, il a mis en place un service OAI-PMH pour une mise à jour automatisée des données par les clients ainsi que des traitements sur mesure comme la vérification complète de lots de données ou l'extraction de titres et données associées par éditeurs.

ROAD : un projet pilote

La réflexion stratégique, lancée en 2012 par la direction et le conseil d'administration, a abouti en 2013 à la création d'une nouvelle offre de service en ligne à partir des données de la base ISSN. ROAD (*Directory of Open Access scholarly Resources*)[2] est le portail des publications scientifiques (revues et dépôts institutionnels d'articles) en accès libre et ouvert. Disponible gratuitement, il a reçu le soutien financier de l'Unesco puisqu'il contribue à la politique des Nations unies en faveur de l'*open access*. Destiné aux chercheurs et étudiants, aux décideurs, journalistes et professionnels de l'information, il signale les ressources ouvertes par pays, sujet, type et ISSN. Son originalité réside dans la possibilité pour l'utilisateur de savoir si une ressource est indexée et évaluée par des services comme Scopus ou Latindex. Il peut également vérifier si celle-ci est conservée en tout ou partie pour un archivage pérenne grâce au lien vers The Keepers Registry[3], initiative internationale menée par Edina et financée par le JISC. Ses données au format Marc XML sont réutilisables sous la licence Creative Commons BY-NC-ND 2.0. Le site évoluera en 2014 avec la mise en œuvre du format RDF basé sur le modèle PRESSoo, développé conjointement par le Centre international ISSN et la Bibliothèque nationale de France.

Open access et linked data

ROAD annonce l'évolution future du registre international ISSN que le conseil d'administration du réseau ISSN, réuni en juin 2014, a validée. Dans un environnement numérique où l'information circule rapidement, peut être modifiée aisément et constitue une source non négligeable de revenus, l'importance des identifiants et des métadonnées associées est renforcée, ne serait-ce que pour des questions d'authenticité et de traçabilité. Depuis le début des années 2000, l'ISO a porté sur les fonts baptismaux de nouveaux identifiants comme l'ISWC (2001) et l'ISAN (2002) ou tout récemment l'ISNI (2012) et bientôt l'ISLI[4]. L'enjeu principal pour les agences qui administrent ces identifiants est d'encourager leur réutilisation la plus large possible dans les transactions entre acteurs de la production, de la consommation et de l'archivage des œuvres numériques.

Diffuser et partager globalement l'ISSN

Le réseau ISSN souhaite faciliter l'accès aux métadonnées ISSN à l'instar de ce que propose, par exemple, l'agence internationale ISAN[5] avec son service Lookup. Il veut renforcer sa capacité de repérage des nouvelles publications en s'ouvrant à un plus grand nombre de pays et en consolidant ses liens avec les éditeurs. L'amélioration de la qualité des métadonnées du registre est aussi un objectif primordial qui se traduit par l'élaboration d'un programme de révision et de correction et la mise en valeur des données dans leur langue d'origine. D'ores et déjà, le projet *Core Resources* vise à repérer en ligne les versions électroniques de périodiques scientifiques imprimés auxquelles des ISSN distincts doivent être attribués. Des métadonnées en caractères non latins sont introduites dans le registre ISSN en complément des éléments translittérés. La coopération avec les promoteurs des bases de connaissance internationales est également un axe de développement de l'ISSN. Enfin, la visibilité du réseau doit être améliorée grâce à son site web, sa présence sur les réseaux sociaux et une campagne de communication plus intense.

La croissance des publications en ligne, le développement de l'archivage numérique et la maturité des outils du web sémantique sont de réelles opportunités pour le centre international et le réseau ISSN qui s'engagent ensemble dans la définition d'un nouveau modèle d'activité fondé sur une offre de services améliorée et une transition technologique maîtrisée.

Gaëlle Béquet
Directrice du Centre international ISSN
gaelle.bequet@issn.org

[1] L'ISSN repose sur la norme ISO 3297 qui entrera en processus de révision en janvier 2015.
[2] http://road.issn.org
[3] http://thekeepers.org
[4] Norme actuellement en vote au sein de l'ISO.
[5] www.isan.org/lookup

La fouille de textes *au service de la documentation*

Les masses de données textuelles aujourd'hui disponibles engendrent un problème spécifique lié à leur traitement automatique. Des méthodes de fouille de textes et de traitement automatique du langage peuvent en partie répondre à cette difficulté. Approche des procédés et des nouveaux défis à relever présentés par deux chercheurs du Cirad, centre de recherche français qui répond, avec les pays du Sud, aux enjeux internationaux de l'agriculture et du développement.

Les processus de fouille de textes sont souvent composés de deux phases successives. Dans un premier temps, ces méthodes consistent à extraire les *descripteurs linguistiques* les plus significatifs à partir de documents (résumés, articles scientifiques, etc.). Les descripteurs linguistiques peuvent être des mots simples (par exemple, « culture »), mais aussi des termes composés (par exemple, « agriculture familiale »). Nous appellerons de tels descripteurs linguistiques des *termes*. Ces derniers représentent le matériau de base qui associe une certaine sémantique aux documents. Par exemple, les termes « riz », « irrigation » et « Madagascar » présents dans un document mettent en lumière une thématique générale liée à l'*agronomie*. La deuxième phase du processus consiste à exploiter ces termes pour, par exemple, classer automatiquement les documents dans des catégories (*agronomie*, *élevage*, etc.). Cette classification repose sur le postulat suivant : si des documents possèdent de nombreux termes en commun alors ils peuvent être regroupés et reliés à une même thématique. Sur la base de cette hypothèse et de mesures de similarité associées, des méthodes d'*apprentissage automatique* permettent, à partir de textes annotés manuellement par les professionnels de la documentation, de construire automatiquement un modèle et un algorithme de classification. Ce dernier pourra alors être appliqué avec succès sur de nouveaux documents, permettant ainsi de traiter des milliers de textes de manière automatique. Ces approches sont bien sûr très dépendantes de la qualité des termes issus des données textuelles. Ce point est discuté dans la section suivante.

L'EXTRACTION DE LA TERMINOLOGIE

Les méthodes d'extraction de la terminologie combinent harmonieusement les informations syntaxiques et statistiques. Les *informations syntaxiques* permettent d'identifier des termes ayant des structures syntagmatiques complexes (patrons de type nom-adjectif, nom-préposition-nom, etc.). Pour extraire des termes respectant de telles structures, il est nécessaire d'apposer, au préalable, une étiquette grammaticale à chacun des mots du texte avec des étiqueteurs dédiés (étiqueteur de Brill, Tree Tagger, etc.).

Par ailleurs, les *informations statistiques* apportent une pondération des termes candidats extraits afin de proposer les plus pertinents aux experts. En effet, la fréquence d'un terme n'est pas nécessairement un critère de sélection adapté. À titre d'exemple, le mot « agriculture » présent dans de très nombreuses publications du Cirad se révèle en fait très général et pas suffisamment *discriminant*. Ainsi, des mesures de discriminance et d'autres méthodes de pondérations qui calculent, par exemple, la dépendance des mots composant les termes complexes peuvent être appliquées. Elles ont notamment été utilisées dans le cadre d'une étude menée en 2014 qui s'est portée sur un corpus de résumés bilingues (français et anglais) extraits d'Agritrop[1], la base de données de publications scientifiques du Cirad. Une synthèse des résultats obtenus est présentée dans la section suivante.

LA FOUILLE DE PUBLICATIONS SCIENTIFIQUES

Les méthodes de fouille de textes décrites précédemment ont été appliquées et adaptées à l'aide de divers pré- et post-traitements. Elles ont permis d'extraire automatiquement des termes en français tout à fait pertinents tels que « développement durable », « développement rural », « sécurité alimentaire », « croissance démographique », « aménagement du territoire », « gouvernance territoriale », etc. Ainsi, 28 % des termes simples et 12 % des termes composés sélectionnés avec nos systèmes sont présents dans un thésaurus du domaine de l'agriculture, Agrovoc[2] (vocabulaire contrôlé issu de la FAO – Food and Agriculture Organization – des Nations unies).

Ceci met en exergue que, d'une part, les termes fournis par les approches de fouille de textes sont pertinents et que, d'autre part, des termes nouveaux qui sont souvent des termes composés sont également mis en avant. Ils peuvent enrichir des thésaurus spécialisés. Une étude qualitative avec

[1] http://agritrop.cirad.fr

[2] http://aims.fao.org/fr/standards/agrovoc

Le terme « Madagascar » peut signifier une entité autant spatiale que politique. Des approches de désambiguïsation sémantique s'avèrent donc nécessaire dans le cadre de la fouille de publications scientifiques.

Luc Legay / Flickr (CC BY-SA 2.0)

un professionnel des ressources documentaires a confirmé ces premières constatations.

Par ailleurs, en partenariat avec les pays du Sud, l'intérêt pour un organisme de recherche comme le Cirad est de disposer d'une représentation spatiale de son activité. Une piste d'exploration a été amorcée en combinant ces techniques de fouille de textes et la prise en compte d'informations spatiales extraites dans les articles. Ces entités sont alors mises en relation avec des référentiels géographiques (par exemple, GeoName[3]). Notons que des approches de désambiguïsation sémantique peuvent, au préalable, distinguer une entité spatiale d'une organisation (par exemple, selon le contexte, le mot « Madagascar » présent dans un texte peut être une entité spatiale ou une entité politique). Ainsi, l'ensemble de ce processus permet de mettre en relief les articles scientifiques (représentés par un identifiant) traitant de sujets précis (termes extraits) en lien avec des entités spatiales (par exemple, un pays ou une ville). Les résultats peuvent alors être diffusés sur des plateformes dédiées à la publication de données en *open data* qui offrent certains modules de visualisation.

VERS DE NOUVEAUX DÉFIS...

Pour traiter les masses de données aujourd'hui disponibles (c'est-à-dire l'infobésité), la problématique de recherche du *big data* est classiquement mise en avant avec les 3 V qui la caractérisent : volume, variété et vélocité. Mais d'autres caractéristiques ne doivent pas être négligées comme la versatilité, la véracité, la visualisation ou la valorisation des données et informations. Toutes ces problématiques ouvrent de nouvelles disciplines de recherche comme la *science des données* qui mêle mathématiques, statistiques, informatique et visualisation.

Dans ce contexte, une voie de recherche tout à fait prometteuse concerne le croisement des données hétérogènes qui met en perspective des connaissances nouvelles. Dans le cadre de nos travaux, ceci revient à mettre en relation les publications scientifiques du Cirad avec d'autres types de données disponibles (enquêtes, images, bases de données, etc.). Aussi, l'extraction et l'exploitation de mots-clés disciplinaires et thématiques sur de la donnée, au sens large du terme, est envisageable dès lors qu'il existe des référentiels grâce auxquels un couplage avec les processus de fouille de textes et de données restitue des résultats qualifiés. Par itération de ce processus sur des données hétérogènes en associant des mesures de similarité, la capitalisation de l'ensemble pourrait constituer demain un résultat majeur.

Mathieu Roche et Sophie Fortuno
Cirad, UMR TETIS, Montpellier
mathieu.roche@cirad.fr
sophie.fortuno@cirad.fr

Remerciements
Nous remercions les documentalistes du Cirad, les étudiants du master IPS (Université Montpellier 2) et de la licence professionnelle Management des ressources numériques (Université Paul Valéry) et Juan Antonio Lossio (doctorant au Lirmm) qui ont activement participé à ces travaux de recherche liés à la fouille de textes.

Metadata mining : *fouiller les données des catalogues ?*

Les grands catalogues de bibliothèques auraient-ils vocation à devenir des sources de *data mining* ? Pourquoi et comment ? Plaidoyer appuyé sur quelques cas d'usage rencontrés à la Bibliothèque nationale de France (BNF).

1

Les débats en cours sur le *text* et *data mining* (TDM), qu'ils soient opérationnels (comment fouiller, pour quelle recherche, quels services et avec quels outils ?) ou juridiques (de quel droit, à quel coût ?) embrassent potentiellement une grande variété de données aux périmètres en réalité très différents. On imagine ou on pratique déjà la fouille de données dans les données brutes de la recherche, dans les publications couvertes par les grandes bases de connaissance éditoriales et commerciales de l'information scientifique et technique, ou encore dans les bibliothèques numériques issues de la numérisation patrimoniale.

Pour le chercheur, il ne s'agit plus seulement de trouver et de lire des documents, mais aussi d'interpréter, d'analyser, de confronter et de « faire parler » par le calcul de grandes masses de données, et ce, dans des domaines disciplinaires qui recouvrent aussi bien les sciences dures que les sciences humaines et sociales, notamment dans le cadre du développement des « humanités numériques ».

LES BIBLIOTHÈQUES, NOUVEAUX LABORATOIRES DE FOUILLE

Pour les bibliothèques, un enjeu stratégique est de trouver leur juste place et leur valeur ajoutée propre dans l'invention de ces nouvelles pratiques et dans les dispositifs au sein desquels elles vont s'élaborer. Leur connaissance intime de la constitution et de la structuration des collections peut les conduire à devenir des fabricants et des fournisseurs de corpus de données à fouiller. Leur maîtrise du signalement et leur pratique du traitement informatique des catalogues et des données bibliographiques peuvent aussi les aider à structurer ces corpus et ces bases en les qualifiant, en les organisant. D'ores et déjà placées, par leurs missions, au plus proche des attentes et des pratiques de recherche des usagers, elles pourraient alors concevoir, gérer et animer par elles-mêmes les services et l'outillage de traitement des données.

Dans la bibliothèque du XXIe siècle, il faut donc imaginer (sur place et/ou en ligne) un laboratoire où un chercheur pourra venir fouiller de grands corpus de données et où le bibliothécaire sera en mesure de l'aider à effectuer et valoriser ces opérations en mettant à sa disposition des outils, des méthodes, des référentiels. Si les types de données à exploiter, par leur masse et leur nature, peuvent, pour certaines, représenter un changement radical de paradigme (on ne parle plus de livres, ni même de bases d'articles scientifiques) pour les professionnels, les fonctions et services à inventer s'inscrivent en réalité dans la continuité de leurs missions et de leurs pratiques.

DES MÉTADONNÉES À REVISITER : CAS D'USAGE

Cet horizon peut paraître lointain et futuriste, et pourtant, très près de nous, quelques cas d'usage m'ont récemment amené à penser que les bibliothèques, ou pour le moins les agences en charge de la maintenance des grands catalogues nationaux comme l'Abes et la BNF, pourraient d'ores et déjà proposer des services de *data mining* à partir de données que nous connaissons particulièrement bien car nous les avons créées : les notices bibliographiques et d'autorité, bref, les métadonnées de nos catalogues. Voici ces cas d'usage.

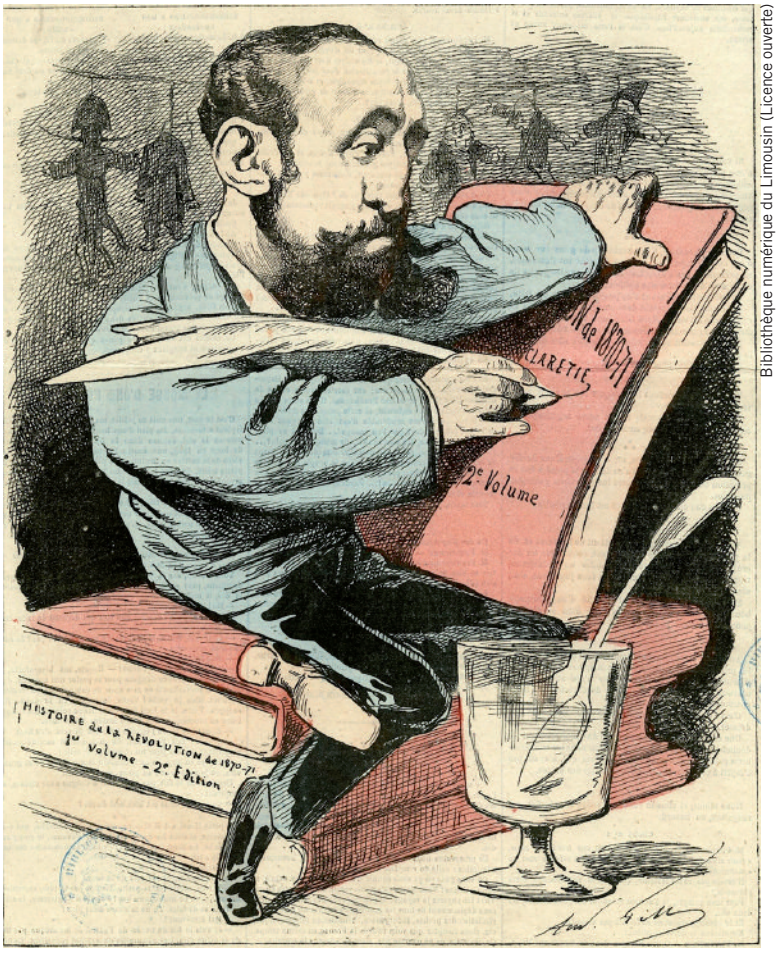
Un maître de conférence en sociologie dans un institut d'études politiques a récemment contacté la BNF car il conduisait une recherche sur la production de livres par les journalistes : pourquoi écrivent-ils ? Comment évolue cette production ? Afin de donner une dimension quantitative et statistique à son étude, il a eu recours au catalogue général de la BNF, dans lequel il a repéré près de 30 000 notices d'autorités correspondant à des journalistes. Pour ce faire, il a cherché à exploiter par lui-même de longues séries de notices d'autorité, mais s'est heurté à différents obstacles dont il s'est ouvert à ses correspondants de la bibliothèque.

À la même époque, une doctorante allemande dont le sujet de thèse est « *La circulation du savoir africain. Présence et réception de la littérature académique africaine en Allemagne et en France* » nous a également sollicités. Elle souhaitait relever et classer par nationalité actuelle l'ensemble des auteurs nés en Afrique, dans le domaine des sciences humaines, publiés en France depuis les années 60, dans n'importe quelle langue. Mais les fonctionnalités de recherche standard du catalogue de la BNF ne suffisent pas pour obtenir des données fiables et en masse, et pour tendre à l'exhaustivité :

il faut lancer des requêtes sur un ensemble de critères croisant les notices bibliographiques et les données d'autorités, requêtes pour lesquelles notre catalogue n'est absolument pas conçu. Dans un premier temps, nous avons essayé de l'assister en utilisant les outils professionnels habituels pour l'exploitation de nos métadonnées par d'autres bibliothèques ou des fournisseurs de systèmes de gestion de bibliothèque. Jusqu'à ce que l'on arrive au constat que ces outils n'étaient ni économiquement ni techniquement adaptés au besoin de cette doctorante.

ADAPTER LES OUTILS SIGNALÉTIQUES AUX NOUVEAUX BESOINS

Ces deux histoires nous ont ouvert les yeux sur une réalité : la demande de *data mining* est déjà à la porte de nos catalogues et nous devons prendre en compte les questions bien légitimes de ces nouveaux usagers dans l'évolution de nos dispositifs de signalement : pourquoi est-il si difficile d'extraire des notices en Marc des informations dans des formats simples d'exploitation bureautique par le commun des mortels ? Pourquoi, comme l'un de nos cobayes nous l'a fait remarquer, y-a-t-il « deux catalogues, l'un pour les autorités, l'autre pour les documents » ? Comment cerner dans le catalogue la population des journalistes alors qu'il n'y a pas de référentiel « emploi » dans les notices d'autorité personne ? Ces questions nous ont plus globalement interrogés sur la nécessité de rendre plus intelligible et plus explicite la structure du catalogue (et pas seulement des informations qu'il contient), de proposer des fonctionnalités d'extraction (sous forme de fichiers tabulés simples) et de retraitement massif de ses données adaptés à ce type d'usage, en donnant par exemple la possibilité aux chercheurs d'exploiter des catégorisations et indexations conçues initialement pour la gestion du catalogue et de la recherche documentaire, mais qui pourraient s'avérer extrêmement utiles pour ce type de traitement.
L'évolution engagée des catalogues vers les standards du web sémantique permettra sans nul doute d'aborder ces fonctionnalités de TDM au moyen d'outils plus adaptés que ceux des systèmes actuels, mais gardons d'ores et déjà cette idée : le catalogue de demain ne sera plus seulement un outil de recherche et de localisation d'information ; il faut aussi l'imaginer comme une vaste base de données à part entière que les chercheurs auront envie de faire parler.



Bibliothèque numérique du Limousin (Licence ouverte)

Jules Clarétie par André Gill, détail de la Une de *L'Eclipse*, n° 245, 6 juillet 1873. De nombreux journalistes, comme ici Jules Clarétie, sont également écrivains. Les métadonnées des catalogues des bibliothèques ne permettent pas d'effectuer un croisement judicieux entre ces 2 statuts à défaut de « référentiel emploi » dans les notices d'autorité personne.

GILDAS ILLIEN
Directeur du département de l'Information bibliographique et numérique, Bibliothèque nationale de France
gildas.illien@bnf.fr

LORSQUE LES MÉGADONNÉES CHASSENT LE BIG DATA

Le *Journal officiel* du 22 août 2014 a publié un avis de la Commission générale de terminologie et de néologie de l'informatique et des composants électroniques (CSTIC) qui propose d'utiliser le mot « mégadonnées », en lieu et place de l'expression anglaise « *big data* ».
Dès lors qu'il faudra parler du *big data*, c'est-à-dire ces « *données structurées ou non dont le très grand volume requiert des outils d'analyse adaptés* », les différentes administrations et autres établissements publics français devront adopter le terme « mégadonnées » proposé par les membres de la commission. L'expression « données massives » reste également admise.
Un peu plus tôt dans le mois (5 août), la même commission a préconisé de remplacer « *crowdfunding* », terme qui désigne le travail collaboratif d'internautes, par « production participative ».

www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029388087&dateTexte=&categorieLien=id

DES DONNÉES OUVERTES SUR L'ENSEIGNEMENT SUPÉRIEUR ET LA RECHERCHE

Lors de la Conférence de Paris organisée par Etalab en avril dernier, Geneviève Fioraso, secrétaire d'État chargée de l'Enseignement supérieur, annonçait l'ouverture d'une plateforme de données ouvertes sur l'enseignement supérieur et la recherche. Modalités et enjeux.

À l'ouverture, cette plateforme, synchronisée avec data.gouv.fr, propose 23 jeux de données : points d'intérêts sur les établissements d'enseignement supérieur, de recherche ou les institutions de transfert de technologie, données issues d'enquêtes statistiques ou de processus de gestion. Cette démarche, novatrice pour l'administration centrale de l'État, signe un engagement fort et structuré en faveur de l'ouverture des données. Pour le ministère de l'Éducation nationale, de l'Enseignement supérieur et de la recherche (MENESR), l'objectif est double : alimenter le débat public et contribuer à l'émergence de nouveaux services ou activités économiques, d'une part ; moderniser son administration en utilisant l'*open data* comme levier, d'autre part. Deux priorités émergentes dans la structuration de son offre de données ouvertes : proposer des jeux de qualité et offrir aux utilisateurs un ensemble de services d'accompagnement.

Des données ouvertes à tous et enrichies

Les efforts sont concentrés sur des jeux de données stratégiques proposés à des niveaux de granularité aussi fins que le permet le cadre légal, historisés et pérennes. Les données sont enrichies d'informations facilitant leur réutilisation et leur association avec d'autres sources, par exemple au moyen de *mashup* de données. Identifiants issus du code officiel géographique, Siret, identifiants Wikidata ou Freebase, sites web et comptes réseaux sociaux des institutions sont associés aux données mises à disposition.
Mais encore faut-il que ces données soient accessibles et que les utilisateurs potentiels puissent rapidement en estimer la qualité et l'intérêt et disposer de points de repère pour en faire un usage pertinent. Le MENESR accompagne ses utilisateurs en proposant un dispositif performant d'exposition des jeux de données. La solution retenue, en mode SaaS, présente des fonctionnalités de prévisualisation et d'exploration, des formats d'exports ouverts et une API. Il met également à disposition des documentations techniques et associe aux jeux de données ses propres réutilisations (notamment grâce aux fonctionnalités sociales de data.gouv.fr) : notes et rapports, *data visualisation*. Il s'agit ainsi de mettre au jour la valeur des jeux de données et de proposer aux utilisateurs des moyens d'étalonner leurs travaux.

Un dispositif numérique diversifié

Dans la stratégie numérique du ministère, les données ouvertes ne constituent qu'un des canaux de diffusion de l'information. À la mise à disposition de données de détail, dans un format permettant des réutilisations industrielles, répondent en effet deux autres modes de diffusion complémentaires : l'offre éditoriale statistique et les outils grand public de *data visualisation*. L'offre éditoriale statistique a entamé sa mue vers le numérique. Le MENESR a ainsi publié une version « numérique d'abord » de sa publication annuelle de synthèse *État de l'enseignement supérieur et de la recherche en France*[1] donnant accès, sous licence libre, à l'ensemble des chiffres et indicateurs mobilisés dans la publication et établissant des ponts entre chapitres de l'ouvrage et jeux de données ouverts disponibles sur la plateforme. Progressivement, l'ensemble de l'offre éditoriale de la statistique publique sur l'enseignement supérieur et la recherche rejoindra ce dispositif. Enfin, le MENESR propose des outils de *data visualisation* grand public, intégrés à son site institutionnel et construits sur ses jeux de données ouverts afin de faciliter l'accès de tous à une information personnalisée (géographie de l'offre de formation et de la population étudiante, statistiques d'insertion des diplômés de l'Université, participations nationales au 7e Programme-cadre de recherche et développement technologique, offre de formation et population étudiante des principaux établissements d'enseignement supérieur).

Avec ces travaux, l'accès de tous aux données sur l'enseignement supérieur et la recherche est nettement amélioré. Mais le chantier est encore vaste : banaliser la démarche d'ouverture des données, encourager des initiatives de réutilisation et faire émerger une communauté de réutilisateurs, étendre leur périmètre en garantissant la qualité des informations mises à disposition.

Emmanuel Weisenburger
Département des outils d'aide à la décision (MENESR-DGESIP/DGRI)
emmanuel.weisenburger@recherche.gouv.fr

POUR EN SAVOIR PLUS

Site de la plateforme de données ouvertes :
http://data.enseignementsup-recherche.gouv.fr

[1] http://publication.enseignementsup-recherche.gouv.fr/eesr/7

La bibliothèque de Clermont Université joue un rôle actif dans le soutien aux circuits de la publication académique en accompagnant les enseignants chercheurs dans leur démarche. Périmètre et enjeux.

Des publiothèques *universitaires ?*

***Publiothèque :* ce néologisme a pour seul mérite de poser deux bornes entre lesquelles va courir mon propos.** D'un côté la proximité entre bibliothèque et chose publiée, de l'autre une nouveauté : l'activité de publication des bibliothèques. De fait, si *publication* désigne *ce qui a été publié*, bibliothèque et publiothèque sont purs synonymes. Mais si l'on désigne la publication comme activité, alors *publiothèque* qualifie bien la bizarrerie de cette chose nouvelle, voire choquante : au nom de quoi les bibliothèques se mêlent-elles de publier ? Or elles s'en mêlent, si nombreuses qu'elles ne s'en cachent même plus. L'édition 2014 du *Library Publishing Directory*[1] américain recense 115 bibliothèques adonnées à ce vice, dont une européenne : la bibliothèque de l'université Humboldt à Berlin est contaminée. En France, combien de bibliothèques sont porteuses du virus sans le savoir ou, pire, mènent sciemment des expérimentations plus ou moins avouées dans ce domaine ?

Quand une nouveauté prend cette ampleur, ce n'est plus une bizarrerie ni une lubie locale, c'est un mouvement de fond qu'il faut comprendre. Les bibliothécaires ayant pour principe de travail d'être à l'écoute des besoins de leurs utilisateurs, s'ils font de la publication, c'est qu'on le leur a instamment demandé. Mais alors pourquoi des chercheurs sollicitent-ils les bibliothèques à ce sujet, pourquoi acceptent-elles, et que peuvent-elles proposer concrètement ?

Leurs préoccupations ont placé les bibliothécaires aux avant-postes du mouvement de l'*open access* qui submerge l'édition scientifique. C'est la véritable raison pour les bibliothèques d'acquiescer à la demande, voire de la devancer.

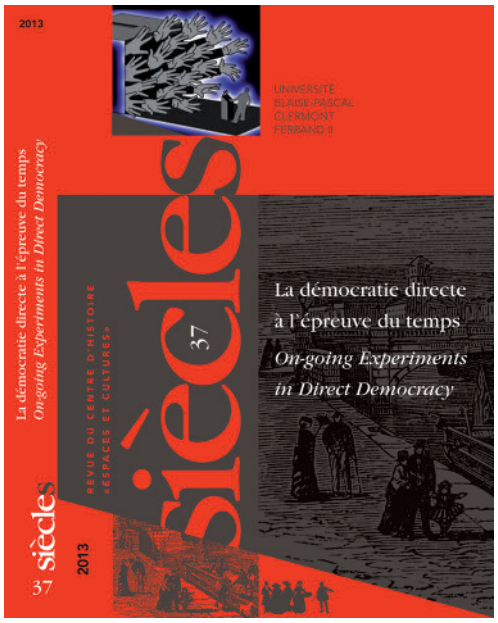
POURQUOI LES BIBLIOTHÈQUES ?

L'introduction du *Library Publishing Directory* situe la réponse du côté de la technicité (p. viii, je traduis) : « *Depuis plus de deux décennies, les universitaires, chercheurs et étudiants, sont venus solliciter, dans leurs bibliothèques de faculté et d'université, un soutien technique et du personnel pour lancer des expérimentations appliquant le numérique au domaine académique* ». Interrogée sur ce qu'apporte la bibliothèque, une collègue enseignant chercheur à Clermont complète en ces termes : « *une aide concrète, logistique ; un interlocuteur, des outils et du dépannage* » ; mais aussi « *un soutien psychologique, de la stimulation et de l'enthousiasme* ». Les termes suggèrent – ce n'est pas déplaisant – que les bibliothèques sont perçues comme familières de l'outil numérique et capables de former à son maniement. De fait, si nous sommes loin d'être aussi à l'aise que nous le voudrions avec le numérique, nous en avons bel et bien une pratique assidue : catalogues en ligne, outils d'accès aux ressources, chaînes de traitement des thèses, archives ouvertes, web social, expérimentation des nouveaux supports… Tout nous pousse vers l'outil numérique de diffusion ; qu'une relative aisance en découle, et que nous soyons pressentis pour la diffuser est plutôt une bonne nouvelle.

Autre atout des bibliothèques : la familiarité avec les enjeux économiques et juridiques. Ferrailler, 15 années durant, avec les éditeurs, a poussé les bibliothécaires, négociateurs et gestionnaires de documentation électronique, à s'intéresser aux circuits de validation, aux modèles économiques, à la chaîne de production. Bon gré mal gré, ils sont aussi montés au front sur les questions de droit, leurs usages traditionnels (prêt entre bibliothèques, conservation pérenne, libre accès) étant remis en cause par l'irruption de contrats d'accès non pérenne. On trouve aujourd'hui en bibliothèque de solides connaisseurs du droit de l'information et du droit d'auteur, sinon des juristes ; de bons connaisseurs des enjeux de l'édition, sinon des éditeurs.

Enfin, leurs préoccupations ont placé les bibliothécaires aux avant-postes du mouvement de l'*open access* qui submerge maintenant l'édition scientifique. Et c'est la véritable raison pour les bibliothèques d'acquiescer à la demande, voire de la devancer. *Green ou Gold*, abonnement, auteur-payeur ou freemium, quelle que soit la réponse les bibliothécaires sont convaincus que l'*open access* doit s'imposer et déterminés à y contribuer. Les

[1] www.librarypublishing.org/resources/directory-library-publishing-services



À Clermont-Ferrand, la collaboration entre le Centre d'histoire « Espaces et cultures » de l'université et la bibliothèque Clermont Université a facilité la migration de la revue *Siècles* sur revues.org.

bibliothèques assurent le dépôt et la diffusion des thèses électroniques ; elles développent et gèrent des dépôts d'archives ouvertes, avec pour effet prévisible d'amincir la frontière entre documentation produite et documentation acquise par les bibliothèques… La publiothèque est déjà là.

LES BIBLIOTHÈQUES, POUR QUOI ?

Par la technicité, l'expérience, la connaissance des enjeux, la conviction, les bibliothèques sont donc un interlocuteur naturel des projets de publication scientifique en *open access*. Que peuvent-elles proposer de concret aux chercheurs qui les sollicitent ? D'abord, l'information à jour sur les dispositifs existants. À titre d'exemple, à Clermont-Ferrand (ailleurs aussi sans nul doute), la bibliothèque aide les chercheurs à instruire les dossiers de revues de sciences humaines pour une diffusion sur la plateforme revues.org[2]. En sciences exactes, elle suit attentivement les débuts des épi-revues et encourage les revues scientifiques naissantes à adopter ce mode de diffusion. Sur le terrain mouvant de la publication scientifique, la bibliothèque dispose d'une carte à jouer et peut servir de guide. Elle peut aussi offrir un soutien technique, par exemple *via* une plateforme comme OJS[3] (*Open Journal Systems*, actuellement testé à Clermont), un logiciel libre qui coordonne le traitement des articles (soumission, relecture, validation, publication), dans une interface propre à chaque revue. La bibliothèque installe, administre, maintient le logiciel et aide les chercheurs pour la prise en main et l'utilisation. On peut aussi penser au balisage XML des articles, nécessaire pour la diffusion sur revues.org ; les bibliothèques peuvent effectuer tout ou partie du travail (une collaboration entre bibliothèque et laboratoire a ainsi permis la migration de la revue *Siècles*), ou dispenser la formation nécessaire.

Dans tous ces cas : conseil, instruction de dossier, formation initiale, administration d'un outil, la bibliothèque agit en soutien de projets naissants. Doit-elle s'aventurer au-delà ? Avec prudence. Le traitement éditorial (relecture, mise en page, etc.) est un métier. S'il devait être effectué en bibliothèque, ce serait par intégration ou reconversion de personnel, ce qui impliquerait une stratégie universitaire mûrie et volontariste. Impliquer une bibliothèque dans un secrétariat de rédaction prête aussi à discussion. L'idée peut séduire parce qu'elle crée un lien très fort entre la bibliothèque et la recherche, mais cette activité exige une disponibilité considérable et une compétence scientifique élevée. Il est peu réaliste d'imaginer la systématiser. De même, si l'aide au lancement de revues entre assez naturellement dans le champ d'action des bibliothèques, le livre électronique exige une autre compétence éditoriale.

LE PAYSAGE ET LA BOUSSOLE

Arstuce / Flickr (CC BY-NC-SA 2.0)

Enfin, il serait malhonnête d'éluder l'inquiétude bien naturelle que suscite l'arrivée des bibliothèques dans le champ de la publication scientifique, par crainte d'une concurrence avec des acteurs déjà présents (laboratoires, presses universitaires etc.). Il me semble que quelques constats rassurants doivent présider aux discussions : il y a du travail pour tout le monde (et au-delà) ; les moyens sont aussi modestes dans les bibliothèques qu'ailleurs – pas seulement en France : les bibliothèques américaines de la *Library Publishing Coalition* comptent en moyenne (calculée sur 55 établissements)… 2,25 équivalents temps plein dévolus à la publication ! Vu l'ampleur de l'enjeu, laisser germer des conflits de compétences serait mortifère. La mise en place d'un système où chacun contribue différemment n'est pas seulement vitale, elle est possible ! On peut imaginer 100 stratégies allant de la répartition des tâches à l'intégration des services, catégoriser par discipline, distinguer livres, thèses et revues… Bref tout est concevable à condition de garder l'objectif en ligne de mire : la mise à disposition des travaux scientifiques la moins coûteuse, la plus rigoureuse, la plus efficace et la plus large possible. Le paysage est mouvant, mais la boussole, c'est l'*open access*.

Olivier Legendre
Responsable de la Bibliothèque numérique, Clermont Université
olivier.legendre@clermont-universite.fr

[2] www.revues.org
[3] https://pkp.sfu.ca/ojs/
[4] http://siecles.revues.org/1176

La veille informationnelle
Enjeux et évolutions

La veille professionnelle est aujourd'hui confrontée à de nouveaux défis étroitement liés au développement des technologies de l'information. Décryptage par Frédéric Martinet, consultant indépendant en intelligence économique.

La veille, démarche qui vise à rationaliser, organiser et systématiser la surveillance de son environnement, et plus particulièrement des mouvements qui pourraient avoir un impact sur son activité, est inévitablement confrontée à la numérisation de l'information et à l'explosion de ses volumes et donc aux aspects techniques qu'il est désormais impossible d'éluder lorsqu'il s'agit d'accéder à l'information et la traiter.

LA VEILLE À L'HEURE DE LA NUMÉRISATION

La numérisation n'en est certes plus à ses débuts, mais c'est un mouvement massif qui vise désormais à supprimer l'information sur support physique ou tout du moins sur support papier. Cette évolution du support de l'information a transformé les métiers du savoir et déplacé l'intervention humaine dans la chaîne de traitement de la valeur. Là où l'humain lisait, sélectionnait, résumait – travail que nombre de documentalistes ont réalisé fort longtemps – il ne subsiste maintenant plus que des algorithmes qu'il utilise et paramètre en y injectant des mots-clés, des périodes, en peaufinant des équations booléennes et en couplant son système avec des sources d'information.

La veille requiert désormais les capacités à :

- détecter de nouvelles sources d'informations ;
- se connecter à ces sources afin d'en extraire l'information de la façon la plus efficace possible ;
- savoir utiliser des outils de traitement de ces flux d'information et maintenir leur efficacité ;
- analyser cette information ;
- animer une communauté de personnes mobilisées autour d'un processus de veille.

LES NOUVEAUX DÉFIS DES ENCLOSURES[1]

Il y eut d'abord le Web 1, un idéal ouvert de documents liés les uns aux autres par une balise unique : la balise <A>. Puis vint le Web 2, un web indéniablement social, favorisant l'explosion des contenus et plus particulièrement des contenus générés par les utilisateurs (*User-generated content*), avec l'apparition d'un nouvel Eldorado : celui de la donnée des utilisateurs, de la donnée personnelle et de sa monétisation.

De ce web normé, unifié, social naquirent des entités dont le slogan désormais lointain était « *don't be evil* » (« ne soyez pas malhonnêtes »). Les univers, ouverts dans un premier temps, se referment pour céder la place à des espaces clos, privatifs ou « communautaires » qui protègent leurs données à coup de conditions générales d'utilisation (CGU), de procès ou d'API de plus en plus restrictives, et le veilleur se trouve ainsi confronté à des murs numériques.

Les options sont alors multiples et leur impact éthique ou juridique souvent loin d'être neutre :

- pénétrer ces espaces en allant à l'encontre des CGU ;
- utiliser des APIs souvent bien restrictives et perdre une part des données ou des éléments de contextualisation pouvant ainsi amener à déformer l'information ;
- acheter la donnée lorsque cela lui est possible.

Il s'agit là du défi le plus important que doit relever le veilleur car ces espaces donnent accès à de l'information blanche, grand public, massive, mais aussi à des informations parfois grises, des indiscrétions, des inattentions à valeur souvent bien plus élevée.

COMPOSER AVEC LA MASSE D'INFORMATION

À lui seul Google possèderait à ce jour plus de deux millions et demi de serveurs informatiques, faisant de lui le premier possesseur privé de serveurs en termes de nombre, de capacité de stockage et probablement en capacité de calcul.

Même si le veilleur limite son attention à une partie du web (celle relative à son activité), il n'en demeure pas moins que la surveillance de millions de pages web nécessite du matériel et de la bande passante. De surcroît, même s'il limite son périmètre d'écoute, il aura toujours besoin d'avoir un radar d'horizon plus large sur le web, afin de détecter de nouvelles sources ou tout simplement des informations qui seraient diffusées sur des sources non surveillées.

La masse d'information disponible et accessible pose donc plusieurs problèmes :

- celui de l'aiguille dans la botte de foin et plus particulièrement lorsque l'on recherche un signal faible ou divergent dont les mots clés descripteurs ou le vecteur sont inconnus ;

[1] Le terme « enclosure » est utilisé pour décrire des processus visant à limiter l'accès ou la réutilisation d'une information.

• celui du bruit informationnel, car quelle que soit l'efficacité des équations de recherche ou de filtre, quelle que soit la qualité des algorithmes de détourage de l'information, de désambiguïsation ou tout simplement d'anti-spam, le veilleur se heurte bien souvent au bruit ;
• celui de la mobilisation des ressources humaines, car les effectifs de veilleurs et de documentalistes sont aujourd'hui rarement à la hausse et je reste convaincu qu'à ce jour la qualité des algorithmes et des technologies accessibles au commun des entreprises ne suffit pas à pallier les limitations imposées en termes de personnels chargés de la collecte, de la diffusion et de l'analyse de l'information

LES CONTRAINTES JURIDIQUES

La veille est soumise à de fortes contraintes réglementaires, juridiques et contractuelles. Et cela ne semble pas évoluer dans un sens qui facilite les activités de veille malgré des mouvements comme les *creative commons* et l'*open data* qui en sont encore à leurs premiers frémissements.
Bien que dans les faits, de nombreuses entreprises sous-estiment, ignorent ou tout simplement décident de ne pas respecter certaines restrictions juridiques, toute information qu'elle soit papier ou numérique (site web, blog, forum…) est potentiellement soumise au droit d'auteur à de rares exceptions près. Par ailleurs, les éditeurs de sites web ajoutent souvent des conditions générales d'utilisation qui interdisent l'extraction automatisée des informations qui y sont diffusées.
La législation est également différente en fonction du type d'information, des structures, mais aussi du pays du détenteur des droits de propriété intellectuelle, de l'hébergeur et du veilleur, complexifiant souvent à outrance les veilles dont la portée est internationale et l'appréciation du risque juridique y afférent. Enfin, la jurisprudence est assez mouvante sur ces sujets et le veilleur devra avoir à cœur de se tenir informé des derniers jugements.

DES DIFFICULTÉS SPÉCIFIQUES AU MULTIMÉDIA

Rechercher des textes à partir de mots-clés devient presque une banalité à tel point que l'on entend souvent l'hérésie « *tous veilleurs* » ou bien alors « *sur Google, on trouve tout* ». Par contre, la recherche d'images et la recherche vidéo comportent encore de nombreuses carences qui limitent leur accessibilité au veilleur. Ainsi la recherche d'images à partir de mots-clés nécessite une indexation manuelle car hormis quelques formes pouvant être reconnues automatiquement ou quelques textes océrisés, l'indexation automatique et la catégorisation des images restent encore limitées. On voit bien apparaître des technologies de recherche d'images par similarité à une autre image, mais celles-ci sont très gourmandes en ressources informatiques et en temps de calcul limitant souvent leur exploitation à quelques sites (Flickr en tête) ou à quelques applications (monuments, étiquettes de vin, produits et marques, véhicules et plus particulièrement véhicules militaires…).
De la même façon, pour les flux audios ou vidéos, les logiciels de reconnaissance vocale (*Speech to Text*) sont également gourmands en ressources et leur exploitation reste coûteuse, limitant leur accès aux organismes gouvernementaux, aux grands groupes ou aux prestataires de services à même de mutualiser ces technologies pour leurs clients.

Pierre (Rennes) / Flickr (CC BY 2.0)

Le veilleur aura toujours besoin d'un radar d'horizon sur le web afin de détecter, dans le flux massif des informations, de nouvelles sources utiles à son activité.

En conclusion, les défis de la veille sont étroitement liés à l'évolution de l'information et ses nouveaux enjeux. Là où certaines sociétés privées cherchent à s'accaparer l'information, le veilleur devra trouver moyen d'y accéder en respectant les contraintes réglementaires. Alors que les ressources humaines dédiées à la veille s'amenuisent, ce dernier devra sélectionner les technologies adéquates et mobiliser les bonnes volontés internes, les connaissances et s'affirmer comme un animateur-coordinateur. Enfin, le veilleur devra garder son esprit ouvert sur un web en mouvement perpétuel. Il évaluera les nouvelles opportunités d'accès à l'information et devra savoir composer avec les évolutions technologiques.

Frédéric Martinet
Fondateur d'Actulligence Consulting
frederic.martinet@actulligence.com
www.actulligence.com

La bibliothèque de Paris 8, labellisée Marianne

Les services au public de la bibliothèque de Paris 8 ont obtenu le label Marianne en mars 2014. Retour sur la démarche suivie, ses enjeux et ses prolongements.

Depuis son installation en 1998 dans le bâtiment de Pierre Riboulet, la bibliothèque de Paris 8 accueille les étudiants et les enseignants-chercheurs de l'université, mais aussi toute personne de plus de 18 ans. Cette politique généreuse s'est étendue au prêt en octobre 2010 : les étudiants d'Île-de-France et les habitants de la Seine-Saint-Denis peuvent emprunter gratuitement deux livres pour deux semaines. Les 1 500 places assises disséminées sur les 11 000 m² d'espace public permettent cette ouverture à tous. L'attachement au service public, entendu comme l'ensemble des activités impliquant un contact direct ou distant avec le public, est profondément ancré dans la mentalité du personnel de la bibliothèque. Tous les agents y consacrent entre 30 et 50 % de leur temps de travail, hormis quelques agents administratifs.

GENÈSE DU PROJET

Lorsque le ministère de l'Enseignement supérieur et de la Recherche a diffusé en février 2011 une circulaire incitant les bibliothèques universitaires à se lancer dans une démarche de labellisation Marianne, l'examen du référentiel nous a très vite convaincus que nous n'avions pas besoin de réformer notre fonctionnement en profondeur pour remplir les 19 engagements Marianne. En revanche, nous tenions là un moyen de nous aider à formaliser et harmoniser nos pratiques quotidiennes d'accueil. Nous pouvions, d'autre part, utiliser ce vecteur pour valoriser un savoir-faire qui peut être considéré comme allant de soi dans une bibliothèque, alors qu'il est peu répandu dans les établissements de l'enseignement supérieur.

En janvier 2012, un chef de projet a été nommé et un groupe de travail constitué avec des représentants des services directement concernés par le respect des engagements du référentiel. Au printemps 2012, une première auto-évaluation a permis d'identifier les actions à conduire en priorité et de bâtir un plan d'action.

L'engagement 15 stipule : « *nous exprimons notre politique d'accueil et nous impliquons notre personnel* ». Cette dimension du projet s'est vite révélée cruciale pour le succès de l'entreprise. De nombreuses opérations de sensibilisation et de communication auprès de l'ensemble du personnel ont jalonné tout le processus. Le port du badge ainsi que des mesures correctives portant en particulier sur l'accueil téléphonique et les réponses/envois/signatures de courriels ont été introduits à la rentrée 2012. Ces procédures n'avaient jamais été mises à plat et ce travail a pointé l'hétérogénéité des pratiques qui ont donc dû être harmonisées et formalisées.

Le projet a marqué une petite pause après le départ en mutation du chef de projet. Le temps que la nouvelle responsable des services de l'accueil et du prêt se familiarise avec le contexte et la démarche qualité. Il s'agit bien en effet d'une opération de cette nature et pas d'un simple engagement à décrocher le téléphone au bout de cinq sonneries comme on l'entend parfois à propos du référentiel Marianne. Des engagements comme : « *nous surveillons régulièrement le respect de nos engagements* » (17), « *nous évaluons régulièrement nos pratiques avec une volonté d'amélioration continue* » (18) induisent un travail en profondeur sur les méthodes d'évaluation de la qualité de l'accueil et amènent à élaborer des tableaux de bord, à programmer régulièrement des auto-évaluations et des enquêtes de satisfaction auprès du public, à dresser chaque année un bilan et à mettre en œuvre des actions d'amélioration. L'engagement 16 : « *nous organisons notre documentation pour apporter des informations exactes et à jour aux usagers* » nous a contraints à revoir la gestion des documents internes que nous produisons à l'intention des agents de la bibliothèque et entreposons dans un intranet. Du côté du personnel, nous avons organisé des formations sur site à l'accueil qui ont touché 30 agents à l'été 2013 et 15 autres à l'été 2014. Nous avons lancé une enquête sur la communication interne qui a permis de cerner les faiblesses de nos dispositifs : pour la qualité de l'accueil, il est indispensable que l'information circule bien entre les services et entre les membres de l'équipe. Du côté des usagers, nous avons aussi développé les services proposés aux personnes handicapées ou en difficulté (engagements 4 et 5).

LA PHASE DE LABELLISATION

La construction des indicateurs et leur suivi sur plusieurs mois (avril à octobre 2013) a été couronnée par une deuxième auto-évaluation en novembre 2013 dont les résultats meilleurs que ceux de 2012 nous ont encouragés à entrer dans la phase de labellisation. Pourquoi aller jusque-là ? Il nous a semblé qu'aller au bout de la démarche, c'est-à-dire solliciter un avis extérieur, celui d'un cabinet d'audit agréé par le Secrétariat général pour la modernisation de l'action publique (SGMAP), donnerait une valeur bien supérieure à la reconnaissance ainsi établie de la qualité des services rendus au public à la bibliothèque de Paris 8. Nous disposerions, en outre, d'un outil de communication beaucoup plus puissant vis-à-vis du public, des acteurs du territoire et de la communauté universitaire.

Le cabinet LNE a été choisi pour auditer nos services d'accueil après mise en concurrence. Passées les étapes de recevabilité administrative et technique de notre dossier, LNE a lancé des enquêtes-mystères à

Ce label est accordé pour 3 ans à condition de répondre de façon satisfaisante à l'audit de suivi au bout de 18 mois. Il ne faut donc pas s'endormir sur ses lauriers.

Thierry Kiefer/SCD Paris 8

Hall d'accueil de la bibliothèque de Paris 8.

Des salles de lecture de qualité.

distance sur les courriels, courriers et appels téléphoniques, consistant à tester nos délais et formes de réponse en se faisant passer pour des usagers. Un audit sur site s'est déroulé le 17 décembre 2013 toute la journée à l'issue duquel un avis favorable a été émis même si quelques non-conformités jugées non bloquantes ont été relevées. Des propositions d'actions correctrices pour ces dernières ont été élaborées en janvier-février 2014. Elles ont été avalisées en comité de lecture du cabinet d'audit et les services au public de la bibliothèque de Paris 8 ont été labellisés Marianne en date du 14 mars 2014. Ce label est accordé pour trois ans à condition de répondre de façon satisfaisante à l'audit de suivi au bout de 18 mois (septembre 2015). Il ne faut donc pas s'endormir sur ses lauriers, mais profiter de la dynamique créée par cette labellisation.

ET AU-DELÀ : L'ÉVOLUTION DES USAGES…

Ce travail sur la qualité des services au public est indispensable, mais ne suffit pas. Les services eux-mêmes doivent aussi évoluer. Le SCD de Paris 8 mène depuis deux ans maintenant deux autres projets qui y concourent. Le projet de learning centre prend en compte l'évolution des usages constatée au sein de la bibliothèque et dans les pratiques pédagogiques. Il est inscrit dans le Schéma directeur du numérique de l'université, ainsi que dans le Schéma directeur du patrimoine. Son nom, « La Pépinière », affirme la volonté de construire et de cultiver de nouvelles formes de connaissance qui impliquent notamment une intégration plus forte de la documentation et de l'enseignement. L'essor des usages du numérique et les possibilités d'expérimentation qu'ouvre ce dernier requièrent un accompagnement et des lieux particuliers. Le passage à la technologie RFID et l'implantation d'automates de prêts-retours à la rentrée 2014 visent à redéployer des agents sur un service public repensé, première étape du projet de learning centre.

… ET DES MÉTIERS

Cette réflexion sur le service public s'inscrit dans une problématique plus large qui constitue le second projet par lequel le SCD cherche à faire évoluer les services proposés par la bibliothèque. Un groupe de travail s'est consacré pendant deux périodes (mars-juillet 2013, mars-juillet 2014) à l'évolution des métiers au sein du SCD de Paris 8. Il a pour objectif de cerner l'impact des attentes du public et celui des services préfigurés dans le cadre de « La Pépinière » ou par ailleurs (développement des services aux chercheurs) sur les activités des personnels de la bibliothèque. Il est animé par l'adjointe à la directrice du SCD et réunit 15 agents très actifs, dont cinq magasiniers. Le produit de leur travail fournit des éléments d'aide à la décision pour l'évolution des profils de poste. Le groupe traite cet été la question : quelles formations pour donner aux agents, en particulier aux magasiniers, les compétences nécessaires à la diversification des activités ?

Ces travaux menés sur la qualité de l'accueil (Marianne), sur les évolutions du service public et plus globalement des métiers en bibliothèque, doivent maintenant être confrontés aux attentes des usagers. L'observation des pratiques ne suffit pas. Le SCD a donc programmé fin 2014 une grande enquête auprès du public (et du non-public) dans la lignée de celle conduite en 2006-2007. Ses résultats seront pris en compte pour la réorganisation du service public qui connaîtra d'autres étapes et pour l'élaboration des phases suivantes du projet « La Pépinière ».

Carole Letrouit
Directrice du SCD de Paris 8
carole.letrouit@univ-paris8.fr
Floriane Berti
Chef de projet Marianne pour le SCD
floriane.berti@univ-paris8.fr

Sudoc II. Année 0

En mai 2014, le cabinet Pleiade a remis son étude sur l'évolution du catalogage et de la production de métadonnées dans le réseau Sudoc, sur fond de projet Système de gestion de bibliothèque (SGB) mutualisé. Ce document clarifie les données du problème et apporte de précieuses informations de terrain, mais il ne propose évidemment pas de scénario clé en main. L'étude doit reprendre, en parallèle du lancement de l'appel d'offres SGB mutualisé, puisqu'il est désormais acquis que le prochain Sudoc – Sudoc II – ne sera pas fatalement une des plateformes retenues par ce marché. Cet article fait le point sur la réflexion en cours.

Sudoc II, ça sonne un peu comme Penduick II. Ou Titanic II. Mais cette connotation est trompeuse car le Sudoc est en pleine mer, vitesse de croisière, sans perspective de quelques années de cale sèche pour travaux de rénovation – voire de reconstruction.
S'il faut tricoter encore un peu l'analogie, c'est plutôt le bateau de Thésée qui ferait l'affaire : entretenu et réparé petit à petit, planche par planche, à l'arrivée, il ne possède plus aucune pièce d'origine (est-ce encore le même ? demandent Plutarque puis la tradition philosophique jusqu'à aujourd'hui).
En effet, le Sudoc a déjà changé depuis son lancement en 2000, grâce aux évolutions logicielles apportées par le prestataire OCLC, mais tout autant grâce aux initiatives de l'Abes : chantier Unicode, scripts, stratégie API Sudoc, etc. À partir de 2008, cette dernière stratégie a cherché à rendre le Sudoc de plus en plus indépendant des progiciels OCLC pour exploiter les données de sa base centrale et pour les modifier (création/modification/suppression de notices bibliographiques ou d'autorité et d'exemplaires). La stratégie API Sudoc a permis aux services suivants d'émerger : IdRef, Self, Périscope, Colodus, micro web services, exposition sur le web de données, etc.

Myrabella / Wikipedia (CC BY-SA 3.0)

Tout comme pour la construction de la réplique de l'Hermione (ci-dessus) de 1779, Sudoc 2 est un vaste chantier pour lequel il ne suffira pas de changer juste quelques planches.

Changer la quille

Tous les services rendus possibles par la stratégie API Sudoc supposent l'existence de la base de données centrale du Sudoc, qui tourne sur un progiciel *ad hoc* maintenu par OCLC. Ce logiciel est performant mais très ancien. D'ici cinq ans, il ne sera plus maintenu. C'est donc une pièce maîtresse du dispositif Sudoc qu'il s'agit de remplacer. À vrai dire, le bateau de Thésée est une analogie qui pèche : Sudoc II ne sera pas seulement Sudoc I avec des pièces de rechange. En raison des évolutions technologiques et normatives, des changements apparus dans son environnement comme, par exemple, l'émergence des systèmes de gestion de bibliothèque dits « de nouvelles génération » (#SGB mutualisé) et, enfin, en raison de l'évaluation critique du dispositif actuel, le nouveau Sudoc sera différent de l'ancien : périmètre, fonctionnement, modèle économique.

Fondamentaux et inconnues

Mais avant de lister les nouveaux principes de conception aujourd'hui actés et les questions, souvent cruciales, encore en suspens, il est utile de rappeler et de réaffirmer les principes fondamentaux qui font l'ambition du Sudoc, l'ancien et le prochain.
Ce qui a été clairement réaffirmé par les membres du réseau participant aux deux groupes de travail, c'est que le Sudoc est plus qu'un catalogue collectif : c'est un réseau de catalogage partagé. Ce qui est partagé, ce sont des règles de catalogage, un dispositif de formation et d'assistance, mais également un outil de catalogage. Le développement des SGB, qui intègrent des fonctions de catalogage au sein du réseau des clients de chaque plateforme, a pu, un moment, laisser penser que le catalogage partagé Sudoc était caduc. Finalement, les membres du réseau ont souhaité réaffirmer la néces-

sité d'un catalogage Sudoc centralisé, hors plateforme commerciale, au moyen d'un outil maîtrisé par l'Abes et sa communauté… à condition que cet outil unique soit intégré de manière transparente dans les systèmes locaux (SGB ou pas). Le cahier des charges est clair. La solution n'est pas encore là : ni la technique ni l'ergonomie de cette intégration fluide. D'ailleurs, si cet outil doit être maîtrisé par l'Abes et son réseau, tous les scénarios restent ouverts : sera-t-il géré directement par l'Abes ou bien sous-traité à une plateforme commerciale ?

Une autre question importante est encore en suspens : celle du modèle de données. Est-il préférable de conserver Marc comme format de travail et de gestion ou est-ce le bon moment pour passer à autre chose, à un modèle de type bases de données (RDF ou pas) ? Ces questions devront être tranchées au printemps 2015, ce qui suppose de poursuivre l'étude engagée par le cabinet Pleiade, notamment en allant voir sur place (Libris en Suède, CIB en Allemagne, etc.).

Dans le marbre

À côté de ces incertitudes de fond, d'autres principes sont d'ores et déjà gravés.

• Sudoc II s'intéressera à tous les documents, quels que soient les supports.

• Sudoc II s'intéressera à tous les documents, quels que soient les modes d'acquisition/accès : *open access*, achat, location (abonnements, droits d'accès). Le signalement de la documentation électronique doit faire partie intégrante de la conception de Sudoc II, mais l'erreur serait de plaquer sur elle les principes du signalement de la documentation imprimée. Précisément, la base de connaissance Bacon aura sa propre autonomie, à côté des autres modules de Sudoc II.

• Sudoc II gère des métadonnées quelle que soit leur origine : catalogage ; dérivation d'autres bases bibliographiques ; imports éditeurs [le hub de métadonnées sera un module de Sudoc II] ; imports catalogues ; génération automatique de métadonnées à partir d'autres métadonnées ou de la fouille de texte.

• Il faut que Sudoc II connaisse à la fois tous les exemplaires d'un document (Sudoc I ne connaît souvent qu'une localisation par bibliothèque) et la disponibilité du document en temps réel (pas seulement par un *deep link* vers le système local, mais *via* une remontée de l'information dans Sudoc II – soit dans la base de données, soit seulement dans l'interface).

Sudoc II émergera progressivement dans la lancée des initiatives apparues ces dernières années, en impliquant les demandes et compétences du réseau.

• Les données d'exemplaires peuvent être gérées soit dans le système local (saisie en local puis remontée vers Sudoc II par web service), soit directement dans Sudoc II. Finie l'inévitable double gestion des exemplaires, bancale et coûteuse.

• Quand un référentiel externe est à la fois ouvert d'un point de vue juridique et technique, alors Sudoc II interrogera ce référentiel à distance, sans en faire une copie locale. Par exemple, la notice d'une thèse sur une obscure abbaye pointera vers l'identifiant de l'autorité BNF si elle existe. Ce n'est qu'en dernier recours (si l'autorité n'est dans aucun référentiel considéré comme fiable) que cette autorité d'abbaye sera créée dans IdRef.

Des sous et des hommes

À côté des questions métier et informatiques, la question du modèle économique reste entièrement ouverte. On devine que le modèle de facturation du Sudoc, qui a lui-même évolué fortement ces dernières années, ne pourra être reconduit tel quel. Au-delà de cette intuition, *terra incognita*.

Ce qui est certain, c'est qu'il n'y aura pas de *deus ex machina* ni d'effet tunnel : Sudoc II émergera progressivement dans la lancée des initiatives apparues ces dernières années, en impliquant autant que possible les demandes et les compétences du réseau. Décidément, Penduick II n'est pas la bonne référence : il fut conçu pour les courses en solitaire.

Groupe de préfiguration
Projet Sudoc II
Benjamin Bober, Christophe Bonnefond, Camille Dumont, Marianne Giloux, Jérôme Kalfon, Isabelle Mauger-Perez, Yann Nicolas, Stéphane Rey

Arabesques à l'Ifla.
Photo publiée sur Twitter par Sophie Agié.

Ar(abes)ques, VOUS EN PENSEZ QUOI ?

Depuis sa création en 1995, le contenu et le format d'*Arabesques* ont changé afin de toujours mieux répondre aux attentes de ses lecteurs.
Alors que la revue s'apprête à souffler ses 20 bougies, il nous a semblé nécessaire de mesurer votre degré de satisfaction, de connaître vos attentes, de recueillir vos suggestions. Vos avis serviront de point de départ à une réflexion de fond sur l'avenir de la revue.

Vous avez déjà été nombreux à nous répondre mais les retardataires auront jusqu'au 20 octobre pour remplir en quelques minutes le questionnaire en ligne sur :
https://fr.surveymonkey.com/s/Arabesques
Bien évidemment, nous publierons dans notre prochaine livraison d'*Arabesques* les résultats de cette enquête pour laquelle nous comptons sur la réponse du plus grand nombre. Merci à vous.

Retour sur les 5e Journées professionnelles du CTLES

La Bulac a accueilli, les 23 et 24 juin derniers, les cinquièmes journées professionnelles du Centre technique du livre de l'enseignement supérieur (CTLES), sur le thème « Quelles collections pour quelles bibliothèques ? Accessibilité, valorisation et conservation ».

S'il ne saurait être question ici de résumer chaque intervention, notons qu'une large part était accordée aux retours d'expériences, forcément divers. Du côté de l'étranger étaient représentés des réseaux très différents : celui de la Bibliothèque nationale du Luxembourg rassemble 65 établissements pour un public de 537 000 habitants, dans un contexte marqué par le pluriculturalisme et le multilinguisme ; le réseau finlandais bénéficie d'une tradition centralisatrice qui permet à la Bibliothèque nationale de dépôt située à Kuopio d'abriter 134 kml de documents et de réaliser 60 % des demandes de communication des bibliothèques universitaires ; quant aux États-Unis, ils évoluent évidemment sur une échelle bien différente et le témoignage du Center for Research Libraries (CRL) de Chicago insiste sur l'éventail très complet de solutions relatives à la conservation partagée, dans un pays peu centralisé où l'initiative vient surtout des réseaux professionnels (OCLC, RLG...).

Des expériences diversifiées

Côté français, deux intervenants – Albert Poirot pour la Bibliothèque nationale et universitaire (BNU) de Strasbourg et Philippe Marcerou pour la Bibliothèque interuniversitaire de la Sorbonne (BIS) – posaient la même question de l'impact d'une restructuration des espaces sur les questions de politique documentaire, de services aux usagers et d'accès, de conservation... Les réponses dans les choix – forcément contraints – d'organisation des collections varient : la BNU a opté pour la pluridisciplinarité, alors que la BIS a choisi une approche thématique. L'intervention de Marie-Madeleine Géroudet (SCD Lille 1) sur les données de la recherche était plus prospective : la problématique est récente (voir les interventions lors du congrès ADBU 2013 ou la création du segment 10 de BSN) et les bibliothécaires doivent s'en saisir. L'exposé, riche et enlevé, a rappelé la légitimité de la profession sur les questions de publications et d'archives ouvertes et la nécessité de procéder de façon pragmatique et progressive.
Brigitte Michel (projet de Learning Centre de Paris-Saclay) a constaté des pratiques très différentes selon les bibliothèques (un degré d'investissement dans le numérique/l'électronique très variable, des logiques de conservation allant de l'hypersélection à l'exhaustivité...) et appelle à ne pas sous-estimer l'investissement symbolique (par les bibliothécaires, mais peut-être et surtout par les usagers !) et la dimension identitaire de ce qui est considéré comme patrimonial.

De nouvelles approches patrimoniales

La problématique patrimoniale a justement été centrale lors de ces journées. Sans aller jusqu'à évoquer un « changement de modèle », Fabien Plazannet (ministère de la Culture et de la Communication) a pointé des évolutions : l'existence d'une double collection nationale, physique et numérique ; la numérisation de masse ; la délocalisation croissante (du stockage, mais aussi des services) ; l'émergence d'un « patrimoine de décision » qui, à la logique d'accumulation, substitue un stockage actif et une gestion dynamique des collections. Dans ce cadre, le CTLES a toute sa place. Une intervention à deux voix – celles de Anthony Roubaud, architecte, et de Bernadette Patte, directrice-adjointe du CTLES – a permis d'apprécier le projet tant architectural que bibliothéconomique, à l'heure où la première tranche des travaux d'extension doit débuter, et alors que le CTLES devient opérateur national pour la conservation et le PEB et que ses missions de stockage distant mais aussi de communication, de conservation partagée et de conseil sont confirmées. Rappelons qu'un processus de conventionnement est en cours entre l'Abes et le CTLES afin de définir les modes de coopération pour les actions visant à soutenir et développer des plans régionaux de conservation partagée des périodiques.

Nombre de questions sont revenues en filigrane durant ces journées : la place (physique et symbolique) des collections imprimées et la saturation des espaces, Collex, les opérateurs nationaux, mais aussi la question de l'interministériel : problématiques et évolutions sont communes, tout comme une partie des publics, alors que la distinction institutionnelle aurait tendance à freiner les initiatives de mutualisation. Terminons en évoquant la synthèse stimulante des journées par Valérie Tesnière et en remerciant le CTLES pour ce riche programme, sans oublier les équipes de la BIS qui ont accueilli les participants pour des visites commentées de leur belle bibliothèque qui a enfin retrouvé ses murs.

Anne-Laurence Mennessier
Abes
mennessier@abes.fr

DROIT D'AUTEUR : POUR UNE RÉSOLUTION INTERNATIONALE

Le 17 août 2014, l'Interassociation Bibliothèques Archives Documentation (IABD) a soutenu la position de l'Ifla en vue d'une adaptation à l'échelle internationale du droit d'auteur et du copyright par une déclaration reprise intégralement ci-dessous.

Phot. Charlotte Hénard (CC BY-SA 2.0)

Quelques signataires de la déclaration de l'IABD lors de l'Ifla. De gauche à droite : Donna Scheeder (Ifla), Sinikka Sikilä (Ifla), Xavier Galaup (IABD), Anne Verneuil (ABF), Christophe Pérales (ADBU), Pascal Sanz (CFIbd).

Les bibliothèques et services d'archives, parmi lesquels la Fédération internationale des associations et institutions de bibliothèques (Ifla), conduisent depuis 2009 une campagne auprès de l'Organisation mondiale de la propriété intellectuelle (OMPI), pour un instrument international officiel qui reconnaîtrait l'importance des exceptions et limitations au droit d'auteur pour les services que rendent les bibliothèques et les services d'archives. L'IABD soutient cette campagne en trois points.

Les lois nationales sur le droit d'auteur ne suffisent plus.

Bibliothèques et archives œuvrent dans un environnement de plus en plus numérique, mondialisé, et transcendant les frontières. Elles visent à donner un accès en ligne à leurs collections, à répondre à des questions et demandes venues de pays étrangers, et cherchent des moyens de partager des collections et des ressources pertinentes pour leurs partenaires dans les autres pays. Les exceptions et limitations au droit d'auteur, à l'inverse, s'interrompent aux frontières. Les bibliothèques et archives qui cherchent à travailler de manière transfrontalière se trouvent dans une zone légale floue, devant choisir entre rendre des services au risque de contrevenir aux lois sur le droit d'auteur, et ne pas du tout rendre ces services. À ce titre, l'IABD a d'ailleurs fait valoir en 2014 la position des bibliothèques dans le cadre de la consultation européenne sur la propriété littéraire et artistique.

Les licences privées ne permettent pas de résoudre ces problèmes.

Les bibliothèques qui achètent des abonnements à des ressources numériques se trouvent souvent forcées d'accepter des termes et conditions des licences, qui se réfèrent à des lois d'autres pays, et/ou qui contreviennent ou restreignent les exceptions et limitations existant dans leurs propres lois nationales. En outre, aucune licence n'est semblable à une autre, ce qui signifie que les bibliothèques qui désirent donner accès à leurs collections sous licence doivent respecter d'innombrables termes, conditions et lois étrangères. Les institutions à l'OMPI demandent un standard international de base pour les exceptions et limitations au droit d'auteur, qui ne pourrait être contourné par les termes des contrats privés.

Ces exceptions et limitations sont essentielles, car elles permettent aux bibliothèques et archives de remplir leurs missions, dans l'environnement imprimé comme dans l'environnement numérique.

Afin de pouvoir préserver un document quel qu'il soit, les bibliothèques et archives s'appuient sur une exception au droit d'auteur qui autorise cette opération. Pour pouvoir fournir des extraits d'œuvres à des fins de recherche ou d'étude, les bibliothèques ont besoin d'exceptions au droit d'auteur. Les exceptions au droit d'auteur permettent aux bibliothèques et archives de donner accès à l'éducation, à la recherche et à l'innovation, à la participation culturelle, à l'inclusion sociale ainsi qu'à de nouvelles formes de créativité.

L'assemblée générale de l'OMPI décidera de la future feuille de route de son comité permanent sur le droit d'auteur et les droits voisins en septembre 2014. Soutenant l'IFLA et les autres associations internationales, l'IABD demande :

- que la feuille de route actuelle, incluant des discussions sur des textes dans l'optique de conclure un ou des instrument(s) international/internationaux pour les bibliothèques et archives, soit prolongée de deux années supplémentaires ;
- que ces discussions bénéficient du même temps de traitement que le traité proposé sur la diffusion audiovisuelle ;
- qu'aucun autre nouveau sujet ne soit ajouté à la feuille de route du comité permanent.

Octobre

LE 13
L'AVENIR DES BIBLIOTHÈQUES : L'EXEMPLE DES BIBLIOTHÈQUES UNIVERSITAIRES

Aix-en-Provence, Campus des lettres

À l'occasion de la publication de *L'avenir des bibliothèques : l'exemple des bibliothèques universitaires*, aux presses de l'Enssib, l'Association des bibliothécaires de France, groupe PACA, organise une journée d'étude sur les bibliothèques universitaires, leur avenir, leur relation au public, dans un contexte en pleine évolution. L'occasion de réfléchir ensemble au sens de la bibliothèque aujourd'hui, y compris comme lieu.

www.abf.asso.fr/23/855/1575/ABF-Region/l-avenir-des-bibliotheques-l-exemple-des-bibliotheques-universitaires

LES 13 ET 14
RENCONTRES HENRI-JEAN MARTIN

Villeurbanne, Enssib

Pour la 8e édition de ces rencontres à destination des bibliothécaires en charge des fonds patrimoniaux, les échanges porteront sur la documentation régionale. La réflexion sur ce sujet riche, vaste et complexe s'appuiera sur trois pivots : la définition mouvante du local ou du régional dans les contextes urbains et numériques d'aujourd'hui ; les modalités de constitution spécifiques à un fonds local ; les enjeux liés à la mémoire locale.

www.enssib.fr/rencontres-Henri-Jean-Martin-2014

LE 17
LA CULTURE NUMÉRIQUE : AU-DELÀ DU DISCOURS, UN OBJET D'ENSEIGNEMENT ?

Rennes

Cette journée d'étude, organisée par le Groupe de recherche sur la culture et la didactique de l'information (GRCDI) et les équipes de recherche en documentation des ESPE de Rouen et de Caen, propose de revenir sur la notion même de « culture numérique » et de l'éclairer notamment sous trois angles : la dimension épistémologique et définitionnelle ; les enjeux éducatifs, de formation ; les enjeux didactiques.

http://culturedel.info/grcdi

DU 30 AU 1ER NOVEMBRE
OPENWORLD FORUM 2014

Paris, Le Beffroi

L'OpenWorld Forum est un sommet européen rassemblant les décideurs, les communautés, les développeurs pour « cross-fertiliser » les initiatives technologiques, économiques et sociétales ouvertes, afin de bâtir le futur du numérique.

www.openworldforum.paris/fr

Novembre

LE 6
7E JOURNÉE NATIONALE RAMEAU

Paris, BNF, Petit auditorium

Journée dédiée à l'actualité du réseau Rameau et à ses projets à l'heure du web sémantique.

www.bnf.fr/fr/professionnels/anx_journees_pro_2014/a.jp_2014_rameau_donnees_bnf.html

LE 7
LA BNF, SES MÉTADONNÉES ET VOUS ?

Paris, BNF, Petit auditorium

Journée consacrée à la réutilisation des métadonnées produites et diffusées par la BNF.

www.bnf.fr/fr/professionnels/anx_journees_pro_2014/a.jp_2014_rameau_donnees_bnf.html

LE 13
QUELLES POLITIQUES DOCUMENTAIRES EN ENVIRONNEMENT HYBRIDE ?

Villeurbanne, Enssib

Le mariage complexe d'une offre numérique et de collections matérielles pose des problèmes de gestion cohérente, de choix stratégiques, de services de médiation, rendus plus complexes du fait de la diversité des usages comme des tensions en moyens. Cette journée d'étude, organisée par le groupe Poldoc, propose de refléter l'approche contemporaine de ces enjeux.

www.enssib.fr/agenda/quelles-politiques-documentaires-en-environnement-hybride-0

DU 20 AU 21
BIENNALE DU LIVRE DE SCIENCES HUMAINES ET SOCIALES

Villeurbanne, amphithéâtre de l'Enssib

« *La fabrique du travail* » est la thématique retenue pour l'édition 2014 : quelles formes, entre flexibilité et rigidité, peut prendre le travail aujourd'hui ? Quelles attitudes face au travail, du refus à l'aliénation, sont à l'œuvre dans notre société ?

www.enssib.fr/biennale-du-livre-de-sciences-humaines-et-sociales

DU 25 AU 26
CARREFOUR DE L'IST

Nancy, Centre Prouvé

Le CNRS, Couperin et l'Université de Lorraine, en partenariat avec l'Abes, les organismes de Recherche, l'ADBU et le réseau Renatis, organisent ce deuxième « Carrefour de l'IST », une occasion de rencontres et d'échanges autour du thème : « *Les services aux chercheurs : innovations, évolutions, nouvelles interactions* ». Conférences et ateliers croiseront les points de vue des acteurs nationaux et européens de l'IST pour permettre de mieux appréhender les évolutions du métier.

www.inist.fr/?Rendez-vous-au-Carrefour-de-l-IST

LE 28
L'ACCÈS AUX DOCUMENTS ET DONNÉES DE L'ENSEIGNEMENT ET DE LA RECHERCHE

Université libre de Bruxelles

Cette journée d'étude, proposée par la section Aurore de l'Association des archivistes français, se penche sur les questions relatives à l'accès aux documents et données de l'enseignement et de la recherche (définition, archivage, notamment électronique, préservation et utilisation comme source).

www.archivistes.org/L-acces-aux-documents-et-donnees

Décembre

DU 8 AU 9 DÉCEMBRE

Paris, Maison des associations

Ce Forum, organisé par le Groupement français de l'industrie de l'information (GFII), rassemble l'ensemble des acteurs du marché de l'information et des connaissances autour d'un parcours de conférences stratégiques et d'ateliers élaboré par les experts des différents groupes de travail. Il favorise la rencontre des points de vue et les regards croisés à travers des échanges entre politiques, industriels, professionnels et chercheurs autour de thématiques transversales. Des temps de co-création avec les participants sont également prévus lors d'ateliers-laboratoires interactifs. Un événement incontournable pour identifier les tendances émergentes du marché, se former aux nouvelles pratiques et échanger avec d'autres professionnels sur les débats qui animent aujourd'hui notre écosystème.

→ http://forum.gfii.fr