Signaler les ressources numérisées : enrichissement, visibilité, dissémination

Dominique Stutzmann, Pauline Moirez

Pauline Moirez et Stutzmann, Dominique. « Signaler les ressources numérisées : enrichissement, visibilité, dissémination ». In *Manuel de constitution de bibliothèques numériques*, édité par Isabelle Westeel et Thierry Claerr, Bibliothèques. Electre-Cercle de la Librairie, 2013, p. 115-171 (version auteur).

MANUEL DE CONSTITUTION DE BIBLIOTHEQUE NUMERIQUE

PARTIE 2 : SIGNALER LES RESSOURCES NUMERISEES : ENRICHISSEMENT, VISIBILITE, DISSEMINATION

La bibliothèque numérique s'inscrit dans le projet d'établissement comme l'un des moyens d'assurer la médiation des ressources documentaires, c'est-à-dire le signalement, la communication et la valorisation, sur place et hors les murs. L'outil principal de la communication, aujourd'hui comme hier, est formé par le catalogue et les « métadonnées », à savoir les informations disponibles sur les documents. Ce sont elles qui vont être au fondement du signalement et de la communication des documents : elles constituent ainsi l'outil du professionnel et il lui appartient de savoir choisir dans sa boîte les instruments adéquats dont il va faire usage selon son projet de service dans un univers documentaire numérique.

La bibliothèque numérique dispose d'atouts pour répondre aux usages actuels du web :

- des métadonnées descriptives, dont la fiabilité est assurée par les normes et standards des bibliothèques : celles-ci permettent aux usagers d'accorder leur confiance aux contenus et favorisent le référencement et l'interopérabilité de la bibliothèque numérique ;
- des contenus (livres, images, sons, etc.) consultables en ligne, parfois même téléchargeables et réutilisables.

Mais ces atouts ne servent à rien si ces données et contenus ne peuvent être trouvés par l'usager. L'objectif n'est en effet pas qu'un internaute *puisse chercher* les ressources de la bibliothèque numérique, ni seulement qu'il *puisse les trouver*, mais qu'il les *trouve* et les *sélectionne* selon ses besoins[1]. Il est nécessaire pour cela que les données des bibliothèques numériques se trouvent sur le chemin de l'internaute, qu'elles s'adaptent à ses usages et à ses attentes. Il convient d'organiser et de baliser l'accès à l'information numérique à travers la surabondance informationnelle du web, pour accroître sa visibilité, sa fréquentation, et permettre également son enrichissement via les possibilités sociales et sémantiques ouvertes par les nouveaux usages et techniques du web.

L'absence de contact physique et de discussion entre le « bibliothécaire » et le « lecteur » ainsi que les spécificités du document numérique imposent de clarifier les moyens d'assurer le signalement des ressources numériques, car la description d'une ressource fait intervenir des informations de natures différentes, qui ont chacune leur utilité et dont le rôle est renforcé dans l'univers numérique.

---

[1] La citation de Roy Tennant « Only librarians like to search. Everyone else likes to find. » (« Digital libraries-cross-database search : one stop shopping », *Library Journal*, 2001, http://www.libraryjournal.com/article/CA170458.html), discutée par Jason R. Neal, (« Do Librarians really like to search », *Pragmatic Librarian*, 2007, http://pragmaticlibrarian.wordpress.com/2007/01/10/do-librarians-really-like-to-search/), rappelle que la bibliothéconomie a toujours considéré que le rôle de l'utilisateur est de trouver et choisir, non de chercher, et que le catalogue n'en est qu'un moyen. Voir aussi Charles A. Cutter, *Rules for a Dictionary Catalog*. 4e éd., Washington, D.C., Government printing office, 1904, p. 12 (« enable to find », « assist in the choice »). Le rapport FRBR décrit ainsi les tâches de l'utilisateur : trouver, identifier, sélectionner, se procurer (http://www.ifla.org/publications/functional-requirements-for-bibliographic-records)

Pauline Moirez et Stutzmann, Dominique. « Signaler les ressources numérisées : enrichissement, visibilité, dissémination ». In *Manuel de constitution de bibliothèques numériques*, édité par Isabelle Westeel et Thierry Claerr, Bibliothèques. Electre-Cercle de la Librairie, 2013, p. 115-171 (version auteur).

Dans ce nouveau paysage documentaire numérique, il importe de présenter les concepts sur lesquels repose le signalement des ressources numérisées, depuis le document en lui-même jusqu'aux différents niveaux de métadonnées (§ 1). Ce préalable conceptuel servira de base à une analyse comparative approfondie des différents formats de métadonnées qui peuvent être utilisés pour le signalement de ces ressources numérisées, en fonction des usages et des fonctionnalités attendus pour la bibliothèque numérique (§ 2).

Mais la bibliothèque numérique n'est pas un lieu unique et fermé : de par leur nature numérique, les contenus et métadonnées peuvent exister à plusieurs endroits en même temps, sur le web comme dans la bibliothèque. Il est donc nécessaire de s'interroger sur son architecture de gestion, et d'analyser les bénéfices d'un enrichissement des métadonnées depuis l'extérieur (partenaires, contributions des internautes, processus automatisés de fouille de données), pour l'amélioration du signalement et de l'expérience de recherche des utilisateurs (§ 3). Échanger les données de la bibliothèque numérique, augmenter leur visibilité et les enrichir de données provenant d'autres sources dans ce vaste écosystème qu'est le web, reposent sur des techniques d'interopérabilité que l'on peut étudier au regard des usages attendus et des publics visés (§ 4).

1. LE NOUVEAU PAYSAGE DOCUMENTAIRE : DOCUMENTS, METADONNEES, DONNEES D'AUTORITE, IDENTIFIANTS

Le paysage documentaire actuel, où l'information numérique a acquis une place dominante, a suscité de nombreux discours sur la révolution numérique et la fin des bibliothèques. Pourtant seuls les moyens ont été modifiés, et non les objectifs, fonctions et missions des bibliothèques. À y regarder de plus près, il apparaît même que les concepts-clefs de l'univers numérique sont ceux-là même que les bibliothèques ont l'habitude de manipuler, même si la matérialité des objets a évolué : « documents », « métadonnées » (anciennes « descriptions »), « données d'autorité » (aussi « référentiels », « ontologies ») et « identifiants » sont au cœur de la communication électronique et relèvent du domaine de compétence des bibliothèques et des métiers de l'information.

Au premier niveau se trouve **le document** lui-même. Celui-ci était la base du catalogage dans la bibliothèque d'autrefois et ne pouvait être communiqué que par le truchement d'un double travail de médiation : celle du catalogage en amont, celle d'interrogation du catalogue en aval. Dans l'univers documentaire, les ressources numériques présentent peu de spécificités par rapport aux ressources traditionnelles, malgré leur nature différente : la bibliothèque peut les conserver, les signaler, les communiquer, les valoriser. Les différences techniques font qu'ils ne sont pas consultables sans le truchement d'un matériel et (au moins) un logiciel intermédiaire, mais c'était déjà le cas des documents sonores et audio-visuels. Ce qui change profondément, c'est 1° la nature ubiquitaire des ressources, qui peuvent être consultées à plusieurs endroits en même temps, 2° leur mutabilité (sans commune mesure avec les publications à feuillets mobiles et leurs mise à jour) et 3° la nature agrégative et la granularité des ressources numériques. Ces trois révolutions impliquent des modifications dans l'emploi des métadonnées et l'organisation des accès.

Il faut en particulier insister sur la révolution que représente la prise en compte accrue des différents niveaux de granularité. Tout d'abord le document n'est plus une unité insécable et l'objet unique de la médiation : chacun des mots contenus ou chacun des énoncés constitue une information particulière, qui peut faire l'objet d'une médiation. En outre, le document numérique

est, plus souvent encore que le texte imprimé, un objet qui inclut et associe en son sein des unités autonomes, qui peuvent avoir une existence indépendante.

Le document numérique jouit, en effet, d'un statut nouveau, car son contenu est directement indexable et le document numérique peut être considéré comme sa propre description à l'échelle 1. C'est en particulier vrai pour le contenu textuel dans le cas d'une recherche dite « plein texte » (souvent indexé à partir du résultat d'un logiciel d'OCR), mais il en est de même pour les images qui deviennent leur propre description, grâce aux moteurs de recherche par le contenu (formes, couleurs, reconnaissance faciale, etc.) ou aux documents audiovisuels avec les logiciels d'analyse vocale.

L'exemple des pages web illustre cette modification profonde intervenue dans le monde documentaire : avec le passage des anciens annuaires de sites (par exemple Altavista, Yahoo! Directory) à l'emploi massif des moteurs de recherche généralistes (par exemple Google, Bing, etc.), il est devenu usuel de rechercher des mots que l'on désire voir apparaître dans une page. Dans ce cas, la page HTML est sa propre description : elle contient les mots cherchés, et l'usage du moteur de recherche impose naturellement le document comme sa propre description, comme si le catalogue de bibliothèque avait été en même temps un dictionnaire des citations ou une concordance textuelle, et avait permis de chercher non seulement les ouvrages sur un thème (Écriture – France – Moyen Âge), mais aussi ceux contenant certains mots (« tablette de cire », « parchemin », « stylet », « plume »).

Ensuite de nouveaux niveaux pertinents de granularité sont créés par la possibilité d'inclure, en apparence, des documents autonomes dans une ressource tierce. De nombreux sites web sont partiellement constitués par des ressources qui ne se trouvent pas physiquement au même endroit[2] : non seulement, les images et fichiers de grande taille peuvent être stockés sur d'autres serveurs pour des raisons techniques, mais surtout il est devenu possible de tirer parti de la nature ubiquitaire du document numérique en incluant des ressources provenant d'autres éditeurs (par exemple lecteur exportable de Gallica ou vidéos embarquées de YouTube).

À un deuxième niveau se trouvent **les métadonnées**, c'est-à-dire *les données sur les données*, ou ce que l'on peut dire *sur* le document pour le décrire. Le signalement par les métadonnées peut être plus ou moins riche et précis, tant sur la forme que sur le contenu. Ainsi une même publication pourra-t-elle être décrite – et éventuellement identifiée – avec le triptyque auteur / titre / date, ou avec la précision « 1 vol. (ix-656 p). – 21 cm » ou avec la mention des lieux et maisons d'édition. Par rapport aux caractéristiques du catalogage classique, de nouvelles formes de description s'insèrent dans l'univers numérique : non seulement les formats et versions des documents numériques (par exemple Epub, PDF 9.2, 145 Mo), mais aussi des éléments relevant d'un autre mode de perception (comme la vignette représentant la couverture du document). Selon la nature de l'information portée, on distingue trois types de métadonnées :

- « métadonnées descriptives » ou informations d'identification, comprenant l'essentiel de la description bibliographique (titre, auteur, mots-clefs et indexation, identifiants de l'édition ou de l'exemplaire),

- les « métadonnées techniques » ou informations de structure, c'est-à-dire les données sur la version du document, la date, et le format, ainsi que les liens vers les ressources

[2] Nous entendons ici par « emplacement physique » le répertoire du serveur d'où le document est consulté.

apparentées (pour un livre numérisé, les métadonnées techniques peuvent comprendre non seulement le format des images et la description de la structure, à savoir l'ordre des images, mais aussi la cote de l'exemplaire numérisé, le modèle de l'appareil de reproduction, les versions des logiciels de traitement d'image),

- les « métadonnées administratives », décrivant la propriété intellectuelle, les droits d'accès, les informations sur la préservation (historique des modifications) et l'archivage pérenne de la ressource.

De même qu'avec le catalogage à la source, avec présence de la description bibliographique à l'intérieur du document lui-même, les métadonnées peuvent être intégrées dans le document numérique. C'est particulièrement le cas dans les documents au format Epub et TEI, mais aussi pour les fichiers d'images qui peuvent comporter un en-tête IPTC ou Exif à l'intérieur d'un fichier JPEG. Les métadonnées peuvent évidemment encore être contenues dans une base extérieure, en particulier un *catalogue de bibliothèque numérique*.

La richesse de description est une problématique primordiale dans l'univers numérique, car l'approche de la granularité est profondément modifiée. Pour rendre compte de celle-ci, les outils traditionnels (catalogage à niveau, par exemple) ne répondent plus aux attentes. Les règles de catalogage excluent le dépouillement des périodiques et ressources continues, ou les reportent à d'autres ressources documentaires (bases bibliographiques, etc.) : cela n'a pas besoin d'être modifié. En revanche l'information numérique permet de réunir les différents degrés d'accès et impose de réfléchir à leur articulation pour des raisons tant intellectuelles que juridiques, matérielles, éditoriales et d'usage. Du point de vue intellectuel, en effet, si un article est muni d'une planche reproduisant un tableau, telle page de tel article sera incluse dans un double ensemble matériel et intellectuel qui doit son existence au processus éditorial, mais qui peut être analysé séparément sous les deux aspects, physique d'une part (« pl. I », entre les pages 46 et 47, du fasc. 2, du vol. 46 de tel périodique, paru en telle année) et logique d'autre part (tableau de tel peintre, illustrant le paragraphe 4 de la partie 2 de l'article de tel auteur). L'accès direct, une fois la ressource numérisée ou si celle-ci est nativement numérique, impose de gérer les différents niveaux de granularité : une œuvre sous droits pourra être munie de droits de diffusion différents de celle dans laquelle elle est incluse à des fins d'illustration ou d'argumentation. (Ainsi, dans Persée, trouve-t-on des espaces blancs pour les œuvres non libres de droits, et dont les droits n'ont été cédés que pour l'édition papier.)

Cette accessibilité multiple, intervenant à des degrés divers, suscite de nouveaux usages à partir d'un même besoin constant, qui est de repérer et de consulter une ressource pertinente. Le catalogue traditionnel ne donnait accès qu'à un unique niveau de description logique car la bibliothèque elle-même ne donnait accès qu'à un niveau matériel : un volume ou une « unité de conservation ». Dans le cas du périodique, la plupart des fascicules n'ayant pas d'unité intellectuelle propre, on décrivait le périodique comme un ensemble unique, celui de l'accès logique, en signalant les hors-série et numéros thématiques, précisément pour offrir l'accès à un niveau jugé pertinent. Dans l'univers numérique, c'est l'accès à l'unité logique de plus petite taille qui prime : dans le cas du périodique, il s'agira de l'article plutôt que du fascicule thématique ou du périodique lui-même, à l'instar des pages web vers lesquelles on crée des liens directs. La nécessité de donner accès à niveau logique de dimension inférieure à l'unité matérielle publiée a, évidemment, des incidences sur la constitution de la bibliothèque numérique et la gestion de ses métadonnées pour rendre compte de documents numériques de nature ubiquitaire, évolutifs et agrégatifs.

À un troisième niveau se situent **les données d'autorité**. Elles sont aussi des métadonnées qui, *in fine*, portent sur les ressources auxquelles les bibliothèques numériques peuvent donner accès, mais elles en sont indépendantes. Qu'elles portent sur les personnes, les œuvres, les noms de lieux ou les sujets, elles servent à identifier correctement les ressources et, ce faisant, à concentrer l'accès et à naviguer dans les données. Elles forment comme une mappemonde documentaire sur laquelle chacune des ressources de la bibliothèque numérique devrait être positionnée pour en permettre le repérage aisé et rapide par les usagers.

Enfin, **au quatrième niveau**, des métadonnées d'un type spécifiques sont constituées par **les identifiants**. Qu'il s'agisse de la cote d'une ressource particulière, de l'ISBN d'une édition, de l'URL d'accès d'une ressource numérique, ces identifiants permettent d'identifier et de créer du lien. De même que le document peut être sa propre métadonnée et être muni d'un identifiant, de même chaque métadonnée et toutes les données d'autorité (un nom de personne avec ses formes rejetées, un sujet ou une construction RAMEAU, etc.) peuvent être désignée par des identifiants. Ceux-ci peuvent être soit « absolus »[3] (valables indépendamment du contexte, comme une URL donnée en entier), soit « relatifs » (comme la cote d'un volume, qui ne le désigne qu'à l'intérieur d'une bibliothèque).

Le document, les métadonnées descriptives, les données d'autorité, et les identifiants forment quatre entités qui permettent le signalement, le repérage et l'accès aux ressources d'une bibliothèque, notamment numérique. La « redocumentarisation »[4], c'est-à-dire la transformation d'un document en un nouveau document à décrire, notamment dans le processus de numérisation, d'une part, ainsi que la transformation du document en sa propre métadonnée et la plus fine granularité de signalement pertinent, d'autre part, engendrent un véritable « continuum descriptif » : depuis l'espace du mot (par exemple, quand, dans une bibliothèque numérique un mot recherché est surligné en jaune lors d'une requête) jusqu'au monde des ressources continues, en passant par les paragraphes, les pages, les chapitres et les volumes, tout niveau peut désormais être décrit dans sa spécificité, aussi bien matérielle que logique. Cependant, afin que le continuum descriptif ne devienne pas un flou documentaire, les bibliothécaires numériques doivent connaître, gérer et maîtriser les métadonnées et leurs caractéristiques. C'est à cette condition que le numérique permet de réaliser une véritable « bibliothèque hors les murs » et non pas seulement une nouvelle bibliothèque virtuelle, qui serait aussi close et intimidante que celles enterrées et gardées par des tours : une bibliothèque numérique, accessible et ouverte, à même de répondre aux besoins documentaires de chaque utilisateur et de redonner à chaque ressource conservée par une bibliothèque la chance de retrouver un public, en la rendant « appelable » depuis l'extérieur, c'est-à-dire qu'on peut demander, recevoir et consulter hors de la bibliothèque.

Dans ce processus, le rôle des métadonnées est central : elles permettent de faire le lien entre les utilisateurs et la bibliothèque numérique, elles sont le passage obligé entre le public desservi et la mise en œuvre de la bibliothèque. Les métadonnées sont de diverses natures et, surtout, elles sont évolutives : elles ne se présentent pas sous une forme figée et ne sont pas uniquement le produit de l'indexation par le bibliothécaire. Elles peuvent être enrichies tant par le

---

[3] Parfois aussi dits « déréférençables ».

[4] Jean-Michel Salaün, « Web, texte, conversation et redocumentarisation », dans JA*DT 2008 : actes des 9es Journées internationales d'Analyse statistique des Données Textuelles, Lyon, 12-14 mars 2008 : proceedings of 9th International Conference on Textual Data statistical Analysis, Lyon, March 12-14*, 2008, éd. par Serge Heiden et Bénédicte Pincemin (Lyon: Presses universitaires de Lyon, 2008), 1198, https://papyrus.bib.umontreal.ca/jspui/bitstream/1866/2226/1/salaun-jm--jadt-2008.pdf, http://jadt2008.ens-lsh.fr/spip.php?article197, http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2008/pdf/salaun.pdf.

bas (par le document lui-même et son contenu, par des processus de fouille de données et d'extraction automatique d'informations[5]) que par le haut (exploitation des anciens systèmes de cotation et de classification) ou encore par la liaison avec des ressources autonomes qui peuvent interagir (autres bases de données, autres indexations, dont sociales[6]). Ces processus dynamiques ne sont pas contradictoires, mais permettent au contraire un heureux mariage : les métadonnées nouvelles auront des degrés de structuration divers, pourront être gérées en interne par la bibliothèque numérique ou y être extérieures ; elles ont toutes leur utilité, leur pertinence et leur usage. En tout cas, elles sont désormais au centre de la focale : il existe un mouvement de balancier entre document et métadonnées, et, à l'heure actuelle, c'est la métadonnée qui est au cœur des réflexions.

2. ÉVALUER LES DIFFERENTS FORMATS

Les méthodes de signalement des ressources numériques sont multiples et correspondent à des usages. De même que le catalogue jouait pleinement son rôle grâce, d'une part, à sa qualité et à son exhaustivité et, d'autre part, à la connaissance et l'accompagnement des professionnels, de même, dans l'univers numérique, le signalement dépend largement de sa qualité propre et des efforts d'accompagnement qui permettent à chaque utilisateur de trouver les documents qui répondront à son besoin, et à chaque document d'atteindre les utilisateurs, comme le préconisent les lois de Ranganathan[7].

Pour structurer les métadonnées, il existe différents **formats**, qui sont autant de grammaires descriptives. Ceux-ci ont des caractéristiques diverses et ont, chacun, été d'abord pensés pour répondre à des besoins précis, de sorte que les uns ont mis l'accent sur la capacité à décrire une structure avant celle de décrire un contenu intellectuel, tandis que d'autres décrivent les ressources isolément et à plat, sans structure et que d'autres encore sont spécifiquement pensés pour la numérisation des ressources analogiques traditionnelles.

Il existe des centaines de formats de métadonnées[8]. Aucun n'est, en soi, parfait, tous sont à évaluer en fonction des usages et fonctionnalités que l'on veut mettre en place. Il convient de préciser ici comment évaluer les différents standards et structurations de métadonnées, car celles-ci sont l'outil principal du signalement des ressources et nous illustrerons le propos en précisant les qualités des principaux formats utilisés dans l'environnement des bibliothèques numériques, en particulier :

---

[5] Voir ci-dessous § 3.B.b) Enrichissement automatique des métadonnées, p. ###.

[6] Voir ci-dessous § 3.B.a) Enrichissement social et *crowdsourcing, p.* ###.

[7] Ranganathan, Shiyali Ramamrita. *The five laws of library science*. Madras : Madras Library Association ; London : Edward Goldston, 1931 [Numérisé et disponible en ligne : < http://arizona.openrepository.com/arizona/handle/10150/105454>) Les livres sont faits pour être utilisés ; 2) À chacun son livre ; 3) À chaque livre son lecteur ; 4) Il faut épargner le temps du lecteur ; 5) Une bibliothèque est un organisme en croissance.

[8] Cf. Jenn Riley, « Seeing Standards: A Visualization of the Metadata Universe », *Indiana University Digital Library Program*, 2010, http://www.dlib.indiana.edu/~jenlrile/metadatamap/.

- des formats d'encodage et de structuration du plein texte des documents[9] : ALTO (*Analyzed Layout and Text Object*)[10], EAD (*Encoded Archival Description*)[11], TEI (*Text Encoding Initiative*)[12],

- des formats de description des ressources numériques : Dublin Core, formats MARC (*MAchine-Readable Cataloging*), METS (*Metadata Encoding and Transmission Standard*)[13], MODS (*Metadata Object Description Schema*)[14], ONIX for Books[15],

- des formats pour exprimer les données d'autorité : MADS (Metadata Authority Description Schema)[16], FOAF (Friends of a Friend)[17].

Plus ou moins riches, elles sont aussi plus ou moins contraignantes techniquement et exigeantes humainement. Aussi doivent-elles être adaptées au projet de service de la bibliothèque numérique (voir partie 4), tant en infrastructure qu'en personnel et selon les données déjà disponibles préalablement à la constitution de la bibliothèque numérique, mais aussi en termes d'usages attendus : les formats de métadonnées ne sont pas des réponses universelles, mais des outils à adapter à l'information disponible et à la nature des objets décrits, en se posant la question « quelles métadonnées pour quels documents et pour quels usages ? ».

Chaque format impose des contraintes de structuration de l'information qui favorisent une certaine forme d'interopérabilité : c'est ainsi que les formats MARC permettent d'échanger des notices bibliographiques. Les formats sont définis, dans leur structure, par des normes qui peuvent, elles-mêmes, renvoyer à d'autres normes pour uniformiser les données et les rendre comparables (les règles de catalogage renvoient aux normes de constitution des titres uniformes et titres de forme, règles d'indexation matière, etc.). Pour les points qui ne sont couverts ni par les normes générales ni par les contraintes du format, il faut s'en remettre aux « bonnes pratiques » et aux guides publiés par des institutions riches d'expérience. Des formats riches et malléables, tels

---

[9] Voir *Mémento sur les langages d'encodage et de structuration de textes*, BnF, 2009, http://www.bnf.fr/documents/formats_textes.pdf.

[10] ALTO est un format utilisé pour la conversion des textes à partir des images suite à un OCR. Il conserve toutes les coordonnées du contenu (texte, illustrations, graphiques) dans l'image et permet la superposition de l'image et du texte ainsi que la surbrillance des mots recherchés lors d'une requête. Voir *Conversion OCR et format ALTO*, http://www.bnf.fr/fr/professionnels/num_conversion_texte/s.num_conversion_texte_ocr.html.

[11] EAD est un format maintenu par la Bibliothèque du Congrès pour l'encodage des répertoires et inventaires d'archives, cf. http://www.loc.gov/ead/.

[12] La TEI permet d'encoder des textes, en particulier les textes littéraires et linguistiques. Elle rend compte de l'organisation logique d'un texte (chapitres, sections, citations, vers, noms propres mentionnés dans le texte, etc.). Voir http://www.tei-c.org/index.xml.

[13] METS est un format spécifiquement conçu pour associer des métadonnées descriptives, administratives et structurelles des objets dans une bibliothèque numérique. Voir Library of Congress, « Metadata Encoding and Transmission Standard (METS) Official Web Site », 2011, http://www.loc.gov/standards/mets/.

[14] Inspiré de Marc21, MODS est un format descriptif dans un formalisme XML, maintenu par la Bibliothèque du Congrès pour les descriptions bibliographiques. Il est développé parallèlement au format MADS. Cf. http://www.loc.gov/standards/mods/.

[15] Format pour le commerce électronique du livre, maintenu, avec des formats associés, par the EDIÆUR, consortium international composé de représentants du monde de l'édition, pour favoriser le commerce électronique des livres, livres électroniques et périodiques, cf. http://www.editeur.org/93/Release-3.0-Downloads/.

[16] MADS est un format maintenu par la Bibliothèque du Congrès pour décrire les autorités. Il est développé parallèlement au format MODS. Cf. http://www.loc.gov/standards/mads/.

[17] FOAF est une ontologie très utilisée par les acteurs du web de données, qui permet de décrire précisément les personnes et les liens entre personnes http://www.foaf-project.org/.

que METS, ou à la capacité descriptive trop floue, tels que EAD et Dublin Core, sont particulièrement concernés[18]..

Au final, le choix d'un format de métadonnées pour une collection numérique ou pour une bibliothèque numérique doit s'appuyer autant sur la nature et le type des documents décrits que sur les usages attendus (fonctionnalités, finesse d'interrogation, interopérabilité et échanges de données avec d'autres réservoirs de métadonnées, etc.). Il n'y a pas de « bon » format de métadonnées, il n'y a de que des formats plus ou moins adaptés à des besoins.

A. QUALITES DES FORMATS DE METADONNEES

Le signalement des ressources nécessite des métadonnées de « qualité » pour permettre un accès correct aux ressources pertinentes. La qualité s'évalue objectivement, selon chacun des critères qui structureront ensuite également les fonctionnalités et les services offerts aux usagers. L'analyse porte ici non pas sur la qualité des informations descriptives contenues dans les métadonnées, mais sur la qualité des formats qui les expriment, au regard des besoins et usages.

Pour mémoire, il faut rappeler ici que le premier critère de qualité des métadonnées est évidemment leur justesse, tant dans leur valeur intrinsèque que dans l'utilisation du format choisi, quel qu'il soit. Il s'agit ici d'un concept simple et évident en apparence.

Dans son aspect le plus élémentaire, cela équivaut à la chasse aux coquilles ou aux fausses informations ; cela peut aussi être le rattachement aux bonnes données d'autorité ou la construction correcte de l'identifiant. La manipulation d'un nombre démultipliés d'entités documentaires (chaque page constitue une image qui a son propre nom) et l'importance des liens, souvent fondés sur une suite semi-aléatoire de chiffres, rendent plus graves les conséquences des erreurs, ou du moins les rendent plus difficiles à détecter et à corriger : réattribuer une page numérisée à un volume imprimé peut se révéler très ardu si le fichier a été mal nommé. Les mécanismes d'automatisation et de contrôle sont donc très importants, et, d'un point de vue concret, il vaut mieux consacrer du temps à paramétrer une procédure d'exportation depuis le catalogue plutôt que de travailler à partir de copier-coller vers un tableur (type *Microsoft Excel*), procédure où se peuvent glisser d'infimes erreurs (oubli du dernier caractère) qui auront des conséquences sur la qualité de l'accès.

À un stade plus élaboré, il s'agit du respect des normes nationales et internationales en vigueur pour la saisie et la structuration des valeurs de métadonnées, par exemple, la transcription du titre[19], les abréviations à employer[20] ou le nom des auteurs[21].

---

[18] Voir ci-dessous. § 2.C. Usages et interopérabilité, p. ###.

[19] Normes sur la translittération et la romanisation : ISO 9:1995 (Translittération des caractères cyrilliques en caractères latins), 233 :1984 (Translittération des caractères arabes en caractères latins), ISO 259:1984 (Translittération des caractères hébraïques en caractères latins), ISO 843:1997 (Conversion des caractères grecs en caractères latins), NF ISO 3602:1989 (Romanisation du japonais), NF ISO 7098:1991 (Romanisation du chinois), ISO 9984:1996 (Translittération des caractères géorgiens en caractères latins), ISO 9985:1996 (Translittération des caractères arméniens en caractères latins), ISO 11940:1998 (Translittération du thaï), ISO 11941:1996 (Translittération de l'écriture coréenne en caractères latins), ISO 15919:2001 (Translittération du Devanagari et des écritures indiennes liées en caractères latins).

Le respect des normes et formats est impératif pour assurer l'interopérabilité et le référencement des ressources présentées par les bibliothèques numériques. Ainsi, pour les personnes physiques, l'on saisira globalement « Nom, Prénom (date-date ; fonction) », indépendamment des règles d'affichage dans un contexte d'usage : pour une bibliothèque numérique à destination des enfants, on préférera peut-être « Jean de La Fontaine » à « La Fontaine, Jean de (1621-1695) ». Les normes ont en effet été élaborées pour répondre à un besoin qui n'a pas cessé, mais s'est au contraire renforcé dans la multiplication des sources et des informations disponibles : permettre le repérage rapide et univoque des ressources, et permettre l'interopérabilité entre bibliothèques. Aussi est-il toujours nécessaire de respecter ces normes.

Les paragraphes qui suivent ne visent pas à donner des indications sur la valeur des métadonnées elles-mêmes pour juger de la qualité des métadonnées, mais incitent à la réflexion sur la qualité du format et l'adéquation de celui-ci au projet de service de la bibliothèque numérique.

FORMAT ET VALEURS

Un préalable indispensable pour assurer la qualité des métadonnées réside dans l'adéquation entre le format choisi (quel qu'il soit) et les valeurs renseignées. Il est parfaitement possible de décrire ainsi le sujet d'une ressource qui porterait sur la construction ferroviaire autour de 1900 : « Réseau ferroviaire en France vers 1900 », mais si l'on est dans une bibliothèque qui utilise RAMEAU, on devra préférer « Chemins de fer -- France -- 1870-1914 ». Un exemple tiré de métadonnées repérées dans des bibliothèques numériques existantes illustrera notre propos : le format Dublin Core prévoit de dissocier le sujet propre (« Chemins de fer », sous le terme « Sujet ») du périmètre géographique et temporel concerné (« France -- 1870-1914 », sous le terme « Couverture »). On a vu aussi bien « Sujet : Chemins de fer -- France -- 1870-1914 » que « Couverture » utilisé pour donner le lien à l'imagette de couverture ! Dans ce cas, les métadonnées ne sont pas fausses en elles-mêmes, mais fausses par rapport au format choisi. La double cohérence, avec l'objet décrit et avec la grammaire descriptive choisie, est un premier impératif de qualité.

Les métadonnées géographiques : formats et gestion

Les informations géographiques sont particulièrement utiles dans l'univers numérique pour créer des interfaces riches (accès par carte) ou des services documentaires mobiles[22]. Dans le domaine des cartes, des initiatives importantes montrent déjà les potentialités tant pour le catalogage que pour la recherche, par exemple CartoMundi, un projet collaboratif de valorisation en ligne du patrimoine cartographique[23].

Les normes pour renseigner les informations géographiques sont NF ISO 3166 (Codes pour la représentation des noms de pays et de leurs subdivisions) et XP Z 44-002 (codes pour la

[20] Règles pour l'abréviation des mots dans les titres et des titres des publications, ISO 4:1997 ; Règles pour l'abréviation des termes bibliographiques, ISO 832:1994.
[21] Forme et structure des vedettes noms de personnes, des vedettes titres, des rubriques de classement et des titres forgés, NF Z44-061 (1986) ; Forme et structure des vedettes de collectivités-auteurs, NF Z 44-061 (1996).
[22] Voir ci-dessous § 3.B.c) Métadonnées et visualisation, p. ###.
[23] http://www.cartomundi.fr/site/.

représentation des noms de pays historiques). Ces normes restent d'actualité, tant dans le domaine bibliographique (UNIMARC champ 123 et 206 pour les coordonnées des cartes ; UNIMARC champs 6XX $y pour les compléments géographiques de RAMEAU) que pour les données d'autorité (UNIMARC champ 123 pour les autorités géographiques) que dans les autres cadres (Exif ou IPTC). En MARC, les codes d'aire géographique (champs 660 du format bibliographique et 160 pour les autorités) sont utilisables également.

Certains formats ne précisent pas quel référentiel utiliser[24]). Ainsi, en Dublin Core, les expressions suivantes sont équivalentes :

Coverage : « France »
Coverage : « W 5°5' - E 9°33' / N 51°4' - N 41°18' »

Dans le profil d'application de Dublin Core pour les bibliothèques[25] a été proposé un moyen de décrire l'ensemble :

Coverage : « name=France; westlimit=5.5; eastlimit= 9.33; northlimit=51.4; southlimit=41.18 »

Les principales difficultés présentées par ces métadonnées géographiques résident dans l'exploitation de l'information. La première difficulté est la gestion différenciée des surfaces (un pays, une région) ou des points (le lieu représenté par une photographie), avec des problèmes importants de méréologie : dans un contexte général, une ville peut être décrite par les coordonnées de son point kilométrique ; dans un contexte local, on différenciera la ville comme surface et les points d'intérêt concernés par les ressources de la bibliothèque numérique (monuments, rues, etc.). Le choix descriptif se fera dans le respect des normes, selon les objectifs de service de la bibliothèque numérique.

La seconde difficulté est la gestion des coordonnées spatiales pour réaliser des accès géographiques, soit dans la présentation des ressources, soit dans les interfaces de recherche. Les métadonnées descriptives formulées en texte (par exemple « France ») ne comportent pas de coordonnées géographiques exploitables pour établir une carte : celles-ci sont soit reportées, au mieux, dans une notice d'autorité liée par un identifiant, soit simplement absentes. Dans ce second cas, la liaison avec des référentiels extérieurs tels que GeoNames[26] permet de pallier partiellement cette absence, en effectuant une requête sur une chaîne de caractère et en rapatriant des coordonnées géographiques pour une exploitation par un système d'information géographique (SIG).

De nombreuses applications mobiles et web exploitent directement les en-têtes des images fixes (Flickr et Panoramio, albums photographiques des smartphones, etc.). Jusqu'à présent, les bibliothèques numériques ont renoncé à insérer dans un en-tête IPTC les coordonnées géographiques disponibles dans les métadonnées descriptives du catalogue, car l'en-tête IPTC comporte normalement le lieu de la prise de vue, c'est-à-dire les coordonnées géographiques de l'atelier de numérisation plutôt que celles du lieu où a été pris le cliché d'origine. Une réflexion

[24] Voir ci-dessous § 2.A.d) Association avec des données d'autorité, p. ###.
[25] Dublin Core – *Library Application Profile* (DC-Lib) http://dublincore.org/documents/library-application-profile/.
[26] GeoNames est une base de données géographique en ligne (http://www.geonames.org) librement réutilisable.

sur le format serait à mener.

STRUCTURATION DES METADONNEES

Ensuite, la qualité des métadonnées se mesure à leur structuration et à la possibilité de rendre compte et d'exploiter les différents niveaux de granularité de l'univers documentaire numérique. Et, sur ce point, les différentes grammaires descriptives ne sont pas égales, car elles ont des caractéristiques fortement divergentes. Il existe, en effet, des formats de métadonnées qui ont une plus ou moins grande force de structuration et hiérarchisation des informations : de ce point de vue, la hiérarchie des formats selon leur qualité place METS en première place, puis, dans l'ordre, EAD, TEI, MODS et Dublin Core.

En effet, un format tel que Dublin Core simple ne permet pas de décrire les relations méréologiques (de tout à partie) entre ressources. Chaque ressource est décrite comme une entité indépendante et si l'on veut décrire à la fois une illustration dans un chapitre, le chapitre et le livre qui les contient, il faudra établir trois notices séparées, qui seront liées entre elle par un champ répété de « relation », qui ne distingue pas les différentes relations possibles ni leur sens (inclut/est inclus ; fait référence à/est référencé dans, etc.). L'utilisation de Dublin Core qualifié permet de préciser les rapports d'inclusion, mais reporte au système de gestion le soin d'établir les liens qui ne sont pas structurels.

À l'opposé se trouve le format METS qui est une grammaire qui se concentre uniquement sur la possibilité de hiérarchiser et de structurer l'information, mais n'impose aucune contrainte sur la description des ressources elle-même[27]. C'est un format aux potentialités extrêmes : il s'agit, en effet, non seulement d'un « format conteneur » qui autorise l'usage des autres formats descriptifs à l'intérieur de sa structure et permet de décrire un même document selon des ordres variés, mais aussi d'un format fondé sur la nature connective de l'information numérique, qui permet de conserver des parties de description en dehors de la description en METS.

Prenons des exemples pour expliciter chacune de ces deux caractéristiques. Dans le format METS, on peut de décrire aussi bien une structure matérielle qu'une structure logique. Ainsi, pour un manuscrit ou un livre mal relié, il est possible de décrire la succession des feuillets et de donner accès à chacun d'eux dans l'ordre où ils sont conservés, mais il est aussi possible de dire dans quel ordre il faut lire les pages et autoriser ainsi une consultation plus aisée. Dans le cas le plus fréquent, il s'agira de distinguer les différentes parties intellectuelles d'un même volume : si la structure matérielle est unique et linéaire (une page suit la précédente et devance la suivante), la structure logique est hiérarchisée, divisée en articles, parties, chapitres, paragraphes, etc. Le format METS permet de décrire aussi bien la structure matérielle plate que des structures à plusieurs niveaux et, notamment, d'indiquer qu'une même page fait partie ou contient deux ou plusieurs ensembles logiques distincts.

Le format METS permet également de pointer vers des descriptions externes, en indiquant un simple lien, relatif ou absolu, à l'intérieur du champ de description. Cette possibilité offre à ce format un avantage de taille : il permet de maintenir la gestion des informations bibliographiques

[27] Voir l'encart sur METS dans le chapitre sur la préservation.

par les applications existantes, tel le catalogue de la bibliothèque, tout en y ajoutant la capacité de structuration pour décrire un document numérique au plus grand degré de granularité, avec des informations qui peuvent être saisies durant les opérations de la numérisation. Par exemple, un prestataire de numérisation pourra indiquer les planches, les pages avec illustration, les pages de titre, voire saisir les tables des matières, pendant la numérisation, sans avoir à redoubler la description bibliographique. L'avantage de cette méthode est que toute modification des informations de la notice maîtresse sera, de fait, reportée dans la notice METS. L'inconvénient est l'éventuelle lourdeur de gestion : en effet, d'une part, la stabilité des liens constitue un pré-requis indispensable pour utiliser cette méthode, de sorte qu'il faut donc munir chaque notice du catalogue d'un identifiant pérenne ; d'autre part, il faut disposer d'applications informatiques capables de suivre les liens si l'on veut disposer en même temps des informations de nature bibliographique et des compléments de structure et de granularité.

Une solution hybride consiste à intégrer les données descriptives essentielles pour faciliter la lecture du fichier METS et l'identification du document concerné, et à renvoyer à la description actualisée dans le catalogue par un lien. Dans le cadre de la numérisation, l'importation de métadonnées descriptives issues du catalogue au sein d'un fichier de structure METS est une bonne pratique, qui rend plus faciles les contrôles de qualité et de cohérence. En bonne logique, ces données ne devraient pas être utilisées autrement que dans ce cadre, puisque les informations de référence restent dans le catalogue et sont davantage susceptibles d'être modifiées. Pourtant, il faut craindre que ces données soient tout de même, à un moment ou à un autre, utilisées par facilité pour renseigner les propriétés des fichiers dérivés (par exemple, un fichier PDF ou un fichier Epub) ; aussi leurs qualité et précision ne doivent-elles pas être négligées, non plus que le choix du format contenu, car ce sont ces métadonnées qui apparaîtront une fois le fichier téléchargé ou exporté hors de la bibliothèque numérique. Ce choix peut être formalisé dans un « profil d'application[28] ».

L'EAD est un format intermédiaire : comme METS, il est très adapté pour décrire les relations méréologiques, mais il est assez pauvre dans sa capacité à structurer et normaliser la description du contenu[29], et sa faiblesse descriptive n'est pas compensée par la possibilité d'inclure des formats plus riches (ce n'est pas un « format conteneur »). Conçu pour structurer des inventaires complets de fonds d'archives, ses forces et faiblesses correspondent aux nécessités de hiérarchiser des informations sur des ressources généralement non publiées et qui ont des rapports d'inclusion entre elles (fonds, carton, liasse, document).

Un format comme la TEI permet de décrire spécifiquement l'objet-texte et les objets patrimoniaux qui le portent, notamment les manuscrits. Il partage avec l'EAD l'excellence dans son champ d'application, la capacité à établir des descriptions hiérarchisées, et permet en outre la description normalisée des caractéristiques physiques (support, mise en page, etc.) ainsi que des liens précis entre le texte et une image numérisée de la ressource d'origine, voire une partie de cette image.

Des formats à structuration hiérarchisée linéaire, tels que l'EAD et la TEI sont très utiles et pratiques pour gérer la granularité de l'information et les héritages, tandis que les formats plus souples et à capacité de structuration plus grande comme METS impliquent des mécanismes plus

[28] Voir ci-dessous § 2.B.b) Profils d'application et guides des bonnes pratiques, p. ###.
[29] Voir ci-dessous § 2.A.c) Capacité descriptive des métadonnées, p. ###.

complexes pour gérer des inférences. ONIX est dans le même cas, avec une structure qui lui permet de décrire précisément les différentes parties d'un tout[30].

Tous présentent l'avantage d'avoir une information structurée plutôt qu'un ensemble mis à plat, qui fait porter tout le poids de la gestion des données et, aussi, de la navigation sur les applications et les surcouches de la bibliothèque numérique, en nuisant à la pérennisation de la bibliothèque numérique.

CAPACITE DESCRIPTIVE DES METADONNEES

Une fois décrite la structure de l'information et les modes de lecture et de navigation dans l'information (tabulaire ou linéaire), l'information descriptive de la ressource s'exprime, elle aussi, selon différents formats qui ont une plus ou moins grande capacité descriptive. Un « format conteneur » tel que METS gère particulièrement bien la « granularité » du document et de l'information, mais n'a de capacité descriptive qu'en fonction des formats qui sont utilisés en lien avec lui[31].

Tandis qu'il est merveilleusement adapté aux fonds d'archives, l'EAD se révèle, sous l'aspect de sa capacité descriptive, toujours insuffisant pour d'autres ressources. Ainsi il ne permet pas de définir de façon standardisée un lieu ou une date d'édition ou un éditeur, ni un ISBN ou un ISSN. En effet, l'élément <publisher> correspond à la publication de l'inventaire lui-même et ne peut être utilisé pour indiquer l'éditeur de pièces conservées dans le fonds. La description physique elle-même est encadrée par des éléments assez rudimentaires : l'élément <physdesc> et son fils <physfacet> ne permettent pas d'indiquer de façon standardisée ce que l'on va décrire (reliure, filigrane, matière, illustration) et reportent l'interopérabilité au niveau des bonnes pratiques. La tendance qui s'est fait jour d'utiliser l'EAD pour autre chose que des inventaires d'archives (par exemple pour des manuscrits) est proprement française et a une origine purement pragmatique qui a fait primer la facilité de traitement sur la qualité des métadonnées.

En tout état de cause, l'emploi de l'EAD est à proscrire absolument pour les livres imprimés[32], et, théoriquement, à déconseiller pour les livres manuscrits pour lesquels TEI ou les formats MARC et MODS sont plus adaptés[33], même si la mise en place de « profils d'application » permet de remédier partiellement aux insuffisances du format et, dans le cadre français, de favoriser l'interopérabilité[34].

---

[30] Ce format n'a pas vocation à être utilisé par les bibliothèques. Très riche, il présente notamment la particularité de rapatrier dans la notice bibliographique des éléments qui doivent lui rester extérieur (des informations concernant les intervenants comme la biobibliographie des auteurs ou le site web de l'éditeur commercial).

[31] La « capacité descriptive » est aussi parfois dite « granularité descriptive », en anglais « *granularity of elements* » (cf. http://conference.ifla.org/sites/default/files/files/papers/wlic2012/80-han-en.pdf). Il faut veiller à éviter les confusions entre la granularité descriptive et la granularité des documents et de l'accès.

[32] On a pu trouver ainsi des livres imprimés décrits en EAD, où l'élément de description matérielle contient les informations de publication (par exemple, « 1 volume, XIV-83 p., Marseille, Archives départementales, 1959 »), qui n'est ainsi non seulement plus interrogeable en tant que tel, mais surtout constitue un énoncé incompréhensible pour le lecteur.

[33] Voir ci-dessous § 2.B. ###usages et interopérabilité des formats de métadonnées, p. ###.

[34] La BnF décrit les manuscrits aussi bien en EAD dans le catalogue BnF Archives et manuscrits (une instance EAD par manuscrit) qu'en format MARC dans le Catalogue général. Le Catalogue général des manuscrits a été rétroconverti en EAD selon des structures différentes (une instance par bibliothèque). La présence conjointe d'archives et de manuscrits dans certains fonds est l'une des raisons de ce choix.

Le format le plus adapté et le plus riche dans sa capacité descriptive est le format MODS, issu de la longue expérience des bibliothèques et de la fine structuration des formats MARC[35]. Il permet non seulement de renseigner de façon standardisée toutes les informations de nature bibliographique, mais également de traiter adéquatement les relations entre les documents originaux et leur version numérisée.

Le format Dublin Core est générique et simple. C'est à la fois sa grande force et sa faiblesse. Il ne dispose que de 15 éléments descriptifs, tous facultatifs et répétables, contre environ deux cents éléments de MODS ; et la description est plate, sans structure interne. Ainsi, tandis que MODS divise le titre en un titre propre et un complément de titre, Dublin Core ne dispose, comme l'EAD, que d'un unique élément titre. En revanche, même dans sa version « simple », il introduit des éléments utiles aux bibliothèques numériques et absents de formats plus structurés, mais à capacité descriptive moindre, comme l'EAD. En particulier, ses éléments « Droits » (rights), « Source » (source) et « Identifiant » (identifier) font place à des nécessités de gestion pragmatique des documents en ligne et prévoient la place pour des informations capitales comme les identifiants. En outre, la définition de types de documents (« type ») et la distinction entre un « Sujet » (« subject ») et une « Couverture » (« coverage ») peut préfigurer des navigations par facettes ou par interfaces graphiques (carte ou frise chronologique).

Des formats spécifiques existent pour répondre à chaque besoin des bibliothèques numériques, comme la gestion des données de pérennisation (PREMIS[36]) ou encore pour aligner le texte et l'image : outre TEI, format polyvalent, il faut mentionner ici le format ALTO, qui permet de stocker de façon standardisée les résultats de l'OCR ainsi que les évaluations de leur qualité. Ce sont des métadonnées qui servent aussi bien à signaler les documents – dans la mesure où le texte d'un document est l'une des formes de sa description – qu'à offrir des services et fonctionnalités aux usagers (repérage des mots sur la page, synthèse vocale, reCAPTCHA, etc.)

Comparaison des capacités descriptives des formats bibliographiques pour les titres (formalisme XML)[37]

Dublin Core (possibilité 1)

<title>The subtle knife</title>

<relation>His dark materials</relation>

Dublin Core (possibilité 2)

[35] Le format ONIX, utilisé dans le monde de l'édition, est encore plus riche, avec près de 400 éléments dans la version 3 (http://www.editeur.org/8/ONIX/) ; à l'heure actuelle, il dépasse de loin les besoins et la maîtrise de la plupart des bibliothèques, notamment pour leur versant numérique, et ne répond pas au besoin d'un format unifié pour tous les types de documents.

[36] Voir chapitre sur la préservation.

[37] EAD n'étant pas fait pour décrire des livres imprimés et, *a fortiori*, une monographie dans un ensemble en plusieurs volumes, nous ne l'indiquons pas ici. Pour décrire un titre, l'EAD dispose de <titleproper> et <subtitle>, mais seulement pour le titre général de l'instrument de recherche, pas pour ses composants.

```
<title>His dark materials, 2: The subtle knife</title>
```

MARCXML

```
<datafield tag="245" ind1="1" ind2="4">
    <subfield code="a">The subtle knife /</subfield>
    <subfield code="c">by Philip Pullman.</subfield>
</datafield>
<datafield tag="490" ind1="0" ind2=" ">
    <subfield code="a">His dark materials ;</subfield>
    <subfield code="v">bk. 2</subfield>
</datafield>
```

UNIMARC en MarcXchange (exemple avec notice propre d'une monographie en plusieurs volumes plutôt qu'un dépouillement dans le champ 327)

```
<datafield tag="200" ind1="1" ind2=" ">
    <subfield code="a">The |subtle knife</subfield>
    <subfield code="b">Texte imprimé</subfield>
    <subfield code="f">Philip Pullman</subfield>
</datafield>
<datafield tag="225" ind1="1" ind2="9">
    <subfield code="a">His dark materials</subfield>
    <subfield code="v">2</subfield>
</datafield>
<datafield tag="461" ind1=" " ind2="0">
    <subfield code="t">His dark materials</subfield>
    <subfield code="v">2</subfield>
</datafield>
```

MODS[38]

```
<titleInfo>
    <nonSort>The </nonSort>
    <title>subtle knife</title>
</titleInfo>
<relatedItem type="series">
<relatedItem type="series">
    <titleInfo>
        <title>His dark materials</title>
        <partNumber>2</partNumber>
    </titleInfo>
</relatedItem>
```

ONIX[39]

```
<TitleDetail>
    <TitleType>01</TitleType>
    <TitleElement>
        <TitleElementLevel>01</TitleElementLevel>
        <TitleText textcase="02">His Dark Materials</TitleText>
    </TitleElement>
</TitleDetail>
[…]
```

[38] Influencée par Marc21, la Bibliothèque du Congrès propose <relatedItem type="series"><titleInfo><title>His dark materials ; bk. 2</title></titleInfo></relatedItem>, mais la séparation en éléments plutôt que par la ponctuation est toujours préférable. La transcription des mentions de responsabilité n'apparaît plus en tant que telle dans MODS.

[39] Dans cet exemple, la logique descriptive est inversée, car la réalité éditoriale et commerciale est la monographie en plusieurs volumes. Les éléments <TextItem> sont un pas en direction de la FRBRisation avec le raccrochement à des identifiants d'œuvres textuelles.

Pauline Moirez et Stutzmann, Dominique. « Signaler les ressources numérisées : enrichissement, visibilité, dissémination ». In *Manuel de constitution de bibliothèques numériques*, édité par Isabelle Westeel et Thierry Claerr, Bibliothèques. Electre-Cercle de la Librairie, 2013, p. 115-171 (version auteur).

```
<ContentItem>
        <LevelSequenceNumber>2</LevelSequenceNumber>
        <TextItem>
                <TextItemType>01</TextItemType>
                <TextItemIdentifier>
                        <TextItemIDType>03</TextItemIDtype>
                        <IDValue>9780375823466</IDValue>
                </TextItemIdentifier>
                <TextItemIdentifier>
                <TextItemIDType>15</TextItemIDtype>
                        <IDValue>9780375823466</IDValue>
                        </TextItemIdentifier>
        </TextItem>
        <ComponentTypeName>Book</ComponentTypeName>
        <ComponentNumber>Two</ComponentNumber>
        <TitleDetail>
                <TitleType>01</TitleType>
                <TitleElement>
                        <TitleElementLevel>04</TitleElementLevel>
                        <TitlePrefix textcase="02">The</TitlePrefix>
                        <TitleWithoutPrefix textcase="02">Subtle Knife</TitleWithoutPrefix>
                </TitleElement>
        </TitleDetail>
</ContentItem>
```

TEI[40]

```
<titleStmt>
        <title level="s" type="main">His dark materials</title>
        <title level="m" type="full">The <title type="main">subtle knife</title></title>
        <respStmt>
                <resp key="aut"/>
                <persName>
                        <forename>Philip</forename><surname>Pullman</surname>
                </persName>
        </respStmt>
</titleStmt>
```

ASSOCIATION AVEC DES DONNEES D'AUTORITE ET D'AUTRES FORMATS

Pour évaluer les formats de métadonnées et voir s'ils correspondent aux objectifs de service d'une bibliothèque numérique, le lien entre un format descriptif et des formats d'autorité ou d'autres formats doit être considéré comme un avantage. Les formats MARC, par exemple, emportent avec eux des liens structurels avec des référentiels, ainsi que la possibilité de décrire de façon cohérente et simultanée les autorités, qu'il s'agisse de personnes physiques ou morales, de lieux ou de sujets. Dans ces formats qui peuvent contenir à la fois des informations bibliographiques et des données d'autorité, la logique de structuration est identique pour l'ensemble des données, et les logiciels de gestion et de publication sont souvent communs, ce qui les rend d'un emploi plus aisé. Héritier des formats bibliothéconomiques, MODS est non seulement riche de nombreux référentiels, mais a en outre aussi un pendant en le format MADS qui permet de structurer les métadonnées d'autorité. Aussi MODS bénéficie-t-il de potentialités multiples, permettant d'associer des ressources à leurs auteurs et autres types de responsables intellectuels.

[40] La logique du format TEI pour une œuvre est de servir de format d'encodage et non de format de description. Aussi aurait-on les informations bibliographiques dans l'en-tête et dans la page de titre. L'exemple donné ici correspondrait à la description de l'en-tête. L'élément titre se subdivise avec des attributs pour indiquer le niveau (périodique, monographie, chapitre,etc.), les sous-titres, titres parallèles, etc.

Si la *Dublin Core Metadata Initiative*, l'agence de maintenance du Dublin Core, encourage bien l'utilisation de référentiels (typologie des documents, classification Dewey, date selon la recommandation du W3C, langues en iso-639), le format lui-même n'est strictement associé à aucun format de structuration des données d'autorité. Les deux chaînes de caractères « Victor Hugo » et « Hugo Victor » sont également possibles pour désigner un « Auteur » et l'absence de structuration plus fine dans le format descriptif, d'une part, ainsi que l'absence de format pour les données d'autorité, d'autre part, empêchent de préciser l'identité des acteurs.

Le format EAD bénéficie, comme MODS, de la présence d'un format associé : EAC-CPF. Celui-ci partage du reste largement les défauts de l'EAD (confusion de l'information et de sa mise en forme, multiplicité des solutions pour signaler une même information, telle que l'association d'une période et d'un lieu d'activité, soit en <place>, soit en <chronList>).

Pour choisir les formats descriptifs à utiliser dans une bibliothèque numérique, il faut en évaluer la qualité et la pertinence en fonction des besoins de la bibliothèque. Les paragraphes précédents ont permis de voir les critères qui fondent la qualité des formats de métadonnées, et, partant, du signalement possible par les bibliothèques : la justesse des informations, la cohérence dans l'usage du format, la capacité de structuration de celui-ci, sa capacité descriptive, et ses liens avec d'autres formats, notamment pour les données d'autorité.

Plusieurs formats peuvent être utilisés conjointement. Ce choix doit évidemment tenir compte également des impératifs techniques (construction des systèmes d'informations, logiciels disponibles) et des objectifs d'interopérabilité.

B. USAGES ET INTEROPERABILITE DES FORMATS DE METADONNEES

L'impératif de stabilité des infrastructures et systèmes d'information ainsi que l'objectif d'interopérabilité doit particulièrement entrer en compte dans le choix des formats de métadonnées, car il est toujours plus simple de faire interopérer des données qui sont structurées de la même façon que de les convertir. Cette réalité, qui a été au fondement même des efforts de normalisation des bibliothèques, demeure dans l'univers numérique, quand bien même les conversions sont plus aisées[41] ou que les technologies du web sémantique permettent de passer outre les différences de format[42].

POPULARITE

(Presque) indépendamment de leurs caractéristiques propres, les formats de métadonnées sont également caractérisés par leur popularité et leurs usages normalisés qui conditionnent le domaine d'interopérabilité. En effet un format, à l'instar d'un logiciel, s'évalue aussi à sa diffusion et à communauté d'utilisateurs : un logiciel très répandu présente des avantages car il évite des conversions ou des incompatibilités de format, même si, par ailleurs, il présente d'autres défauts. Une large dissémination et des utilisateurs actifs assurent que le format – toujours à comme un logiciel – ne deviendra pas obsolète, ni en lui-même (absence de prise en compte de nouvelles

[41] Voir ci-dessous § 2.C. Le cycle de vie des métadonnées : continuum descriptif et conversions, p. ###.
[42] Voir ci-dessous § 4.C. Bibliothèques numériques et web de données, p. ###.

formes de ressources), ni dans les outils associés (logiciels, protocoles d'échanges de données, documentation, autres formats).

Ainsi, les formats MARC ont pu intégrer les nouveautés dues à l'émergence des documents numériques car la communauté d'utilisateurs est importante, bien structurée, et par nature attentive à l'émergence de nouvelles ressources, et car elle présente ses propres instances de normalisation internationale. Aussi peut-on être certain que ces formats continueront d'être maintenus, ou que les outils d'une migration seront créés pour les utilisateurs s'il est décidé de migrer vers de nouveaux formats, comme cela a déjà été le cas avec la création des formats MARCXML et MODS, et sera sans doute le cas avec RDA[43].

Les usages du format constituent aussi l'une des forces du Dublin Core, dont la faiblesse de structuration et de capacité descriptive pourrait sembler rédhibitoire aux professionnels de l'information bibliothéconomique. Pourtant les avantages et inconvénients se compensent, car sa simplicité, qui le rend généralement insuffisant pour les professionnels des bibliothèques, est aussi ce qui en permet l'utilisation par de très nombreux utilisateurs hors du monde des bibliothèques, et qui en assure une adoption à la fois rapide et aisée. C'est grâce à sa simplicité que les webmestres ou les chercheurs sont en mesure de le mettre en œuvre aussi bien que les musées, les archives, les universités ou les gestionnaires de bases de données. C'est un format qui est très répandu et utilisé dans de très nombreuses circonstances. Il est même obligatoire à l'intérieur d'autres formats, tel que l'Epub, dont l'un des fichiers contient obligatoirement la description du livre au format Dublin Core. Il est aussi un format obligatoire dans le cadre des protocoles d'échanges de métadonnées, en particulier le protocole OAI-PMH (*Open Archive Initiative- Protocol for Metadata Harvesting*)[44], et il est largement utilisé dans le cadre du web de données[45]. Que ce format soit extrêmement répandu lui assure un intérêt qui est indépendant de ses qualités intrinsèques : toutes les bibliothèques numériques au monde, ou presque, peuvent s'enrichir et exporter des données en Dublin Core, mais ne peuvent pas toujours coopérer sur la base d'autres formats. Dublin Core bénéficie aussi d'une agence de maintenance active, la Dublin Core Metadata Initiative, dont les groupes d'intérêt et les groupes de travail permettent de formaliser l'emploi de Dublin Core dans des contextes toujours plus variés — notamment avec les profils d'application — et de façon à atteindre toujours l'objectif principal qui est l'interopérabilité. L'emploi de ce format pauvre n'invalide en aucun cas les règles internationales de description ni l'intérêt d'avoir des données plus richement structurées ou informées ; en revanche il s'impose pour certains usages de dissémination d'information de bas niveau et à grande échelle.

PROFILS D'APPLICATION ET GUIDES DES BONNES PRATIQUES

Des formats à la fois riches et populaires peuvent faire l'objet d'emplois variés et, *in fine*, difficilement interopérables. Pour obvier à cette difficulté, des réponses de deux natures sont apportées.

La première intervient au niveau des utilisateurs et consiste en la publication de « guides des bonnes pratiques » (en anglais : *best practice*) ou de documents qui servent de modèles, qui

---

[43] RDA (*Ressources : Description et Accès)*, code de catalogage destiné à remplacer les Règles de catalogage anglo-américaines (AACR2), se présente comme « une nouvelle norme pour la description des ressources et les accès, conçue pour le monde numérique », voir http://www.bnf.fr/fr/professionnels/rda/s.rda_objectifs.html.

[44] Voir ci-dessous annexe 1 Le protocole OAI-PMH, p. ###.

[45] Voir ci-dessous §4.C . Bibliothèques numériques et web de données, p. ###.

permettent à un établissement d'expliciter ses choix pour les cas complexes, et d'inciter ses partenaires à les suivre. La Bibliothèque nationale de France (BnF) a, par exemple, publié un Guide du Dublin Core[46] pour expliquer comment ses données sont structurées, en particulier dans les entrepôts OAI-PMH. Malgré leur nom, certains de ces guides font des préconisations qui sont tout à fait contraires au format dont ils tendent à propager l'utilisation.

La seconde réponse est la production de « profils d'application » : ceux-ci réglementent et restreignent l'usage possible des formats, en imposant une structure plus rigide ou l'emploi de référentiels qui ne sont que facultatifs dans le format d'origine ; parfois les « profils d'application » enrichissent le format en lui adjoignant des éléments exogènes.

Ces profils peuvent être directement créés par des agences de normalisation ou des agences de maintenance du format, tel que cela se fait dans le cadre de la *Dublin Core Metadata Initiative*, dont les groupes de travail ont par exemple créé un « profil d'application » de Dublin Core pour les bibliothèques[47]. Celui-ci rend obligatoire la présence d'un titre (éventuellement forgé) et, quand il y lieu, l'indication des identifiants, langue et sujet ; il préconise par exemple également l'emploi de référentiels sujets tels que la CDU, la classification Dewey, ou les LCSH (*Library of Congress Subject Headings*).

Surtout, les profils d'application peuvent enrichir un format d'éléments provenant d'autres formats : ainsi le profil d'application du Dublin Core pour les bibliothèques ajoute l'élément <edition> de MODS pour indiquer s'il s'agit d'une réédition et la caractériser.

Parfois, ce sont les utilisateurs eux-mêmes qui se créent un profil d'application afin d'assurer un emploi homogène et stable du format. Des formats riches et malléables tels que METS font l'objet de plusieurs « profils d'application » selon leur domaine d'utilisation. Ainsi la BnF a créé un profil d'application générique pour l'emploi de METS dans le cadre de l'archivage et de la pérennisation à long terme des données numériques[48]. Ce profil peut être suivi par toute bibliothèque désireuse de s'en inspirer, mais il ne s'agit cependant pas d'un document normatif.

Selon les coopérations envisagées, chaque bibliothèque numérique est libre de faire usage des guides de bonnes pratiques et des profils d'application qui lui permettent d'accomplir son projet de service, en mutualisant au mieux les développements informatiques.

C. LE CYCLE DE VIE DES METADONNEES : CONTINUUM DESCRIPTIF ET CONVERSIONS

Les métadonnées descriptives dont il est ici question forment une description des documents à une échelle inférieure à 1, c'est-à-dire que leur teneur est toujours inférieure à celle du document décrit ; et la description tend vers l'échelle 1 quand le « plein texte » est joint à la description, comme dans de nombreuses bibliothèques numériques. Il faut pourtant bien noter que les catégorisations et l'indexation sujet constituent un ajout d'information qui ne se trouve pas à proprement parler dans le document, mais qui résulte de son analyse intellectuelle.

---

[46] Guide d'utilisation du Dublin Core (DC) à la BnF
http://www.bnf.fr/documents/guide_dublin_core_bnf_2008.pdf

[47] Dublin Core – *Library Application Profile* (DC-Lib) http://dublincore.org/documents/library-application-profile/.

[48] Voir le chapitre sur la preservation. Voir aussi l'index des profils METS enregistrés http://www.loc.gov/standards/mets/mets-registered-profiles.html.

L'ensemble des métadonnées descriptives s'inscrit donc dans un espace à plusieurs dimensions : description de l'information portée par le document (titre, nombre de pages, etc.), analyse intellectuelle (formation d'un titre uniforme ou de forme, indexation, etc.) et rattachement à des référentiels indépendants (par exemple, identification de l'auteur et rattachement aux données d'autorité). Dans ces trois directions, il appartient toujours à la bibliothèque de décider jusqu'où aller. Les différents formats et leurs qualités décrites en dans les paragraphes précédents, permettent de formaliser ce choix. Pourtant, notamment avec l'introduction de profils d'application dans Dublin Core et la possibilité d'utiliser conjointement des descriptions dans différents formats avec l'emploi de formats conteneurs comme METS, il ne faut plus voir l'espace descriptif comme essentiellement discret, au contraire, il existe un continuum des capacités descriptives.

Une fois constituée une description, elle peut être formulée dans différents formats. Une même information sera structurée de façon plus ou moins riche et cette richesse correspondra à l'adéquation de la description au format choisi. L'information contenue dans un format descriptif n'est pas enfermée dans ce format : elle peut toujours être convertie vers un autre format, de même que les différents formats nationaux peuvent être convertis vers l'UNIMARC.

Pourtant, il faut retenir qu'il est plus aisé de dégrader une information que de la qualifier et de l'enrichir. L'une des voies pour transformer une donnée et l'affaiblir est la suppression de l'information propre à la capacité descriptive d'un format.

Prenons un premier exemple, les mentions du lieu d'édition et de la date :

- En INTERMARC, on aura deux champs
  - 260 $aParis$cB. Grasset$dimpr. 2007
  - 270 $a18-Saint-Amand-Montrond$cImpr. Bussière
- En UNIMARC, un seul champ, mais la même teneur et une structuration de même qualité :
  - 210 $aParis$cB. Grasset$dimpr. 2007$e18-Saint-Amand-Montrond$gImpr. Bussière
- En revanche, en Dublin Core, le champ « éditeur » (*publisher*) désigne l'« entité responsable de la mise à disposition de la ressource ». On aura donc, normalement, l'information suivante :
  - <dc:publisher>B. Grasset</dc:publisher>

La BnF adopte la forme <dc:publisher>B. Grasset (Paris)</dc:publisher> à la fois pour conserver l'information du lieu d'édition et pour assurer l'identification d'éditeurs homonymes. Beaucoup de bibliothèques, en créant une incohérence entre un format et des normes qui ne lui sont pas applicables, telles que l'ISBD, utilisent la forme <dc:publisher>Paris : B. Grasset</dc:publisher>[49]. En tout cas, les indications de l'imprimeur et des lieux tant d'édition que d'impression, n'ont pas de place dans une description en Dublin Core, et l'impression ne trouve pas sa place non plus en MODS. Et cet exemple montre bien qu'il est possible de passer

[49] Voir http://www.loc.gov/standards/mods/userguide/origininfo.html.

d'un format riche à un format pauvre, alors que l'inverse n'est pas vrai sans avoir à *qualifier* les données, c'est-à-dire à les enrichir avec des informations extérieures.

À l'heure actuelle, l'adoption du modèle FRBR pose un problème similaire. Il est aisé de créer une notice UNIMARC à partir de descriptions précises des Œuvres, Expressions, Manifestations et Documents, en revanche, la création des différentes Œuvres et Expressions à partir des notices actuelles ne pourra pas être effectuée correctement sans un apport intellectuel nouveau, donc une qualification des données, ainsi qu'on le voit parfaitement dans l'historique des Œuvres de la Bibliothèque nationale d'Australie[50].

NOTICE ET INFORMATION

On peut passer d'un formalisme à un autre sans perte d'information (Dublin Core en format tabulaire ou Dublin Core en XML, UNIMARC en iso2309 ou MARCXML). Il est plus ardu de passer d'un format à un autre, c'est-à-dire d'une structure d'information à une autre : cette conversion intervient généralement avec perte ou ajout nécessaire d'information. La structure d'une notice dans un format prédéterminé contraint les informations disponibles, aussi bien en incitant le catalogueur à renseigner des champs dont il n'aurait pas conservé l'information dans d'autres formats, qu'en l'empêchant de les saisir dans d'autres.

Les promesses du web sémantique et de modèles comme RDF (*Resource Description Framework*) sont d'obtenir une interopérabilité de haut niveau en sortant d'un modèle arborescent où la notice constitue une unité insécable, pour passer à une fragmentation ultime des métadonnées en un ensemble d'informations élémentaires que l'on peut mettre en réseau, comme dans le web de données[51]. Ce faisant, l'on passe théoriquement de la problématique de format, axée sur la structuration des informations, à celle de modèle, fondée sur l'explicitation des rapports logiques : toute information peut être conservée avec l'indication précise de sa signification et peut être exploitée dans d'autres contextes grâce aux rapprochements, alignements et ontologies.

Les technologies du web sémantique abolissent ainsi en théorie le besoin de conversion. L'agence de maintenance d'un format doit en effet « déclarer » celui-ci et peut créer une ontologie qui établit les « alignements » ou points de correspondance avec d'autres formats. Il est, par exemple, possible de déclarer que la relation 200$f de UNIMARC-Autorité peut être équivalente aux relations DateOfBirth et DateOfDeath de DBPedia[52] pour les autorités personnes physiques. En pratique, la réalité est un peu plus complexe, car pour suivre de lien en lien les données présentées dans un ensemble externe, il faut à la fois connaître extrêmement bien le modèle des

[50] La Bibliothèque nationale d'Australie donne accès à une liste des dernières modifications intervenues sur les « œuvres » au sens que les FRBR donnent à ce mot. La tâche à accomplir se révèle gigantesque, cf. http://trove.nla.gov.au/recentMergesOrSplits.
[51] Voir ci-dessous § 4.C. Bibliothèques numériques et web de données, p. ###.
[52] Le projet DBpedia réalise une extraction des données de Wikipédia pour les publier sur le web de données.

données que l'on explore et disposer des capacités informatiques (logicielles ou de programmation) pour automatiser cette exploration.

L'évaluation et le choix du format de métadonnées qui sera utilisé par la bibliothèque numérique, du côté de l'infrastructure ou auprès de l'utilisateur final, dépendent des objectifs de service de la bibliothèque. Plusieurs formats peuvent être utilisés conjointement, aussi bien pour gérer la différence entre les formats de préservation et ceux de présentation ou d'exportation que dans le cas d'un format conteneur comme METS. Chaque format a sa logique et ses capacités de structuration et de description propres qui sont à considérer au regard de sa popularité, de sa facilité d'emploi, et des promesses d'interopérabilité, sachant que les conversions sont toujours possibles, mais également délicates pour éviter la perte d'information. De façon générale, et dans la mesure du possible, on préférera toujours un format riche et bien structuré permettant une réutilisation optimale, car la dissémination et l'utilisation potentielle des métadonnées hors de la bibliothèque numérique d'origine doivent toujours être envisagées.

3. PLURALITE DES BASES ET DES LIEUX D'UTILISATION

Le document numérique et ses métadonnées ont tous deux une nature ubiquitaire, c'est-à-dire qu'ils peuvent être présents en un ou plusieurs lieux en même temps, et exploités différemment par des systèmes variés. Dans le cadre du présent manuel, il ne nous appartient pas de préconiser une solution en particulier, mais il faut insister sur les potentialités et les difficultés de la gestion du document numérique et des métadonnées dans leur nature ubiquitaire, en notant que les métadonnées peuvent être créées par la bibliothèque, ou importées, ou simplement affichées.

Rares sont les bibliothèques qui adoptent délibérément et explicitement des stratégies d'informatique dans les nuages (*cloud computing*) pour les métadonnées, même quand les documents numérisés eux-mêmes sont hébergés par des prestataires. Le processus de dérivation de notice et l'exemple du Sudoc, qui assure une synchronisation régulière des métadonnées et permet, *in fine*, à chaque bibliothèque de disposer de sa base locale, même si elle est en lien avec une base externe, reste le modèle fondamental de la catalographie en France. Pourtant les architectures évoluent. D'un côté, l'ABES développe un projet de « système de gestion des bibliothèques mutualisé » (SGBm), qui serait un catalogue partagé dans les nuages et avec des données ouvertes[53]. De l'autre, les ressources numériques en accès payant et la numérisation modifient profondément les catalogues, car la construction d'une bibliothèque numérique ou l'abonnement à des ressources complémentaires engendre généralement l'existence de bases spécifiques, qu'il s'agit de maîtriser au mieux et de faire dialoguer avec le catalogue des ressources présentes sur place.

[53] Voir http://sgbm.abes.fr.

A. GERER SES BASES EN INTERNE

Pour des raisons qui tiennent au développement des techniques ou à la nature variée des documents conservés, de nombreuses institutions porteuses de bibliothèques numériques alimentent et gèrent plusieurs bases de données différentes. La constitution d'une bibliothèque numérique amène une complexité supplémentaire et conduit généralement à créer au moins une base spécifique supplémentaire.

La BnF dispose, par exemple, de nombreux catalogues de documents :

- le « Catalogue général » pour les documents imprimés, cartographiques, audiovisuels, une partie des documents non publiés, etc.,
- « BnF Archives et Manuscrits » pour les documents non publiés de certains départements,
- le catalogue du Centre national de la littérature pour la jeunesse - « La Joie par les Livres »,
- Gallica, pour les documents numérisés libres de droit, ainsi que ceux des bibliothèques partenaires et les documents sous droits des éditeurs, et son pendant Gallica IntraMuros qui permet, dans l'enceinte de la bibliothèque, de consulter également des ressources non libres de droit,
- Mandragore, pour les miniatures des manuscrits,
- la Banque d'image, pour les documents disponibles auprès des services de la Reproduction,
- BnF-Ressources électroniques, qui donne accès aux périodiques et livres électroniques ainsi qu'aux bases de données,
- les archives de l'Internet (*WayBack Machine*),
- les Signets de la BnF, sélection commentée de ressources web,
- les expositions virtuelles et dossiers pédagogiques,
- les conférences en ligne.

Il s'agit, bien évidemment, d'un cas extrême. Pourtant il ne s'agit nullement d'un cas isolé. Les bibliothèques universitaires ou municipales sont nombreuses dont une partie du fonds est décrit dans une base séparée, soit pour les ressources les plus anciennes (ne serait-ce que les manuscrits dont la description du Catalogue général des manuscrits est accessible via le Catalogue collectif de France ou via la base Calames), soit pour les ressources les plus récentes (abonnement à des périodiques en ligne ou accès distant à des bases de données documentaires).

Nous traiterons plus loin des avantages d'une architecture éclatée, où les ressources sont mises en ligne selon leur typologie et, le cas échéant, avec des offres de service d'hébergement par des prestataires ou des services gratuits, tels que Flickr[54].

L'existence de plusieurs bases entraîne des lourdeurs techniques, car il faut des techniciens aptes à gérer chacune des bases, leurs différents formats et leur articulation. La multiplication des

[54] Voir ci-dessous. § 3.B.a) Enrichissement social et *crowdsourcing*, p. ###.

bases est également pénalisante pour l'usager qui doit démultiplier sa recherche. Pour gérer l'ensemble des métadonnées, plusieurs solutions sont envisageables.

Tout d'abord, la fusion des différentes bases. L'avantage est évident : c'est la simplicité de gestion et la fluidité de navigation qui doit en découler. C'est le principe qui présidait à la création des SIGB (systèmes « intégrés ») et qui motive les systèmes de GED (« gestion électronique des documents »). L'inconvénient est que cette fusion impose aux données une structuration selon un même format, de sorte que des opérations de conversions ou de qualification des données s'avèrent indispensables, qui se font soit au détriment des spécificités de catalogage, soit à celui de l'uniformité du catalogue. Pourtant, bien menée, une telle unification est sans aucun doute la solution à préférer et à mener pour les ressources de même nature, sans préjuger de la fusion des interfaces de catalogage et de consultation.

Elle n'est cependant pas toujours possible : intégrer des ressources électroniques externes et leur plein texte demande parfois des passerelles techniques que d'anciens SIGB n'offrent pas. Dans ce cas, il est absolument indispensable d'offrir aux utilisateurs un accès unifié aux données, même si celles-ci sont gérées par plusieurs systèmes différents. Aussi faut-il impérativement mettre en place un moteur de recherche fédérée, qui permette aux internautes de découvrir les ressources accessibles. Chaque ressource doit à la fois pouvoir y être présentée de façon autonome et demeurer rattachée à son contexte intellectuel ou de conservation, pour les fonds où cela se justifie. La bibliothèque numérique doit constituer, pour les documents numérisés, cette interface de recherche fédérée. Celle-ci doit être soigneusement préparée et paramétrée, tant du point de vue technique que de l'analyse intellectuelle et de l'élaboration des critères de pertinence.

data.bnf.fr : un « pivot documentaire » pour le signalement des ressources de la BnF

Mis en ligne en 2011, le projet data.bnf.fr[55] a pour objectif de rendre les ressources de la BnF plus accessibles et utilisables par les usagers, en s'appuyant sur les technologies du web sémantique et sur la FRBRisation des données.

La BnF propose en effet différents catalogues et ressources adaptés aux besoins spécifiques des données diffusées en termes d'interfaces comme de formats : en particulier le catalogue général pour les notices bibliographiques et d'autorité (format MARC), le catalogue BnF-Archives et manuscrits (format EAD), la bibliothèque numérique Gallica (notices descriptives en format Dublin Core), ou encore des sites web multimédia pour les expositions virtuelles.

Ces bases et ressources multiples, qui constituent autant de silos de données indépendants, sont à la fois complexes et peu lisibles pour les usagers, et inaccessibles pour les moteurs de recherche. L'identification de l'information la plus pertinente et la plus qualifiée pour une

[55] Voir *Le Web de données à la BnF : data.bnf.fr* http://www.bnf.fr/fr/professionnels/modelisation_ontologies/a.web_semantique_bnf.html, et A. Simon, *De la description des documents à l'exploitation des données : data.bnf.fr* http://www.musees-mediterranee.org/pdf_breve/reseauDoc_103_4RCS.pdf.

recherche est brouillée.

C'est pourquoi data.bnf.fr regroupe et expose les données bibliographiques provenant de ces différents catalogues et ressources, en créant, dans une perspective FRBRisée, des pages sur les auteurs, les œuvres et les thèmes. Par exemple, le lecteur pourra trouver sur la page consacrée à l'*Histoire de ma vie*, de Casanova, des rebonds vers :

- les notices bibliographiques des différentes éditions de cette œuvre (issues du catalogue général),
- la description du manuscrit original (issue du catalogue BnF-Archives et manuscrits),
- les ouvrages qui traitent ou étudient cette œuvre (issues du catalogue général, en accès matière),
- les fichiers de celles de ces éditions qui sont numérisées et disponibles sur Gallica,
- le manuscrit original numérisé (disponible sur Gallica),
- et bientôt l'exposition virtuelle consacrée à Casanova.

Grâce à l'utilisation des techniques du web sémantique[56], data.bnf.fr extrait et articule les données produites dans chaque catalogue, en leur appliquant des traitements automatiques d'alignement et de regroupement. Le projet agit ainsi comme un véritable « pivot » pour les données de la BnF, à la fois en entrée en récupérant et regroupant les données des différents catalogues, mais aussi en sortie, en produisant en parallèle des pages lisibles par les usagers et des données structurées selon le modèle RDF (*Resource Description Framework*) du web sémantique qui peuvent être utilisées par les moteurs de recherche comme par des réutilisateurs potentiels, ce que les bases de données d'origine ne permettaient pas.

B. AU DELA DU SIGNALEMENT PAR LES PROFESSIONNELS, UN ENRICHISSEMENT SOCIAL ET SEMANTIQUE DES METADONNEES

L'intégration de la bibliothèque numérique dans l'écosystème du web permet d'envisager des possibilités inédites d'enrichissement des métadonnées descriptives, en s'appuyant sur la participation des internautes, mais aussi sur des techniques automatiques de traitement des données.

Ces informations « non-professionnelles » n'ont certes pas le même niveau de qualité et de fiabilité que les données produites par les bibliothécaires, et l'interface de consultation peut

[56] Voir ci-dessous § 4.C. Bibliothèques et web de données, p. ###.

rendre compte clairement de ces deux niveaux de qualité : informations fiables produites ou vérifiées par les bibliothécaires / informations complémentaires produites par les internautes ou par traitement automatique. Mais ces données permettent d'une part d'enrichir les catalogues dans des proportions que les institutions ne peuvent pas envisager seules, et également de répondre à des besoins et usages spécifiques des internautes, dont les recherches sont alors facilitées par la multiplication des mots-clefs et par l'utilisation du « langage naturel ».

Ces opérations doivent s'intégrer dans la stratégie de diffusion numérique de l'établissement, afin de prendre en compte le plus en amont possible les moyens à mettre en œuvre, tant sur le plan technique qu'humain.

A) ENRICHISSEMENT SOCIAL ET *CROWDSOURCING*

Les bibliothèques numériques s'inscrivent dans un écosystème où l'interaction est la norme : l'internaute s'attend à pouvoir intervenir sur les données et sur les contenus, que ce soit pour les commenter, les partager ou les enrichir. Même lorsqu'il n'utilise pas ces fonctionnalités[57], elles lui sont familières dans sa pratique courante du web, sur les réseaux sociaux ou les sites marchands. Elles constituent son cadre de référence, il se sentira enfermé et exclu s'il ne les a pas à sa disposition[58]. Et pourtant, l'expérience montre que l'intégration de fonctionnalités d'enrichissement collaboratif dans les catalogues ou bibliothèques numériques françaises rencontre rarement le succès escompté, et peine à atteindre la masse critique nécessaire pour améliorer notablement l'expérience de recherche des usagers[59], alors que des bibliothèques anglo-saxonnes (comme la Bibliothèque du Congrès ou la Bibliothèque nationale d'Australie[60]) ou d'autres institutions culturelles, en particulier les services d'archives[61], parviennent à mettre en place des projets particulièrement réussis.

L'enjeu des bibliothèques est donc, au-delà d'une réponse à cette attente inconsciente des internautes, de faire le meilleur usage possible de ces technologies et des contributions des

---

[57] La règle du « 1-9-90 » veut que seul 1% des internautes participe activement à l'enrichissement de contenus en ligne, 9 % y contribuent occasionnellement, et 90 % soient des consommateurs passifs (http://fr.wikipedia.org/wiki/R%C3%A8gle_du_1_%25). On assiste toutefois à une remise en cause progressive de cette règle, vers une participation accrue des internautes (jusqu'aux ¾ de contributeurs au moins occasionnels au Royaume-Uni, par exemple), voir Aref Jdey, « La règle des 90/9/1 est désormais dépassée », *Demain la veille*, 2012, http://www.demainlaveille.fr/2012/07/02/la-regle-des-9091-est-desormais-depassee/.

[58] Etienne Cavalié, « Les tags dans les OPAC : ce n'est pas parce que personne ne s'en sert que ça ne sert à rien », *Bibliothèques (reloaded)*, 2010, http://bibliotheques.wordpress.com/2010/02/19/les-tags-dans-les-opac-ce-nest-pas-parce-que-personne-ne-sen-sert-que-ca-ne-sert-a-rien/.

[59] Lionel Dujol, « Le catalogue 2.0 ou le mythe de l'usager participatif ? », *La bibliothèque apprivoisée*, 2009, http://labibapprivoisee.wordpress.com/2009/10/14/le-catalogue-2-0-ou-le-mythe-de-lusager-participatif/.
Bertrand Calenge, « Des publics utilisateurs aux publics collaborateurs : une fausse bonne idée ? », *Bertrand Calenge : carnet de notes*, 2012, http://bccn.wordpress.com/2012/02/11/des-utilisateurs-aux-collaborateurs-une-fausse-bonne-idee/.

[60] Voir ci-dessous les encarts « La Bibliothèque du Congrès sur Flickr : disséminer pour mieux tagguer… », p. ###, et « Trove, ou la correction collaborative d'OCR des périodiques de la Bibliothèque nationale d'Australie », p. ###.

[61] Voir Pauline Moirez, « Archives participatives », dans *Bibliothèques 2.0 à l'heure des médias sociaux*, dir. Muriel Amar et Véronique Mesguich, 2012, p. 187-197. Voir aussi ci-dessous l'encart « L'indexation collaborative de l'état civil », p. ###.

usagers, pour enrichir les métadonnées descriptives de leurs collections numériques et pour améliorer l'expérience de recherche et de navigation des utilisateurs. Il serait en effet dommage de n'utiliser les potentialités du web social que de façon « cosmétique », pour répondre aux codes du web, sans en faire véritablement bénéficier le signalement des collections et l'interface de recherche de la bibliothèque numérique.

Les bibliothèques peuvent en effet chercher à susciter une « participation[62] » des usagers, c'est-à-dire la mise en œuvre de véritables compétences et connaissances des usagers, une interaction de haut niveau, de caractère scientifique, qui contribue à l'enrichissement du signalement des collections numériques. La fluctuation de la terminologie désignant ces opérations participatives souligne d'ailleurs la frontière de plus en plus floue entre signalement et valorisation des collections sur le web. On parlera de « *crowdsourcing*[63] » pour désigner des projets collaboratifs de grande ampleur, mais l'accent sera davantage mis sur le nombre des participants, sur la notoriété du projet, sur la constitution de communautés de contributeurs, que sur la valeur scientifique de leurs contributions. L'expression « métadonnées sociales[64] », ou données produites par les utilisateurs, insiste quant à elle davantage sur l'enrichissement et l'amélioration de la description bibliographique.

Cette participation des usagers, qui peut exister sur de simples données bibliographiques, est renforcée par la mise en ligne des bibliothèques numériques. En effet, la mise à disposition des usagers de documents numérisés, images voire textes OCRisés, permet des opérations de *crowdsourcing* plus ambitieuses qui enrichissent notablement la description des documents : indexation, identification de photographies, correction d'OCR, ou encore transcription collaborative.

Les bibliothèques bénéficient pour la mise en place de programmes de *crowdsourcing* d'atouts significatifs. Tout d'abord, la masse, la richesse et la variété des collections numérisées (manuscrits, livres, images fixes, images animées, documents sonores) multiplient les opportunités d'expérimentations et d'interventions d'utilisateurs aux intérêts, formations et qualifications divers. De plus, pour les contenus édités, l'existence de plusieurs exemplaires d'un même document ouvre la voie à des réutilisations d'enrichissements sociaux réalisés sur d'autres exemplaires.

C'est pourquoi les **fonctionnalités de recommandation**, bien que relativement sommaires et reflets d'opinions subjectives, sont souvent utilisées par les bibliothèques. Qu'il s'agisse de l'évaluation d'un ouvrage (notation, commentaires), d'un partage sur des réseaux

---

[62] L'archiviste américaine Kate Theimer définit ainsi les « archives participatives » : « un organisme, un site ou une collection auxquels des personnes qui ne sont pas des professionnels des archives apportent leur connaissance ou ajoutent des contenus, généralement dans un contexte numérique en ligne. Il en résulte une meilleure compréhension des documents d'archives. » Kate Theimer, « The participatory archives », *Archives Next*, 2011, http://www.archivesnext.com/?p=2319.

[63] Littéralement « apport de contenu par la foule », http://fr.wikipedia.org/wiki/Crowdsourcing. Voir aussi Rose Holley, « Crowdsourcing: How and Why Should Libraries Do It?", dans *D-Lib Magazine*, 2010, vol. 16, n° 3-4, http://www.dlib.org/dlib/march10/holley/03holley.html.

[64] OCLC, Sharing and Aggregating Social Metadata, http://www.oclc.org/research/activities/aggregating/default.htm.

sociaux (par exemple via le bouton « J'aime » de Facebook) ou d'une véritable critique argumentée, il est souhaitable d'intégrer des outils de partage et de recommandation dans la bibliothèque numérique, afin de s'adapter aux usages du web et aux attentes des usagers. Toutefois, il est difficile de parvenir à la masse critique de participations nécessaires à améliorer notablement l'expérience de consultation.

En revanche, des médias sociaux spécialisés dans les échanges autour des livres, des films ou de la musique, comme Babelio, Sens Critique, LibFly ou encore LibraryThing, disposent d'un large vivier de contributeurs, dont l'intense activité de recommandation peut bénéficier à l'enrichissement du signalement des bibliothèques. Des ponts peuvent être créés entre ces médias et la bibliothèque, pour enrichir les catalogues et bibliothèques numériques avec les critiques et avis des utilisateurs de ces sites : des widgets intégrables sur une page web, ou encore l'outil Babelthèque de Babelio, utilisé en particulier par la bibliothèque municipale de Toulouse. Ce dernier permet d'une part aux usagers du catalogue de contribuer à l'enrichissement des notices (sans grand succès), mais surtout intègre les commentaires des usagers de Babelio à l'intérieur même du catalogue (sans toutefois les rendre interrogeables, ni permettre une recherche par livres bien notés, ce qui réduit l'intérêt du service pour l'utilisateur). Au final, « un outil d'enrichissement plutôt que de participation[65] », sans doute, mais qui répond parfaitement à un objectif double : s'appuyer sur les contributions des internautes et enrichir les métadonnées bibliographiques.

L'usager peut également être invité à **enrichir l'indexation** des ressources numériques, par l'ajout de mot-clefs ou « *tags* ». Ce processus d'indexation et de classification collaborative, par des mots-clés librement choisis par chaque internaute, est appelé « folksonomie[66] ». Celle-ci n'apporte évidemment pas la qualité d'une indexation professionnelle normalisée et appuyée sur des référentiels contrôlés ; elle pose même des problèmes de polysémie, d'orthographe, d'absence de hiérarchie, ou encore de personnalisation des vocabulaires. Mais le *tagging* social fournit une indexation simple, gratuite et rapide, appuyée sur une large communauté d'usagers, qui couvre potentiellement tous les domaines de la bibliothèque numérique et tous les types de documents. De plus, ces folksonomies sont conformes aux usages du web, elles s'expriment dans des vocabulaires simples et intuitifs qui correspondent aux modes de recherche en langage naturel des usagers[67].

Si le *tagging* des collections peut être intégré au sein de la bibliothèque numérique, malgré les risques liés à la faiblesse des interactions et à la difficulté d'obtenir une masse critique, les médias sociaux de partage de contenu (Flickr pour les photographies, Youtube ou Dailymotion pour les documents audiovisuels) restent le lieu privilégié pour ce type de service.

---

[65] Eymeric Manzinali, « Babelthèque à la bibliothèque de Toulouse : observations sur les OPAC 2.0. », *Le Monde du livre*, 2012, http://mondedulivre.hypotheses.org/477.

[66] Olivier Le Deuff, "Folksonomies. Les usagers indexent le web », *Bulletin des bibliothèques de France* (2006 - t. 51, n° 4), http://bbf.enssib.fr/consulter/bbf-2006-04-0066-002.

[67] Olivier Ertzscheid, *Folksonomies et indexation sociale : le monde comme catalogue*, 2008, http://fr.slideshare.net/olivier/oe-abes-mai2008.

La Bibliothèque du Congrès sur Flickr : disséminer pour mieux tagguer…

Afin d'accroître la visibilité de ses collections sur le web, de s'intégrer dans des communautés d'usages collaboratifs, et d'étudier les impacts potentiels des folksonomies sur l'enrichissement du signalement et des modes de recherche des usagers, la Bibliothèque du Congrès diffuse depuis 2008 environ 4600 photographies anciennes sur Flickr[68], alliant ainsi la dissémination des contenus et l'ouverture à la participation des usagers.

En un peu moins d'une année, ces photographies ont été vues plus de 10 millions de fois, 7000 commentaires ont été saisis, et 67 000 tags ajoutés. La fréquentation de la bibliothèque numérique a augmenté de 20 % pendant cette période. La qualité des commentaires a permis la mise à jour et l'enrichissement de 500 notices bibliographiques, tandis que les tags apportent des compléments notables à l'indexation professionnelle (par exemple, des informations géographiques, des traductions, des relevés d'objets ou de couleurs présents sur les photos).

Malgré leur pertinence en termes d'usages, ces folksonomies sont souvent bien loin de constituer un véritable travail d'indexation scientifique. Mais certaines initiatives peuvent servir de base indispensable à une description bibliographique, par exemple l'identification collaborative de photographies, comme à la bibliothèque municipale de Lyon[69] ; ou permettent d'encadrer plus étroitement l'enrichissement social pour produite de véritables données structurées utilisables dans les catalogues, par exemple une géolocalisation de cartes anciennes à la *New York Public Library*[70].

Au-delà des documents iconographiques, le tagging permet aussi d'améliorer l'accès aux documents audiovisuels, comme le montre le projet *Waisda?* de l'Institut néerlandais pour le Son et l'Image, qui propose, sous forme ludique, l'indexation collaborative des archives de la télévision, et qui a rencontré un excellent succès public (plus de 340 000 tags ajoutés pendant les 6 premiers mois)[71].

L'indexation collaborative de l'état civil : une expérience réussie

En janvier 2012, 21 services d'archives français ont développé sur leurs sites web des modules d'indexation collaborative de documents sériels à caractère nominatif (état-civil le plus souvent, mais aussi registres matricules militaires, recensements de population, etc.). Les internautes sont invités à indexer ces documents grâce à des formulaires simples et structurés qui limitent les possibilités d'erreurs (indexation de la date, des noms et prénoms, des professions exercées, etc.). La qualité de cette indexation est assurée, suivant les cas, par des

[68] *For the Common Good: The Library of Congress Flickr Pilot Project*, 2008, http://www.loc.gov/rr/print/flickr_report_final.pdf.
[69] http://collections.bm-lyon.fr/photo-rhone-alpes/.
[70] http://maps.nypl.org/warper.
[71] Maarten Brinkerink, « Waisda? Video Labeling Game: Evaluation Report », *Images for the future*, 2010, http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/.

tests de paléographie en ligne préalables, par une double indexation pour croiser les données, et par des vérifications par échantillonnage par les archivistes.

Le succès de ces initiatives repose largement sur l'animation de communautés très actives de généalogistes et amateurs éclairés, et sur une interface d'annotation ergonomique et structurée, qui permet la réutilisation des données produites. Les chiffres sont parlants : aux Archives départementales de l'Ain, 500 000 pages ont été indexées en 2 ans ; aux Archives départementales du Cantal, 1000 indexations sont réalisées chaque jour[72]…

Au-delà des métadonnées descriptives, les utilisateurs peuvent également contribuer à **enrichir le document numérique lui-même**, en corrigeant un texte préalablement OCRisé, voire en le transcrivant *ex nihilo*. Les techniques d'OCR automatique ne peuvent en effet pas obtenir des résultats complètement parfaits, seule une relecture humaine permet d'atteindre un taux de reconnaissance de 100 %. De plus, l'OCR n'est à ce jour efficace ni sur les écritures manuscrites anciennes ni sur les livres imprimés avant le XVII^{e} siècle ; là encore, seul l'œil humain permet de réaliser une transcription de ces documents, afin de disposer d'un mode texte nécessaire à la recherche plein texte, à la synthèse vocale pour les non-voyants ou encore à la réalisation de livres numériques.

Trove, ou la correction collaborative d'OCR des périodiques de la Bibliothèque nationale d'Australie

La bibliothèque numérique Trove[73] propose une stratégie globale et cohérente de *crowdsourcing* (tagging et commentaires) sur l'ensemble des collections. Le programme de correction collaborative d'OCR sur les périodiques numérisés reste toutefois l'aspect le plus innovant de l'ensemble. Mis en place depuis 2008, il propose aux internautes de participer à l'amélioration de la transcription de plus de 6 millions de pages (chiffres de mars 2012). 2 millions de lignes de texte sont ainsi corrigées chaque mois par environ 30 000 volontaires. L'intégration de ce service au cœur même de la bibliothèque numérique permet de rendre immédiatement disponibles aux internautes les enrichissements apportés.

Le succès de l'opération repose notamment sur une bonne animation de la communauté des contributeurs (valorisés par la mise en avant chaque mois des « *top correctors* »), sur une interface ergonomique et agréable, et sur l'intégration des contributions des internautes aux fonctionnalités de recherche, ce qui met en avant leur richesse et leurs apports et améliore notablement l'aisance de recherche dans les collections.

[72] Voir Edouard Bouyé, « Le web collaboratif dans les services d'archives publics : un pari sur l'intelligence et la motivation des publics », *La Gazette des Archives*, n°227 (2012-3) (à paraître).

[73] Rose Holley, *Many Hands Make Light Work: Public Collaborative OCR Text Correction in Australian Historic Newspapers*, 2009, http://www.nla.gov.au/ndp/project_details/documents/ANDP_ManyHands.pdf.

Les fonctionnalités d'enrichissement participatif des données peuvent être intégrées dans la bibliothèque numérique elle-même, permettant le regroupement de l'ensemble des métadonnées, bibliographiques et sociales, au sein d'une même interface. Elles peuvent également s'appuyer sur une dissémination des contenus numériques hors du site de la bibliothèque, sur des médias spécialisés : sites de partage (de photographies comme Flickr, de documents audiovisuels comme Dailymotion ou Youtube), sites collaboratifs (tels que la bibliothèque libre Wikisource), portails dédiés (LibraryThing, Babelio). Cette dissémination permet de bénéficier d'une mutualisation des développements techniques et ergonomiques, mais également d'une plus grande visibilité sur des sites très fréquentés, et donc d'un vivier de contributeurs plus important. Toutefois, se pose alors la question de la réintégration des métadonnées sociales à l'intérieur de la bibliothèque numérique qui, pour des raisons techniques ou de condition d'utilisation des services, n'est pas toujours simple.

La BnF a ainsi conclu en 2010 un partenariat avec Wikimedia France pour le versement dans la bibliothèque numérique Wikisource de 1416 ouvrages ouverts à la transcription collaborative des usagers du site[74]. Toutefois, le format DjVu utilisé par Wikisource ne contient pas d'informations de structure comme le format ALTO utilisé par Gallica, et les fichiers corrigés ne peuvent donc pas être réintégrés automatiquement.

Dans tous les cas, un investissement important est demandé aux internautes, en temps et/ou en compétences scientifiques, et un tel projet ne peut rencontrer le succès (forte participation et création de données de bonne qualité) sans **un investissement de la part de la bibliothèque** également :

- une communication soutenue auprès des communautés d'utilisateurs potentiels (usagers des sites collaboratifs concernés, publics physiques et virtuels de la bibliothèque, populations locales, communautés d'intérêt comme par exemple les associations historiques ou généalogiques locales, etc.), pour les sensibiliser, les recruter, les former, les mobiliser par des actions de suivi et de reconnaissance (remerciements, invitations, mise en avant des contributeurs les plus prolixes).
- une ergonomie de saisie simple et intuitive qui facilite la prise en main et les travaux de correction ou transcription. Une utilité évidente jointe à un maniement intuitif, tel pour le reCAPTCHA[75], ou une interface ludique peuvent tout à fait constituer un moyen de motivation des contributeurs, comme le montre le succès de Digitalkoot[76], service de correction collaborative d'OCR de la Bibliothèque nationale de Finlande (en un an, 101 614 visiteurs ont passé 328 376 minutes pour réaliser 6.461.659 micro-tâches de correction).

---

[74] Voir http://fr.wikisource.org/wiki/Wikisource:Partenariats/Biblioth%C3%A8que_nationale_de_France.

[75] Le reCAPTCHA (http://www.google.com/recaptcha) est un service anti-spam qui demande à l'internaute de transcrire deux mots qui lui sont soumis ; l'un est un mot test, et l'autre un mot mal reconnu par un logiciel d'OCR ; en transcrivant les deux mots, l'internaute contribue à améliorer la qualité du plein texte. Racheté par Google en 2009, cet outil est notamment utilisé pour la numérisation des archives du *New York Times*, et pour les ouvrages de *Google Books*.

[76] Nora Daly, *IMPACT Final Conference-Crowdsourcing in the Digitalkoot Project*, 2011, http://impactocr.wordpress.com/2011/10/24/impact-final-conference-crowdsourcing-in-the-digitalkoot-project/

- une stratégie de gestion de la qualité des contenus produits[77] : en croisant des doubles (ou triples) corrections, en appelant les internautes à signaler ou corriger les erreurs, en faisant vérifier les données par des experts, etc.
- une réflexion stratégique sur les collections concernées par les programmes de *crowdsourcing* que l'on veut mettre en place : suivant les moyens de l'institution, les choix techniques réalisés (hébergement sur le site de la bibliothèque ou dissémination), il pourra être judicieux de se limiter dans un premier à un type de contenu, quelques collections iconographiques, par exemple, ou bien un fonds spécifique. C'est ainsi que la bibliothèque municipale de Toulouse a disséminé dès 2008 le fonds du photographe Eugène Trutat[78] sur Flickr, afin d'augmenter sa visibilité, d'améliorer l'identification et les descriptions des photographies, mais aussi de reconstituer virtuellement un fonds divisé entre 3 institutions de conservation.

La mise en place de ces programmes collaboratifs doit s'appuyer sur une stratégie clairement définie : quel enrichissement des données, quelle implémentation dans les catalogues, quels nouveaux services offrir aux usagers ? Seule cette stratégie permettra d'éviter l'écueil d'un collaboratif « cosmétique » qui n'améliore pas véritablement la qualité et les fonctionnalités de la bibliothèque numérique et trompe finalement l'internaute qui croit contribuer à cette amélioration.

Il est ainsi souhaitable de prévoir la réintégration des contenus enrichis (tags, fichiers textuels corrigés ou transcrits) dans les catalogues et de les intégrer aux index pour qu'ils améliorent l'expérience de recherche de l'usager, que ces enrichissements collaboratifs aient été produits sur le site de la bibliothèque ou déportés sur des médias externes. Cela implique la mise à disposition de moyens techniques (processus de traitement et d'indexation des données – tags ou commentaires – créés dans l'interface d'une bibliothèque numérique, par exemple) et humains (animation de la communauté d'utilisateurs et enrichissement des métadonnées du catalogue, après analyse du contenu généré par les utilisateurs, comme les tags créés sur Flickr pour la Bibliothèque du Congrès). Il est en effet nécessaire de prouver aux contributeurs leur « retour sur investissement » : leur participation doit véritablement améliorer et enrichir les catalogues et bibliothèques numériques, au risque de les voir cesser leurs contributions si elles restent inutiles et inutilisées.

B) ENRICHISSEMENT AUTOMATIQUE DES METADONNEES

Si les pratiques d'enrichissement social s'enracinent progressivement dans les bibliothèques depuis 5 ou 6 ans, les techniques d'enrichissement automatique ou sémantique restent encore

---

[77] Voir Ben W. Brumfield, « Quality Control for Crowdsourced Transcription », *Collaborative Manuscript Transcription*, 2012, http://manuscripttranscription.blogspot.com/2012/03/quality-control-for-crowdsourced.html.

[78] http://www.flickr.com/groups/eugene_trutat/.
Retour d'expérience et statistiques sur *Des bibliothèques 2.0*, articles taggués « theme-flickrcommon » http://bibliotheque20.wordpress.com/tag/theme-flickrcommon/.

largement expérimentales et innovantes. Elles ouvrent toutefois des possibilités remarquables pour le signalement et l'ergonomie de recherche des bibliothèques numériques, et devraient donc rapidement se démocratiser.

Ces technologies, dites de « fouille de données » (*data mining*[79]), permettent d'extraire de l'information à partir de grandes quantités de données, en s'appuyant sur la spécificité du corpus documentaire (récurrence de termes, référentiels spécifiques, etc.)[80]. Elles permettent en particulier, dans le domaine bibliographique, l'indexation automatique de collections numérisées. La fouille de données porte avant tout sur les documents textuels (fouille de texte ou *text mining*) et iconographiques (fouille d'image ou *image mining*), même si des projets de recherche commencent à explorer les possibilités de traitement automatique des données sonores et audiovisuelles[81].

La **fouille de texte** permet notamment d'extraire des données structurées à partir du plein texte : reconnaissance d'entités nommées (personnes, lieux), indexation automatique. Elle s'appuie largement sur l'utilisation de vocabulaires ou de référentiels d'autorité : utilisation des formes retenues et rejetées pour identifier un auteur, appui sur la hiérarchisation des concepts pour relever les thèmes traités dans un texte, etc.

La plateforme de recherche Isidore[82], qui offre depuis 2011 un accès unifié aux données et documents numériques des sciences humaines et sociales, s'appuie largement sur les technologies de fouille textuelle pour enrichir son interface de recherche. Isidore récupère les métadonnées et le texte intégral issus de bases de données, de ressources électroniques et de sites web ; ces informations sont ensuite enrichies en s'appuyant sur des référentiels (notamment RAMEAU) pour fournir une indexation matière et géographique unifiée sur l'ensemble de ces ressources très hétérogènes. L'interface de recherche s'appuie sur la richesse de ces données structurées pour offrir de nouveaux services à l'utilisateur, grâce en particulier à la catégorisation (regroupement par concepts) des différentes ressources.

Dans le monde des bibliothèques, le réseau des bibliothèques de Suisse occidentale, RERO, a lui aussi annoncé le projet de réaliser une indexation et une classification automatique de ses ressources en s'appuyant sur le référentiel RAMEAU[83].

Le Centre pour l'édition électronique ouverte (CLEO), en partenariat avec le Laboratoire Informatique d'Avignon (LIA), a lancé le projet de recherche BILBO[84] sur la reconnaissance et la structuration automatique des références bibliographiques dans les publications électroniques

---

[79] Katherine Chiang, *Data mining, data fusion, and libraries*, 2010, http://docs.lib.purdue.edu/iatul2010/conf/day1/4.

[80] Christian Fauré, « Introduction au Text-mining », *Hypomnemata : supports de mémoire*, 2007, http://www.christian-faure.net/2007/05/30/introduction-au-text-mining/.

[81] Sam Davies, « Multimedia Classification », *BBC Research and development blog*, 2011, http://www.bbc.co.uk/blogs/researchanddevelopment/2011/10/multimedia-classification.shtml. Yves Raimond, « Automatically tagging the World Service archive », *BBC Research and development blog*, 2012, http://www.bbc.co.uk/blogs/researchanddevelopment/2012/03/automatically-tagging-the-worl.shtml.

[82] Réalisée par le très grand équipement Adonis (CNRS), http://www.rechercheisidore.fr/. Voir aussi http://www.tge-adonis.fr/service/isidore.

[83] Voir http://www.rero.ch/page.php?section=news&pageid=ind&fid=68.

[84] Projet « Robust and Language Independent Machine Learning Approaches for Automatic Annotation of bibliographical References in DH Books, Articles and Blogs », http://bilbo.hypotheses.org.

académiques (qu'elles se trouvent dans les bibliographies, les notes de bas de page ou dans le corps du texte), afin en particulier de créer des liens automatiques entre publications, par exemple un blog qui cite un article.

La **fouille d'image** analyse les documents iconographiques d'un corpus pour reconnaître des formes (identification d'objets, reconnaissance de similarités entre images) et pour classer ou regrouper les images par type ou par sujet. Très utilisée par les grands acteurs du web (moteurs de recherche, réseaux sociaux) et de la photographie numérique, elle reste expérimentale dans les institutions culturelles, qui pourtant conservent et diffusent de larges volumes de documents iconographiques sous forme numérique.

La Bibliothèque nationale d'Australie expérimente ainsi une interface de recherche par couleur d'image, appuyée sur des technologies de reconnaissance de couleurs[85]. Le projet *Invisible Australians*[86], développé par deux chercheurs à partir de documents conservés aux Archives nationales d'Australie, utilise des techniques de reconnaissance faciale pour identifier et extraire des portraits photographiques à partir de milliers de documents administratifs numérisés en mode image.

C) METADONNEES ET VISUALISATION

Au-delà du strict enrichissement des métadonnées descriptives des objets numériques, des traitements automatiques, ou parfois manuels, permettent d'améliorer la représentation et la visualisation des données, et d'offrir aux usagers de la bibliothèque numérique des interfaces de recherche plus intuitives (comme le nuage de tags) et davantage porteuses de sens (traduction des contenus, de leur structuration, de leurs relations[87]). Finalement, le traitement des métadonnées structurées permet d'offrir de nouveaux modes de représentation graphique de l'information et donc de nouvelles façons de rechercher dans les données.

Il est ainsi possible d'offrir aux internautes des représentations cartographiques des collections, appuyées sur une géolocalisation des contenus (ajout de métadonnées spatiales générées à partir des métadonnées, de l'indexation géographique en particulier, ou par fouille dans le plein texte). Les médiathèques du Pays de Romans proposent un « cartoguide[88] » sélective des guides touristiques qu'elles conservent, tandis que la Bibliothèque numérique mondiale[89],

---

[85] *Search by color*, http://ll04.nla.gov.au/.
Paul Hagon, *Everything I know about cataloguing I learned from watching James Bond*, 2010, http://www.slideshare.net/paulhagon/everything-i-know-about-cataloguing-i-learned-from-watching-james-bond.

[86] Projet *Invisible Australians, Living under the White Australia policy*, http://invisibleaustralians.org ; l'interface de consultation des images reconnues automatiquement se trouve sous le titre *The real face of white Australia* http://invisibleaustralians.org/faces/.
Tim Sherrat, "The real face of white Australia", *Invisible Australians*, 2011, http://invisibleaustralians.org/blog/2011/09/the-real-face-of-white-australia/.

[87] Sophie Chauvin, *Visualisations heuristiques pour la recherche et l'exploration de données dynamiques : l'art informationnel en tant que révélateur de sens*, thèse Paris 8, 2005, http://tel.archives-ouvertes.fr/docs/00/06/91/27/PDF/SChauvin_these_avril06.pdf.

[88] http://maps.google.fr/maps/ms?ie=UTF8&oe=UTF8&msa=0&msid=111921843359991125734.000446bf8fe3bf754fc09. Voir Lionel Dujol, « Géolocalisons nos collections ! », 2008, *La bibliothèque apprivoisée*, http://labibapprivoisee.wordpress.com/2008/07/17/geolocalisons-nos-collections/.

[89] http://www.wdl.org/fr/

coordonnée par l'UNESCO et la Bibliothèque du Congrès, expose une représentation cartographique de l'ensemble des documents numériques qu'elle diffuse. Le portail Europeana offre quant à lui une cartographie dynamique des résultats de recherche[90], qui représente la répartition des lieux de conservation d'objets numérisés sur un thème donné.

Outre l'évident confort de recherche, cette représentation de l'information spatiale ajoute du sens aux données, en permettant d'un simple coup d'œil de repérer les zones les plus « peuplées » ou au contraire celles auxquelles aucun contenu ne se rapporte. Par ailleurs, la géolocalisation des contenus et leur représentation cartographique permettent d'envisager la création d'usages mobiles fondés sur la découverte d'un territoire.

De même, des représentations sous forme de frises chronologiques améliorent nettement la visualisation des données et apportent une aide à la recherche pour les usagers. Les fresques multimédias thématiques de l'INA[91] permettent ainsi d'explorer chronologiquement les extraits vidéo sélectionnés. Le site *Timelines: Sources from history*[92] de la British Library juxtapose, pour permettre leur comparaison, plusieurs frises chronologiques thématiques (politique, littérature, sciences, etc.) et les documents numérisés correspondant à chacun de ces thèmes. La bibliothèque numérique de la médiathèque de Roubaix propose elle aussi une navigation chronologique parmi ses collections[93]. Le projet d'OCLC *WorldCat Identities*[94] propose pour chaque auteur une représentation sous forme de frise chronologique de ses publications, rééditions et des publications à son sujet, qui offre une vue globale de l'évolution de sa production et de sa notoriété dans le temps.

Des visualisations sous forme de carte heuristique ou de graphe de données mettent particulièrement en valeur les liens entre concepts, et sont surtout utilisées pour naviguer entre thèmes d'indexation. L'interface *Aquabrowser*, intégrée en particulier dans le catalogue de la médiathèque de la *Skema business school* de Lille[95], permet lors d'une recherche de rebondir vers des thèmes liés dans les thesaurus matière ou dans les autres champs d'indexation. L'INA fournit aussi une interface de recherche par graphe, appelée Mediagraph[96] : des « résultats proches » sont proposés pour chaque vidéo (grâce à une recherche par similarité des images), et l'on peut en suivant le graphe naviguer de « résultat proche » en « résultat proche ».

C. ABOLIR LES FRONTIERES : LA DISSEMINATION SOUS LES DEUX ESPECES

Afin d'accroître la visibilité et la fréquentation de la bibliothèque numérique, il est nécessaire de positionner les contenus là où se trouve l'usager, c'est-à-dire sur les sites les plus

---

[90] http://europeana.eu/portal/map.html.
« New Feature: Map Search and Display », *Europeana Blog*, 2012, http://blog.europeana.eu/2012/02/new-feature-map-search-display/.
[91] Par exemple les *Jalons pour l'histoire du temps présent* (http://www.ina.fr/fresques/jalons/accueil) ou *L'Ouest en mémoire* (http://www.ina.fr/fresques/ouest-en-memoire/accueil).
[92] http://www.bl.uk/learning/timeline/index.html.
[93] http://www.bn-r.fr/fr/decouvrir-periode.php
[94] http://www.worldcat.org/identities/
[95] http://mediatheque.skema.edu/index.php?id=362.
[96] http://www.ina.fr/graph/accueil.

fréquentés du web, réseaux et médias sociaux en tête, et d'organiser, de baliser, l'accès de l'internaute au document numérique.

La dissémination des contenus peut prendre plusieurs formes, qui auront toutes en commun de signaler le document numérique « hors les murs » du site de la bibliothèque :

- **duplication des documents**, qui sont alors hébergés à deux endroits différents, sur le site de la bibliothèque et sur un média social de partage (par exemple Wikisource pour les documents textuels, Flickr pour les images fixes, Youtube pour les documents audiovisuels) ou un réseau social (Facebook, Pinterest). L'atout majeur de la duplication, on l'a vu, repose sur la possibilité de bénéficier de l'audience incomparable de ces sites et des possibilités d'enrichissement social des données qu'ils proposent. Mais l'enjeu repose sur la réintégration, le regroupement des informations ainsi générées sur le site de la bibliothèque. Dupliquer ne signifie pas disperser.

  La duplication des contenus numériques est également utilisée pour le développement d'applications mobiles par les bibliothèques, qui offrent ainsi à leurs usagers la possibilité de consulter tout ou partie de la bibliothèque numérique via des terminaux de lecture portables (téléphones mobiles, tablettes), comme à la British Library[97]. L'insertion de contenus dans des applications mobiles (par exemple Culture Clic) ou une offre de livres numériques en format Epub (à la British Library, encore, ou sur Gallica[98]), téléchargeables par les internautes, est également une forme de dissémination des contenus, appuyée sur une appropriation par les usagers qui renforce son efficacité.

- **« apparence » de duplication** du document via un lecteur exportable : les documents sont hébergés sur un site de publication (par exemple Calameo[99] ou Scribd[100] pour les documents textuels, Youtube ou Dailymotion pour les documents audiovisuels) ; et sur la bibliothèque numérique le document est consultable grâce à une « fenêtre » qui montre en réalité le contenu de ce site de publication. Ce système permet de regrouper simplement sur la bibliothèque numérique des documents qui sont hébergés ailleurs, et d'offrir aux internautes d'intégrer à leur tour ces documents sur leurs sites et sur leurs blogs, augmentant ainsi la diffusion des contenus numériques. Mais il pose des problèmes de pérennité, et également de statistiques de consultation puisque c'est le site hébergeur qui est vu par les internautes et qui comptabilise donc une partie des visites.

  Certaines grandes institutions de conservation ont également développé leurs propres lecteurs exportables (c'est le cas de la BnF et de l'INA, par exemple), qui bénéficient alors de tous les avantages de ces lecteurs (non duplication des contenus, intégration facile sur les sites et blogs des usagers ou d'autres institutions) sans présenter de risques de pérennité puisqu'au final c'est bien sur la bibliothèque numérique, et seulement sur la bibliothèque numérique, que les documents sont hébergés. Des logiciels *open source* existent et sont utilisables par des bibliothèques moins équipées, même s'ils ne présentent pas les mêmes garanties d'accès aux contenus et de maintenance face à l'obsolescence logicielle.

---

[97] http://www.bl.uk/app/.
[98] http://gallica.bnf.fr/ebooks?lang=FR.
[99] Comme le proposent les médiathèques de La Roche sur Yon : http://abcd.ville-larochesuryon.fr/lrsy/node/1152.
[100] Par exemple sur la bibliothèque numérique scientifique BibNum http://www.bibnum.education.fr/.

- **dissémination des seules métadonnées** descriptives (par exemple via un moissonnage selon le protocole OAI-PMH) et publication sur un portail regroupant des données de provenances diverses, comme Gallica ou Europeana. L'inconvénient en est une navigation moins agréable pour l'internaute, renvoyé à la bibliothèque d'origine pour la consultation effective des documents.

- simple **mention des contenus** via des liens hypertextes, sur les réseaux sociaux par exemple, qui renvoient l'usager vers la bibliothèque numérique. Il s'agit alors moins de signalement que de valorisation ou de médiation numérique, mais ces opérations concourent souvent à une stratégie de dissémination plus vaste.

Les choix stratégiques de dissémination peuvent alors s'appuyer sur la nature des documents (texte, image, image animée) mais aussi sur le type de publics que l'on cherche à toucher et les interactions que l'on veut mettre en place avec ceux-ci (simple consultation, dialogue, voire opérations de *crowdsourcing*).

La Bibliothèque du Congrès fait office d'établissement précurseur dans la mise en place d'une véritable stratégie de dissémination systématique de ses contenus sur le web : « Notre stratégie est de pêcher là où se trouve le poisson[101] ». La bibliothèque numérique se duplique ainsi progressivement depuis 2009 : photographies sur Flickr, archives sonores sur iTunes, vidéos sur Youtube, médiation numérique sur Facebook et Twitter, applications mobiles, etc.

La bibliothèque peut également développer des pratiques et outils permettant aux internautes de contribuer à la stratégie de dissémination[102] : mise en place de boutons de partage vers les principaux médias sociaux sur chaque document, développement de flux rss sur les nouveautés, possibilités de téléchargement des contenus, création d'un lecteur exportable, etc. On s'appuie alors sur les usagers pour disséminer l'information numérique, pour propulser le signalement des collections numériques « hors les murs », en profitant de l'effet viral des médias sociaux.

C'est la stratégie mise en œuvre par Gallica qui, outre une dissémination active sur les médias sociaux (blog, Facebook, Twitter) et via un partenariat avec Wikimédia France pour la publication de près de 1500 ouvrages sur la bibliothèque libre Wikisource, propose une vignette (pour les images) et un lecteur (pour les livres) exportables qui permettent ainsi aux internautes d'insérer sur leurs sites ou blogs personnels ou encore sur leur page Facebook des contenus de la bibliothèque numérique, qui sont alors consultables et feuilletables par leurs propres visiteurs[103].

---

[101] *« Fishing where the fish are », Web 2.0 and mobile technology at the Library of Congress* https://grahamschool.uchicago.edu/programs/museumpublishingseminar/documents/Jim%20Karamanis%20Presentation.pdf.

[102] Lionel Maurel, *Bibliothèques numériques : quels enjeux, quels modèles ?*, 2011, http://www.bnf.fr/documents/definition_bibnum.pdf.

[103] « Un lecteur exportable pour consulter les livres sur votre site », *Blog Gallica*, 2010, http://blog.bnf.fr/gallica/?p=1579. « Nouveauté Gallica : publier le lecteur exportable sur son mur Facebook », *Blog Gallica*, 2011, http://blog.bnf.fr/gallica/?p=2416.

4. ÉCHANGER ET ENRICHIR : L'INTEROPERABILITE DES BIBLIOTHEQUES NUMERIQUES SUR LE WEB

L'ouverture sur le web permet, comme on l'a vu, de donner une ampleur et un positionnement nouveaux au signalement des bibliothèques numériques. Encore faut-il pouvoir sortir du carcan du site de la bibliothèque pour positionner ses contenus et ses métadonnées sur le chemin des internautes et améliorer leur visibilité, pour croiser ses données avec celles produites par d'autres acteurs (institutions patrimoniales, acteurs de la culture et du web, internautes), et pour bénéficier en retour des enrichissements permis par ces échanges et regroupements de données.

En effet, dans le monde global du web, la segmentation des données et ressources numériques n'a plus beaucoup de sens pour les internautes, qui ne comprennent pas les barrières institutionnelles, qui recherchent contenu et information indépendamment des institutions qui les ont produites. C'est pourquoi les projets conjoints de diffusion de contenus numériques entre institutions culturelles se développent un peu partout, au niveau international (par exemple à l'échelle européenne avec Europeana) aussi bien que régional ou local (multiplication des portails culturels départementaux ou régionaux comme la BnsA[104] ou Manioc[105]), afin d'apporter aux usagers des points d'accès unifiés et plus visibles au patrimoine. Ce rapprochement des institutions patrimoniales pour la diffusion, le signalement et la valorisation de leurs collections est particulièrement soutenu et conceptualisé dans les pays anglo-saxons[106] qui multiplient les projets de collaboration entre ce qu'ils appellent les « LAM[107] » (*Libraries, Archives, Museums*).

Mais l'échange des données des bibliothèques numériques sur le web va bien au-delà du monde patrimonial et culturel. C'est finalement une véritable ouverture sur le « vaste web » (comme on dit le « vaste monde ») qui est permise par l'interopérabilité de leurs ressources. En effet, la mise en place de protocoles et de formats ouverts d'échange de données, ainsi que, dans la mesure du possible, de licences ouvertes de réutilisation, ouvre la voie à des utilisations nouvelles de nos bibliothèques numériques, à des usages que l'on n'aurait pas forcément imaginés ou pu susciter, par exemple des applications mobiles ou des outils de visualisation innovants[108]. Des bibliothèques et d'autres institutions culturelles ont même pu encourager de façon encore plus incitative les réutilisations originales de leurs données, en organisant des « hackathons »,

---

[104] Banque numérique du savoir d'Aquitaine http://bnsa.patrimoines.aquitaine.fr/.

[105] Bibliothèque numérique Caraïbes Amazonie Plateau des Guyanes http://www.manioc.org/.

[106] Voir par exemple Chuck Leddy, « *Linking libraries, museums, archives: U.S. archivist says interactivity must rise because users demand it* », *Havard Gazette*, 2012, http://news.harvard.edu/gazette/story/2012/04/linking-libraries-museums-archives/.

[107] "*Library, Archives and Museum Collaboration*", OCLC Research, http://www.oclc.org/research/activities/lamsurvey/.

[108] Voir par exemple les utilisations des données statistiques de la bibliothèque municipale de Rennes, décrites par Thomas Chaimbault, « Bibliothèques et open data », *Vagabondages*, 2012, http://www.vagabondages.org/post/2012/07/17/Biblioth%C3%A8ques-et-Open-data.

c'est-à-dire des concours de développeurs invités à manipuler et à enrichir les données pour produire de nouveaux services innovants[109].

Il ne faut pas non plus oublier les enjeux de référencement par les moteurs de recherche, dont on connaît les impacts sur la fréquentation des ressources de bibliothèques : l'indexation par les robots des moteurs de recherche s'appuie aujourd'hui très largement sur les liens entrants vers les sites et sur l'exploitation de métadonnées structurées et ouvertes sur le web[110]. Puisque les actions de diffusion et de dissémination des données favorisent la création des liens entrants et comme les robots ne peuvent pas entrer dans les bases de données et les indexer, l'ouverture se révèle, en termes de signalement, synonyme d'amélioration de la visibilité et de la consultation des ressources numériques des bibliothèques.

L'interopérabilité des systèmes documentaires est nécessaire à l'échange des données entre ces systèmes. Pour se parler, pour échanger, il faut se comprendre. Mais avant d'être un ensemble de recommandations sémantiques et d'outils et protocoles techniques, l'interopérabilité est bien un moyen : un moyen pour les bibliothèques d'améliorer le signalement de leurs collections en favorisant l'enrichissement des métadonnées et leur dissémination sur le web, et par conséquent d'augmenter la visibilité, la notoriété et la consultation de leurs ressources numériques ; un moyen pour les usagers de trouver plus facilement des contenus pertinents, et de bénéficier de services plus efficaces.

A. LES QUATRE PILLIERS DE L'INTEROPERABILITE

Dans la perspective d'échanges de données avec d'autres producteurs (autres bibliothèques, institutions culturelles, administrations, éditeurs, grands acteurs du web commercial – moteurs de recherche en particulier – ou non commercial – projets Wikimédia[111] par exemple), les bibliothèques doivent donc positionner leurs données dans un écosystème plus vaste qu'elles. L'interopérabilité des systèmes entre ces différents acteurs est assurée par un ensemble de normes et standards qui garantissent le transport des données (interopérabilité technique) et la compréhension des informations échangées (interopérabilité sémantique)[112]. Elle s'appuie tout particulièrement sur :

- **des protocoles d'échange**, c'est-à-dire les règles qui régissent les échanges d'information entre les machines, par exemple HTTP, le protocole qui définit les échanges entre les machines sur le web, et particulièrement entre les serveurs et les navigateurs.

---

[109] Par exemple le « Library hack » de bibliothèques australiennes et néo-zélandaises http://libraryhack.org/, la série de hackathons d'Europeana http://pro.europeana.eu/reuse/hackathons, ou encore le « hack day » des Archives nationales des Etats-Unis http://labs.nationalarchives.gov.uk/wordpress/index.php/2012/04/the-hacks.

[110] Voir ci-dessous l'encart « Mettre des données structurées dans les pages web : les données dites « embarquées », p. ###.

[111] Voir Rémi Mathis, « Wikipédia et les bibliothèques : dix ans après », dans *Bibliothèques 2.0 : à l'heure des médias sociaux*, 2012, p. 33-40, http://archivesic.ccsd.cnrs.fr/index.php?halsid=k1nojaopp4mb7lc0v1j50fkkc1&view_this_doc=sic_00710428&version=1.

[112] Voir *Référentiel général d'interopérabilité*, Direction générale de la modernisation de l'Etat, 2009, http://references.modernisation.gouv.fr/sites/default/files/RGI_Version1%200.pdf.

- **des formats de données** : des langages d'échange pour se comprendre, une grammaire. Certains formats sont plus adaptés que d'autres à assurer une véritable interopérabilité sémantique[113].

- **des référentiels**, des vocabulaires communs. Là où les moteurs de recherche avaient, aux début du web, fragilisé l'intérêt des référentiels en s'appuyant sur une recherche en plein texte, la multiplication des ressources et les besoins d'interopérabilité et d'échanges de données remettent sur le devant de la scène les vocabulaires contrôlés[114]. Les référentiels sont en effet le seul moyen de s'entendre sur le sens des données échangées. Par exemple, l'utilisation large du référentiel matières RAMEAU dans les bibliothèques françaises, territoriales comme universitaires, permet de construire des services interrogeant conjointement les données de ces institutions (CCFr, Sudoc, Gallica), en étant sûrs que les ressources indexées avec les mêmes termes traitent bien des mêmes sujets.

- **des identifiants** uniques, et si possible pérennes, pour désigner les objets échangés. Lorsque l'on échange des données, qu'on les dissémine sur le web, il est nécessaire de s'assurer que les ressources décrites par ces données (notices de catalogue, notices d'autorité, objets numérisés, etc.) soient dotées d'un identifiant qui permet de les nommer de façon univoque. Pour éviter tout risque de redondance ou d'ambiguïté et assurer la « citabilité » des ressources par les internautes ou par d'autres acteurs du web, il est nécessaire que ces identifiants soient uniques, c'est-à-dire qu'ils désignent une seule ressource et qu'ils aient la même signification partout, et si possible pérennes[115].

  Il est inutile de disséminer ses métadonnées, de les échanger pour accroître leur visibilité, si l'on ne s'est assuré que les internautes retrouveront le chemin vers les ressources de la bibliothèque numérique. Par ailleurs, les moteurs de recherche indexent le web en suivant les liens, il faut donc leur fournir des liens stables vers des ressources dotées de références stables. Les identifiants ont également un rôle central à jouer dans la préservation des ressources numériques[116].

  On pourra citer en particulier le système d'identification ARK[117] (*Archival Ressource Key*), mis en place par la *California Digital Library* et utilisé par la BnF pour les notices bibliographiques et d'autorité et pour les ressources numériques[118], mais aussi le mécanisme DOI[119] (*Digital object identifier*), largement utilisé dans le monde de la recherche pour désigner les publications scientifiques.

---

[113] Voir ci-dessus 2.B. Usages et interopérabilité des formats de métadonnées, p. ###.

[114] Voir actes de la journée d'étude AFNOR/BnF *Référentiels et données d'autorité à l'heure du Web sémantique* (Paris, 27 mai 2011) http://www.bnf.fr/fr/professionnels/autres_journees_professionnelles/a.referentiel_donnees_autorites_110527.html

[115] Voir E. Bermès, *Des identifiants pérennes pour les ressources numériques, L'expérience de la BnF*, 2006, http://bibnum.bnf.fr/identifiants/identifiants-200605.pdf

[116] Voir chapitre sur la préservation.

[117] Site officiel des identifiants ARK https://wiki.ucop.edu/display/Curation/ARK.

[118] http://www.bnf.fr/fr/professionnels/s_informer_autres_numeros/a.ark_autres_numeros.html.

[119] http://www.doi.org/.

B. DES PROTOCOLES ET DES USAGES

Les bibliothèques numériques diffusent des contenus numérisés sur le web, mais ces ressources sont difficilement accessibles car les bases de données se situent dans ce que l'on appelle le « web profond » ou « web caché » : les contenus ne sont accessibles que ponctuellement, à la demande, via un formulaire de recherche. Les machines, comme les moteurs de recherche, ainsi que les réutilisateurs éventuels, ne peuvent pas récupérer, utiliser, indexer ou regrouper ces contenus pour fournir de nouveaux services ou des accès simplifiés. C'est pourquoi des protocoles spécifiques d'échange de données ont progressivement été mis en place. A chaque usage, à chaque type de données, à chaque type d'utilisateurs son protocole.

Le protocole « historique » d'échange de données dans les bibliothèques, **Z 39.50**[120], et la « nouvelle génération » qui l'adapte au monde du web, **SRU**[121], sont tout particulièrement adaptés à l'échange de notices bibliographiques. Les ressources de bibliothèques sont en effet généralement éditées et existent en multiples exemplaires : l'enjeu premier de l'échange de données entre bibliothèques est donc de mutualiser les efforts de catalogage en récupérant les notices d'un catalogue à l'autre.

Utilisés également pour la mise en place de catalogues collectifs et moteurs de recherche fédérés, ces protocoles permettent une interrogation synchrone des catalogues et offrent un grand éventail de fonctionnalités, peut-être trop riches et trop complexes, et qui, à l'expérience, se révèlent souvent mal paramétrées et mises en œuvre par des bibliothèques aux moyens limités.

Développés par et pour des bibliothèques, ces protocoles reposent sur l'utilisation des formats MARC et permettent donc l'échange de données riches, structurées et complètes. En revanche, ils sont complètement incompréhensibles dès que l'on sort du monde des bibliothèques, et ne sont donc pas adaptés aux échanges de données avec d'autres acteurs de la culture et du web.

Le protocole OAI-PMH[122] en revanche, a été spécifiquement créé pour la description de ressources numériques et pour la mise en place de portails d'accès centralisés à ces ressources hétérogènes par nature. Né dans le monde de l'université pour l'échange de données concernant les publications scientifiques, il a rapidement été adopté par les domaines culturels et patrimoniaux, car il s'adapte parfaitement à des besoins génériques transdisciplinaires et permet le partage de données entre bibliothèques, musées, archives, cinémathèques, etc.

Conçu spécifiquement pour la mise en place de portails et méta-moteurs fournissant un accès unique à des bases de données et ressources de natures et de formats divers, il a été choisi par le portail européen Europeana, et s'est très largement imposé comme le protocole d'échange de données incontournable dans le domaine culturel.

Très souple et simple techniquement, il s'appuie sur l'utilisation du format de métadonnées Dublin Core pour faire dialoguer entre eux des jeux de données d'origines diverses. Toutefois, la

---

[120] Publié en 1988 (donc avant la création du web), le protocole Z 39.50 est normalisé à l'ISO en 1998 : « Recherche d'information (Z39.50) - Définition du service de l'application et spécification du protocole ». Son évolution est coordonnée par la Bibliothèque du Congrès http://www.loc.gov/z3950/agency/. Voir aussi http://www.bnf.fr/fr/professionnels/protocoles_echange_donnees/a.proto_z3950.html.

[121] SRU (Search/Retrieval via URL) est l'équivalent fonctionnel du protocole Z39.50 adapté aux standards technologiques du web (HTTP, XML). La dernière version en date, la version 1.2, date de 2007. Ce protocole est lui aussi maintenu par la Bibliothèque du Congrès http://www.loc.gov/standards/sru/. Voir aussi http://www.bnf.fr/fr/professionnels/proto_sru/s.proto_sru_intro.html.

[122] Voir annexe « Le protocole OAI-PMH », p. ###).

pauvreté sémantique et la faiblesse de structuration des éléments du Dublin Core (qui s'appuie sur le « plus petit dénominateur commun » pour la description de ressources aussi diverses que des articles scientifiques, des inventaires d'archives, des œuvres d'art, des films ou des livres) peinent à assurer une véritable interopérabilité sémantique et à s'ouvrir à des usages autres que la constitution de portails. Même dans ce contexte, OAI-PMH s'avère rapidement insuffisant, en raison de l'impossibilité d'indiquer le plein texte ou les vignettes de présentation à afficher, par exemple. Bien que demeurant obligatoire dans la spécification de l'OAI-PMH, le Dublin Core tend déjà à être remplacé par d'autres formats à peine plus riches pour des usages spécifiques. Par exemple, Europeana a créé un format spécifique, ESE[123] (*Europeana Semantic Elements*) pour les besoins d'affichage des données dans le portail, et ce format lui-même, encore trop pauvre, sera progressivement abandonné ; le nouveau modèle de données EDM[124] (*Europeana Data Model*) s'appuie sur les technologies du web sémantique pour accepter une plus grande diversité de formats de diffusion de données.

Europeana : le défi de l'interopérabilité au niveau européen

Europeana a été conçue dès 2004, et mise en ligne en 2008, comme un portail permettant un accès unifié à toutes les ressources culturelles et patrimoniales européennes numérisées, qu'il s'agisse de livres, d'archives, d'œuvres d'art, de monuments historiques, de documents sonores ou audiovisuels, etc. Devant le défi que représentait l'interopérabilité technique et sémantique de bases de données produites par un très grand nombre d'organismes publics ou privés très différents, et dans des contextes géographiques variés, il a fallu faire des choix de processus de chargement de données, de protocoles et de formats d'échange qui ont eu et auront des conséquences durables sur le paysage informationnel européen :

- **un modèle de chargement de données fondé sur la mise en place d' « agrégateurs » intermédiaires**.

Les bases de données décrivant des ressources numérisées en Europe, produites par des bibliothèques, services d'archives, musées, services patrimoniaux, cinémathèques, etc., sont beaucoup trop nombreuses (et en nombre toujours croissant) pour qu'Europeana puisse imaginer être en contact direct avec chacune d'entre elles. Europeana suscite donc la mise en place de portails intermédiaires, qui agrègent chacun un certain nombre de bases de données : au lieu de milliers de bases de données, Europeana moissonne seulement un nombre limité d'agrégateurs.

Ces agrégateurs peuvent être nationaux, ils récupèrent alors des données provenant de tous types d'institutions, mais limitées à un contexte national (par exemple, le moteur français Collections[125], ou le portail espagnol Hispana[126]). Des agrégateurs par type d'institution ont également vu le jour, comme *The European Library* (TEL)[127] pour les bibliothèques (d'abord les

[123] http://www.europeana.eu/schemas/ese/.
[124] http://pro.europeana.eu/edm-documentation.
[125] Le moteur Collections donne accès à 5 millions de documents provenant de 52 bases de données et publications en ligne (chiffres d'août 2012) http://collections.culture.fr/.
[126] http://roai.mcu.es/es/inicio/inicio.cmd.
[127] http://www.theeuropeanlibrary.org/tel4/.

bibliothèques nationales, puis les bibliothèques de recherche et les fonds patrimoniaux des bibliothèques de lecture publique), *Archives Portal Europe*[128] pour les archives ou encore *The European Film Gateway*[129] pour les cinémathèques.

Outre leur rôle de fournisseurs de données à Europeana, ces portails se dotent d'interfaces enrichies qui leur confèrent un intérêt propre. Ainsi, ils proposent généralement à leurs publics davantage de données que ce qui sera transmis à Europeana : Collections, TEL ou *Archives Portal Europe* donnent ainsi accès non seulement aux collections numérisées, mais aussi aux catalogues décrivant les collections physiques. Ils peuvent également cibler des usages différents de ceux du grand public visé par Europeana : TEL ou *Archives Portal Europe* s'adressent un public de chercheurs, d'étudiants, d'érudits, et développent des services et fonctionnalités spécifiquement dédiés à ces publics.

- **le choix du protocole d'échange OAI-PMH**, qui semblait le plus approprié pour traiter simplement et souplement des jeux de données très hétérogènes.

L'utilisation du protocole OAI-PMH s'est progressivement diffusée dans toutes les institutions patrimoniales qui, pour améliorer le signalement de leurs collections, veulent participer à Europeana et/ou à l'un ou l'autre de ses agrégateurs. Et lorsque des portails nationaux, territoriaux (par exemple la BnsA) ou thématiques (par exemple Isidore[130], la plateforme de recherche des sciences humaines et sociales), voient le jour, ils se basent également sur ce protocole, d'abord parce qu'il est tout à fait adapté à ce type d'usage, mais aussi afin de pouvoir être à leur tour moissonnés par un agrégateur et signalés dans Europeana.

Si bien que pour une bibliothèque, diffuser les données descriptives de ses ressources numériques dans un entrepôt OAI-PMH permet de donner une grande ampleur à la dissémination de ces données, dans des portails géographiques ou thématiques nationaux, dans des agrégateurs d'Europeana et finalement dans Europeana elle-même, sans effort supplémentaire. Par exemple, une bibliothèque numérique partenaire de Gallica pourra bientôt être moissonnée, via Gallica, par TEL puis par Europeana.

Toutefois, Europeana regarde maintenant vers les technologies du web sémantique, et il est probable que d'autres moyens de fournir des données seront mis en œuvre dans les années à venir.

- **la question épineuse du format de données**.

Europeana et ses agrégateurs ont rapidement souligné les insuffisances du Dublin Core pour répondre aux besoins fonctionnels d'affichage dans les portails (voir ci-dessus). Certains agrégateurs ont donc fait le choix d'utiliser des formats plus riches (par exemple, *Archives Portal Europe* utilise l'EAD, TEL a mis en place un profil d'application spécifique du Dublin Core, appelé *TEL – Application profile*, abandonné depuis), tandis qu'Europeana elle-même créait son format ESE.

Outre qu'ils sont souvent spécifiques à un projet et donc à l'opposé d'une interopérabilité ouverte, ces formats ne sont pas complètement satisfaisants, car ils appauvrissent les données

[128] http://www.archivesportaleurope.eu/Portal/index.action.
[129] http://www.europeanfilmgateway.eu/.
[130] http://www.rechercheisidore.fr/.

d'origine et les nivellent au « plus petit dénominateur commun ». C'est pourquoi Europeana et ses agrégateurs mènent actuellement une expérimentation sur la mise en place d'un nouveau modèle de données, EDM (*Europeana Data Model*). Celui-ci permettra de décrire de façon plus riche et plus conceptuelle les objets culturels numérisés, en séparant les éléments descriptifs de l'objet original et ceux qui s'appliquent à sa version numérisée fournie par une institution. EDM permettra de renforcer les liens entre objets, liens d'inclusion et de séquence en particulier, et donc d'affiner la granularité des données, mais aussi de faire des liens vers des données d'autorité personnes ou matières. Des regroupements pourront également être réalisés, par exemple entre le fichier numérique d'un manuscrit et les images d'enluminures extraites de ce manuscrit qui se trouveraient dans une base de données distincte. L'utilisation d'EDM ouvrira également la voie à la diffusion des données enrichies par Europeana (dédoublonnages, FRBRisation, etc.) sur le web de données.

Le travail de transformation des données fournies par les bases de données d'origine vers le modèle EDM sera réalisé par les agrégateurs. Ainsi, pour les bibliothèques, un groupe de travail rattaché à TEL a fait des propositions pour la représentation des données bibliographiques dans EDM, qu'elles soient fournies en Dublin Core, en ESE ou dans un format MARC, en intégrant une réflexion sur la FRBRisation de ces données[131].

Les protocoles d'échange traditionnellement utilisés en bibliothèques (Z 39.50, SRU, OAI-PMH) répondent à des besoins spécifiques, pour lesquels ils ont été créés (récupération, enrichissement, dissémination). Leur emploi reste cependant limité à certaines communautés métier spécifiques : les bibliothèques seulement pour Z 39.50 et SRU, les mondes de la recherche et de la culture pour OAI-PMH, et ils peinent à s'adapter à la multiplication des usages : échanges de données complexes, multiplication des acteurs, multiplication des réutilisations. C'est pourquoi, pour s'ouvrir à de nouveaux usages, pour répondre à de nouveaux besoins, d'autres outils sont peu à peu utilisés par les bibliothèques, que ce soit pour diffuser leurs données ou récupérer et utiliser les données diffusées par d'autres.

Les w*eb services* (appelés aussi « *web API* » - pour « *Application programming interfaces* », ou simplement « API » par généralisation) sont des technologies qui permettent à des applications, à des machines, de dialoguer entre elles sur le web de façon synchrone[132].

Ces outils, qui s'appuient sur les standards du web (HTTP, XML, JSON[133]), sont très largement utilisés par les grands acteurs du web, moteurs de recherche, réseaux sociaux, sites commerciaux. Ils permettent d'intégrer dans un site web des services, requêtes, données ou applications provenant d'un autre site. Ils sont particulièrement utiles pour réaliser des *mash-ups*, c'est-à-dire des applications composites qui combinent des contenus et/ou services provenant de

---

[131] http://pro.europeana.eu/web/guest/europeana-libraries-edm.

[132] Voir E. Bermès, *Web services et bibliothèques,* Figoblog, 2005, http://www.figoblog.org/document1057.phpS; S. Mercier, *Les services web ou l'espéranto numérique*, Bibliobsession, 2007, http://www.bibliobsession.net/2007/06/07/les-services-web-ou-l-esperanto-numerique/; E. Cavallié, *Qu'est-ce qu'une API ?*, 2009, http://bibliotheques.wordpress.com/2009/06/25/quest-ce-quune-api/.

[133] JSON (*JavaScript Object Notation*) est un formalisme d'écriture de données générique qui tend, grâce sa simplicité (surtout par rapport à XML : JSON est parfois appelé « l'alternative allégée à XML »), à s'imposer dans les *web services*, en particulier pour le développement d'applications mobiles. Ce n'est pas un standard officiel du web, plutôt un standard de fait (décrit par la RFC 4627 de l'IETF - *Internet Engineering Task Force* - http://tools.ietf.org/html/rfc4627), mais le W3C s'y intéresse de plus en plus. Voir le site officiel : http://www.json.org/.

plusieurs applications, comme par exemple la création d'une carte géolocalisant des livres conservés dans une bibliothèque en combinant des données de cette bibliothèque avec une application *Google Maps*.

Ainsi, par exemple, l'ABES développe et propose de nombreux *web services*[134], et micro *web services* dédiés à une fonctionnalité spécifique, à destination des bibliothèques et autres acteurs du monde du livre intéressés par les données du Sudoc, de theses.fr, de Calames ou de IdRef. Par exemple, le *web service Multiwhere* permet d'identifier automatiquement, à partir de son identifiant PPN, les lieux de conservation d'un document et de les intégrer dans les résultats de recherche d'un catalogue distant, dans le catalogue d'une bibliothèque universitaire qui ne conserve pas ce document, par exemple, pour indiquer directement à l'usager où il pourra le trouver, sans avoir besoin de le renvoyer sur le Sudoc.

Pour avoir une idée de la richesse des réutilisations permises par des *web services* proposés par des institutions culturelles, on peut se tourner vers le Brooklyn Museum, qui tient à jour une liste des applications réalisées en utilisant son API[135] : intégration des données décrivant les collections numérisées du musée dans de grandes bases de données artistiques, dans des applications mobiles (smartphones, tablettes) ; mais aussi constitution de jeux de données exploités par des chercheurs en histoire de l'art.

La plupart des bibliothèques de petite taille sont toutefois plutôt utilisatrices de *web services* que productrices. Outre ceux de l'ABES, les *web services* d'Amazon[136], LibraryThing ou WorldCat sont ainsi très utilisés par les bibliothèques pour enrichir leurs catalogues, avec l'intégration de résumés des ouvrages, de commentaires ou encore de vignettes d'illustration.

Les API permettent d'interroger des données structurées, et donc de créer de nouvelles applications à partir de ces données, et aussi de fournir des fonctionnalités de services à distance sur ces données, mais chaque API est propriétaire, et n'est valable que pour un seul et unique silo de données. Et ce sont bien des fonctionnalités de requête sur les données qui sont permises, les données elles-mêmes ne sont pas exposées telles quelles sur le web, ce qui rend difficile la création de services ou d'usages qui n'auraient pas été imaginés par le fournisseur de données. De plus les *web services* n'utilisent pas tous les mêmes formats et langages. Par exemple, SRU, dont nous avons déjà parlé, est une forme de *web service* portant sur des catalogues de bibliothèques ; il permet certes l'utilisation d'autres formats que MARC pour interroger grâce aux « *context sets* », mais les concordances doivent être créées par la bibliothèque sur le serveur SRU et ne peuvent pas être proposées par l'utilisateur. Il s'agit donc d'un système qui reste fermé et n'est, au demeurant, pas du tout utilisé hors du monde du livre.

C'est donc dans d'autres technologies, véritablement ouvertes, qu'il faut rechercher une meilleure interopérabilité technique et sémantique : celles du web de données.

C. BIBLIOTHEQUES NUMERIQUES ET WEB DE DONNEES

[134] « Les micro *web services* ABES », *Punktokomo*, 2011, http://punktokomo.abes.fr/2011/07/04/les-micro-web-services-abes/; « L'ABES pour les développeurs » http://m.abes.fr/Acces-direct-a/Pour-les-developpeurs.
[135] Brooklyn Museum Application Gallery http://www.brooklynmuseum.org/opencollection/api/docs/application_gallery.
[136] Voir par exemple le catalogue de la Bibliothèque municipale de Rennes http://www.bibliotheques.rennes.fr/.

Le web est par nature normalisé et interopérable, il est fondé sur des standards de publication et d'échange d'information coordonnés par le W3C (le *World Wide Web Consortium*, l'organisme de normalisation du web). Plutôt que d'utiliser des protocoles spécifiques, adaptés à des usages et à des milieux spécifiques, pour ouvrir les données des bibliothèques sur le web, pourquoi ne pas plutôt mettre directement ces données SUR le web, DANS le web ? C'est le rôle du web dit « de données ».

Au-delà d'une complexité conceptuelle et technique initiale pour la « prise en main » du web de données, les bibliothèques trouvent d'ailleurs dans cet écosystème un monde qui leur est parfaitement naturel et où elles ont parfaitement leur place : un monde de données structurées, reliées entre elles, liées à des référentiels, et dont la qualité est primordiale... c'est-à-dire justement des données qui correspondent à ce que proposent les catalogues de bibliothèques.

Comme l'affirme le rapport du W3C sur les bibliothèques et le web de données[137], les bibliothécaires et les bibliothèques ont un rôle majeur à jouer dans le développement du web de données précisément parce qu'ils disposent de larges jeux de données structurées, contextualisées, pérennes et validées, et de référentiels d'autorité encyclopédiques.

Mais ces données sont aujourd'hui enfermées dans des bases de données étanches et inaccessibles depuis l'extérieur, que l'on compare souvent à des « silos », riches de grains, mais clos.

IL FAUT « DECLOISONNER LES SILOS » !

Les machines (les moteurs de recherche, applications et services désirant utiliser les données des bibliothèques, mais aussi les catalogues de bibliothèques désirant réutiliser des données extérieures pour l'enrichissement de leurs ressources) ne peuvent pas entrer dans ces silos. Elles ne savent pas inventer des questions pour aller au-delà des formulaires de recherche qui sont la porte d'accès à ces bases.

Pour ce faire, le W3C a standardisé un ensemble de technologies, protocoles, langages, formats, modèles, qui doit permettre à ces machines d'utiliser des données structurées pour fournir des services plus riches aux utilisateurs. L'ensemble de ces techniques est désigné sous le nom de « web sémantique », tandis que le résultat de ce qui est publié sur le web grâce à ces techniques sera appelé « web de données » ou encore « *linked data* » (données liées). Il ne s'agit pas de créer un nouveau protocole ou des *web services* spécialisés, mais précisément d'utiliser les technologies du web, et rien que les technologies du web, pour diffuser et échanger les données.

Le web de données crée ainsi un espace documentaire unifié, global, commun, où les données elles-mêmes peuvent être reliées l'une à l'autre (d'où le nom de « *linked data* »), et non plus les documents dans leur globalité comme dans le web « traditionnel » où un lien pointait simplement d'une page web vers une autre. Par exemple, on pourra déclarer que le photographe Gustave Le Gray, décrit par une notice d'autorité dans le catalogue de la BnF, est exactement la même personne que le Gustave Le Gray décrit dans les autorités auteurs du Sudoc ou que celui décrit dans Wikipédia.

---

[137] *Library Linked Data Incubator Group Final Report*, 2011, http://www.w3.org/2005/Incubator/lld/XGR-lld-20111025/. Voir la traduction française, 2012 : http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html.

Là où les protocoles d'interopérabilité de type Z 39.50, SRU, OAI-PMH imposent des formats de données (formats MARC, Dublin Core), le web de données permet en revanche l'utilisation et le mélange de multiples formats. Il n'est plus nécessaire d'appauvrir ses données au « plus petit dénominateur commun » pour pouvoir les rendre interopérables avec d'autres communautés. Les données ainsi ouvertes sur le web de données seront donc plus riches, plus précises et finalement plus propres.

De plus, à la différence des *web services* qui constituent un filtre entre les données et l'utilisateur, sur le web de données ce ne sont pas des fonctionnalités de requêtes qui sont ouvertes, mais bel et bien les données elles-mêmes, si bien que l'on ne préjuge pas des usages de réutilisation qui pourront en être faits. Et puisque ces données sont diffusées et interrogeables selon des standards du web normalisés au niveau international par le W3C (HTTP, HTML, XML, ou encore les standards spécifiques du web sémantique : RDF, SPARQL), elles sont susceptibles d'être utilisées par tous les acteurs du web, sans limitation à une communauté particulière.

Les bibliothèques peuvent grâce à ces techniques enrichir leurs catalogues avec des données d'origine très diverse, à condition qu'elles soient elles-mêmes ouvertes sur le web de données (comme des données bibliographiques proposées par une autre bibliothèque, des données géographiques[138], des thesaurus scientifiques, des informations ou images de Wikipédia, etc.) et susciter en échange des utilisations nouvelles, innovantes, inattendues, de leurs ressources. Par exemple, les données de data.bnf.fr[139] diffusées sur le web de données sont réutilisées aussi bien par des développeurs d'applications mobiles culturelles, par des portails bibliographiques utilisant la FRBRisation des données de data.bnf.fr, par des chercheurs qui développent des outils expérimentaux de visualisation de données, ou encore par des bibliothèques qui récupèrent ces données pour faire de la fouille textuelle et de l'indexation automatique de leurs propres collections.

C'est finalement l'intégration dans un cycle vertueux de l'échange de données que permet la diffusion de données bibliographiques sur le web de données, où les bibliothèques peuvent à la fois diffuser leurs données, en accroître la visibilité et la notoriété, et en retour récupérer elles-mêmes des données d'autres origines. Elles contribuent à améliorer les services rendus aux usagers ou à développer de nouveaux services, et elles ouvrent largement leurs ressources au référencement par les moteurs de recherche qui indexent facilement ces données structurées.

Afin de permettre une réelle ouverture de ces données, l'utilisation des technologies du web sémantique n'est pas tout à fait suffisante, et il est nécessaire de doubler cette ouverture technique d'une ouverture juridique[140] (par exemple par la mise en place de licences libres de réutilisation de type CC0, CC-BY[141] ou la Licence Ouverte d'Etalab[142]), on parle alors de « *linked open data* » (données ouvertes et liées).

Et c'est ainsi que l'on pourra réellement faire tomber les silos, pour ouvrir sur le web le riche grain des données structurées de bibliothèques et le faire fructifier au contact des autres données ouvertes.

OUI MAIS COMMENT ÇA MARCHE ?

---

[138] Voir ci-dessus l'encart « Les métadonnées géographiques : formats et gestion », p. ###.

[139] Voir ci-dessus l'encart « data.bnf.fr : un « pivot documentaire » pour le signalement des ressources de la BnF », p. ###.

[140] Voir le chapitre sur les questions juridiques.

[141] http://creativecommons.fr/.

[142] http://www.etalab.gouv.fr/pages/Licence_ouverte_Open_licence-5899923.html.

Le web de données repose sur les quatre composantes de base du web :

- le protocole HTTP, qui assure la communication entre machines sur le web, par exemple entre un navigateur web et un serveur hébergeant les données,
- un mécanisme d'identification : les URL (*uniform ressource locators*) sont les adresses qui désignent une ressource sur Internet (page ou fichier, par exemple),
- le principe de l'hypertexte, c'est-à-dire des liens entre ressources,
- le langage HTML pour l'affichage des contenus des pages web.

A ces principes de base s'ajoutent des standards plus spécifiques au web de données[143] :

- **des identifiants, les URI**[144] (*uniform ressource identifiers*)

Afin de nommer précisément et uniformément les données que l'on va publier, et de pouvoir les lier entre elles et avec des données d'autres jeux de données, le web de données s'appuie sur des identifiants, que l'on appelle URI. Ces URI sont de préférence des identifiants web (commençant par http[145]), pour que les machines comme les usagers puissent obtenir des informations sur ces ressources ; mais il peut aussi s'agir d'identifiants qui n'ont de sens que dans le contexte particulier d'un jeu de données. On a déjà dit toute l'importance des identifiants uniques et pérennes pour l'interopérabilité et la citabilité des ressources numériques[146] ; dans le web de données, ce ne sont plus seulement les objets numériques qui sont dotés d'identifiants mais chaque donnée, chaque information, chaque concept : par exemple l'objet physique d'origine, sa représentation numérique, son auteur, ses thèmes d'indexation, etc. Finalement, les pratiques des bibliothécaires sont souvent déjà très proches de ces objectifs, en attribuant un identifiant à l'objet physique (cote) et un à la représentation numérique, et en proposant des liens vers les identifiants des notices d'autorité (auteur, thèmes). Mais dans les catalogues, les identifiants et les liens sont enfermés dans des notices et dans des bases de données.

- **un cadre de description, RDF** (Ressource description Framework)

Le modèle RDF est un modèle de graphe, qui « atomise » la notice descriptive pour la diviser en un ensemble de déclarations minimales autonomes. Ces déclarations, ou « triplets », se présentent sous une forme « sujet – verbe – complément » (« sujet – prédicat – objet » dans le langage RDF), où chaque élément est désigné par une URI. Ces triplets sont reliés entre eux pour constituer un graphe, ce qui permet une très grande souplesse descriptive.

Par exemple, une notice bibliographique :

> Identifiant de la notice : http://catalogue.bnf.fr/ark:/12148/cb35347035r/
> Auteur : Melville, Herman (1819-1891)
> Titre(s) : Moby-Dick, or The Whale
> Titre d'ensemble : The Writings of Herman Melville. ; 6
> Publication : Evanston, Ill. : Northwestern university press ; Chicago, Ill. : Newberry library, 1988

peut être exprimée sous forme de graphe RDF par un ensemble de triplets reliés entre eux, notamment :

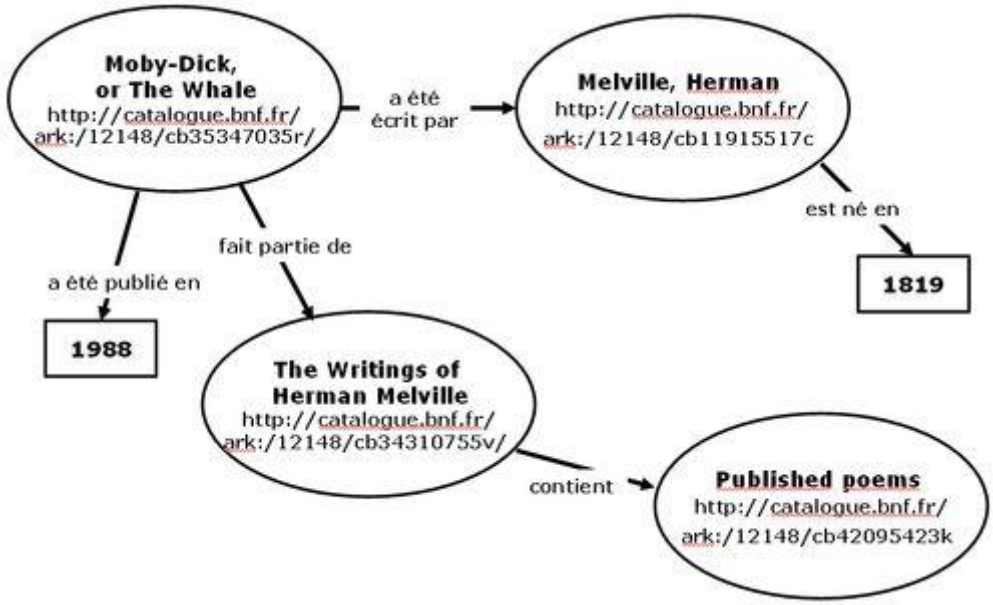
---

[143] Voir les pages consacrées au web sémantique sur le site du W3C http://www.w3.org/standards/semanticweb.

[144] Le W3C fournit des recommandations sur l'utilisation et la forme des URI : *Cool URIs for the Semantic Web* (2008) http://www.w3.org/TR/cooluris/.

[145] Les URL sont une forme d'URI.

[146] Voir ci-dessus 4.A. Les quatre piliers de l'interopérabilité, p. ###.



Les différents concepts (par exemple un livre, un auteur, un titre propre, un ISBN) ainsi que les relations qualifiées entre eux (par exemple « a été écrit par », « contient », « a pour exemplaire numérisé ») doivent être définis par des ontologies pour pouvoir être compris par les machines[147]. Dans un souci de normalisation et pour favoriser une véritable interopérabilité sémantique, il est recommandé d'utiliser des ontologies documentées et largement adoptées[148].

Pour décrire une ressource numérisée par une bibliothèque, on pourra ainsi piocher des éléments de description dans différentes ontologies en fonction des besoins[149], par exemple :

- des éléments de Dublin Core pour décrire la ressource numérique elle-même : le type de document, le format, l'identifiant du fichier,
- des éléments de l'ISBD[150], qui a été publié comme ontologie du web sémantique par l'IFLA, pour décrire le document d'origine : le titre, les mentions de responsabilité et de publication, l'ISBN, etc., avec toute leur richesse et leur précision bibliographique,
- pour décrire les personnes, auteurs ou contributeurs, on pourra utiliser une ontologie spécifiquement adaptée à ce type d'information, comme FOAF pour décrire les relations entre personnes, ou BIO[151] pour des éléments biographiques plus précis (date de naissance, date de décès), ou encore une ontologie spécifique au monde des bibliothèques comme MADS/RDF[152].

- **des langages de présentation et d'interrogation des données** : les langages de représentation des ontologies RDFS et OWL, le langage d'interrogation SPARQL.

---

[147] Une ontologie est un système d'organisation des connaissances qui se présente comme un ensemble structuré de concepts liés par des relations sémantiques d'inclusion ou d'héritage.

[148] Le projet LOV (Linked Open Vocabularies) recense ainsi les différentes ontologies de métadonnées, publiées et documentées, utilisables sur le web de données, et leur notoriété http://labs.mondeca.com/dataset/lov/.

[149] Voir aussi ci-dessus 2.C. Le cycle de vie des métadonnées.

[150] http://iflastandards.info/ns/isbd/elements/

[151] BIO est une ontologie pour les informations biographiques http://vocab.org/bio/0.1/.html.

[152] La Bibliothèque du Congrès a publié le format MADS comme ontologie sur le web de données www.loc.gov/standards/mads/rdf/.

Pour résumer, le web de données c'est :

- des ressources à décrire,
- des URI pour les identifier,
- un modèle RDF pour les représenter et les relier,
- des langages pour les exprimer et les interroger.

MAIS A QUOI ÇA SERT CONCRETEMENT, POUR LES BIBLIOTHEQUES ?

Ouvertes sur le web de données, les données des bibliothèques s'offrent à des **réutilisations multiples et innovantes** qui leur donneront une notoriété inédite, et contribueront à améliorer la consultation des ressources numériques tout en positionnant clairement les bibliothèques dans le flux informationnel du web et dans l'économie de la connaissance.

Les réutilisateurs potentiels sont autant les entreprises et utilisateurs privés que les autres bibliothèques et institutions culturelles[153], ou encore les grands acteurs du web (moteurs de recherche, réseaux sociaux).

Mettre des données structurées dans les pages web : les données dites « embarquées »

Les données en RDF peuvent être exposées de différentes manières :

- par navigation, on dit par « **négociation de contenu** », c'est-à-dire que pour une même page web un être humain pourra consulter normalement la page qui s'affichera en HTML, alors qu'une machine (le robot d'un réutilisateur ou d'un moteur de recherche) sera renvoyée vers les mêmes données en RDF[154],
- par téléchargement massif de l'ensemble des données (« ***dump*** »),
- par interrogation fine selon le langage normalisé **SPARQL**,
- par insertion directement dans le code HTML des pages, on parle alors de **données « embarquées »** dans le HTML.

Ces données structurées directement lisibles dans les pages web sont comme un premier pas simplifié vers le web de données : elles sont généralement peu complexes et peu nombreuses sur chaque page, pour un « retour sur investissement » particulièrement puissant puisqu'elles sont très facilement utilisables par les moteurs de recherche pour indexer et qualifier le contenu des

[153] Le monde de la culture et du patrimoine attend beaucoup du web de données pour faire tomber les barrières entre institutions, et parvenir enfin à interopérer véritablement les données des bibliothèques, archives et musées : la communauté informelle « LOD-LAM » (*Linked open data in Libraries, Archives and Museums* http://lodlam.net/) organise des rencontres entre professionnels de la culture sur ces sujets, propose des outils et des règles de bonnes pratiques techniques et juridiques.

[154] Voir par exemple la page de data.bnf.fr sur Molière, en HTML pour les humains http://data.bnf.fr/11916418/moliere/fr.html et en RDF pour les machines http://data.bnf.fr/11916418/moliere/rdf.xml

pages.

C'est pourquoi on assiste à une explosion de l'utilisation de ces données embarquées par de nombreux acteurs du web, institutions publiques comme acteurs privés et commerciaux. Toutefois, cette utilité et cette simplicité évidentes ne doivent pas occulter le fait que ces données sont appauvries et limitées à des usages spécifiques de référencement, elles peuvent difficilement être utilisées par des développeurs ou des chercheurs qui voudraient travailler sur l'ensemble du jeu de données[155].

Différents formats plus ou moins normalisés permettent d'embarquer ces données : « microformats », « microdata » ou encore « RDFa ». Ils ont tous en commun de signaler des données structurées dans le texte d'une page HTML : par exemple, sur une page web décrivant un film, on pourra indiquer, dans un langage compréhensible par une machine et en particulier un moteur de recherche, qu'il s'agit d'un film, et quel est son titre, son réalisateur, ses acteurs principaux, etc.

Les principaux moteurs de recherche (Google, Yahoo! et Bing) se sont mis d'accord sur un format privilégié pour ces données embarquées, appelé *schema.org*[156]. Ce format à vocation encyclopédique permet de décrire très simplement tous types d'objets, comme des livres, des films, de la musique, des personnes, des lieux, etc. Devant l'assurance d'un meilleur référencement et d'un meilleur affichage des résultats, de nombreux fournisseurs de contenus sur le web ont intégré *schema.org* dans leurs pages, et les bibliothèques ne sont pas en reste : la BnF pour data.bnf.fr, l'ABES pour IdRef, ou encore OCLC pour WorldCat. Pour une bibliothèque aux moyens limités qui désirerait avancer vers le web de données, la mise en place sur ses pages web de données structurées selon le format *schema.org* pourrait constituer une première expérimentation relativement simple.

Les moteurs et robots qui indexeront le contenu de la page pourront non seulement comprendre et analyser son contenu, mais aussi l'afficher de façon plus claire pour les utilisateurs. Les moteurs de recherche commencent ainsi à proposer directement dans le résultat des requêtes une extraction de ces données structurées embarquées : par exemple, pour les recettes de cuisine, on voit apparaître la liste des ingrédients, le temps de cuisson, les votes des internautes, etc.

Depuis peu, Google utilise également ces données pour proposer une nouvelle forme de représentation de l'information, le « *Knowledge Graph*[157] », déployé tout d'abord sur la version anglophone du moteur. Google exploite les données structurées dans les pages ainsi que d'autres ensembles plus riches du web de données, comme Freebase[158], pour désambiguïser les requêtes (faites-vous une recherche sur Jules César le personnage historique, ou sur *Jules César* la tragédie de Shakespeare ?), pour proposer dès l'affichage des réponses une fiche signalétique regroupant toutes les informations structurées concernant l'objet de la recherche (pour une personne : sa biographie, ses dates, ses liens familiaux, ses activités, etc.), et pour créer des réseaux de relations entre les concepts, sous forme de graphe, permettant aux usagers de rebondir le long de ce

[155] De nouveaux usages de ces données apparaissent tout de même : ainsi, le portail Isidore indexe certaines des ressources numériques qu'il signale grâce à des données en RDFa publiées dans leurs pages.
[156] http://schema.org/
[157] Voir *Introducing the Knowledge Graph: things, not strings*, 2012, http://googleblog.blogspot.fr/2012/05/introducing-knowledge-graph-things-not.html
[158] Freebase est une encyclopédie collaborative sous forme sémantique, qui a été rachetée par Google en 2010 http://www.freebase.com/.

graphe (par exemple, lors d'une recherche sur Marie Curie, l'usager se voit proposer de rebondir non seulement vers Pierre Curie ou Irène Joliot-Curie, mais aussi vers d'autres scientifiques ayant reçu le prix Nobel de physique ou de chimie).

Toutefois, les moteurs de recherche ne sont pas les seuls grands acteurs du web à s'intéresser de près aux données embarquées, les réseaux sociaux s'y investissent également. Facebook a ainsi développé son format spécifique fondé sur RDFa, l'*Open Graph Protocol*[159] (OGP). Un fournisseur de contenu peut ainsi ajouter dans ses pages web des données structurées décrivant ses ressources (type de la ressource – livre, image, film, etc. –, titre, description, identifiant, adresse d'une vignette d'illustration, etc.), afin de permettre leur intégration optimale dans Facebook : lorsqu'un usager clique sur le bouton « J'aime » intégré dans un site pour recommander une ressource, ou lorsqu'il signale cette ressource directement sur le réseau social, Facebook va chercher dans les données OGP les informations qui seront affichées. L'utilisation d'OGP est simple et nécessaire à l'optimisation de la dissémination des ressources numériques et donc de leur visibilité sur le web social. La BnF intègre ainsi ce format dans les pages de Gallica et de data.bnf.fr pour s'assurer de la qualité de la dissémination de ses ressources.

Depuis un an ou deux, les expériences de publications sur le web de données par les bibliothèques se multiplient, en particulier au niveau des bibliothèques nationales. Les travaux du groupe de travail du W3C sur les bibliothèques et le web de données[160], qui a rendu son rapport final en octobre 2011, autant que les réflexions et réalisations de l'IFLA sur le sujet[161], donnent un élan et un soutien technique et méthodologique aux initiatives locales.

C'est ainsi que le DataHub, qui recense les données publiées sur le web de données, compte pas moins de 57 jeux de données dans la catégorie « *Library Linked Data*[162] » :

- de nombreux référentiels[163] : VIAF[164], RAMEAU, les autorités du Sudoc (IdRef), les autorités de la Bibliothèque du Congrès, etc.
- des catalogues, parfois FRBRisés (comme data.bnf.fr, le catalogue collectif suédois Libris ou le catalogue espagnol datos.bne.es) mais pas toujours (British Library et Bibliothèque nationale d'Allemagne),
- en revanche, très peu de bibliothèques seulement numériques sont à ce jour disponibles sur le web de données, à l'exception d'Europeana. En effet, l'objectif du modèle RDF et du web de données étant précisément de supprimer les silos, y compris au sein des institutions, pourquoi diffuser seulement la bibliothèque numérique et pas le reste des collections (catalogues, expositions virtuelles, ressources en ligne, etc.) ? C'est précisément

[159] Site officiel de l'Open Graph Protocol http://ogp.me/.

[160] *Library Linked Data Incubator Group* http://www.w3.org/2005/Incubator/lld/. Voir la traduction française du rapport final : http://mediatheque.cite-musique.fr/MediaComposite/ARTICLES/W3C/XGR-lld-fr.html.

[161] En particulier dans le cadre du *Semantic Web Special Interest Group* http://www.ifla.org/en/swsig et du *Namespaces Task Group* http://www.ifla.org/en/node/5353.

[162] Chiffre de septembre 2012 http://thedatahub.org/group/lld?page=1

[163] Sur le rôle central des référentiels pour l'interopérabilité des ressources culturelles, voir Emmanuelle Bermès, « Convergence et interopérabilité : l'apport du Web de données », IFLA 2011, http://www.buw.uw.edu.pl/images/mapa/IFLA_2011/papers/149-bermes-fr.pdf.

[164] Le projet VIAF (http://viaf.org/), coordonné par l'OCLC et auquel participent en particulier la Bibliothèque du Congrès, la BnF, la British Library et la Bibliothèque nationale d'Allemagne, a pour objectif de relier et d'aligner les référentiels d'autorité des différentes bibliothèques, en commençant par les noms de personnes. Cela permet par exemple de disposer pour chaque nom des équivalents en caractères non latins : la publication de ces données sur le web de données permet à chacun de récupérer les données enrichies par VIAF.

> le sens du projet data.bnf.fr, qui utilise les techniques du web sémantique pour le traitement des données des différents catalogues et ressources numériques de la BnF, Gallica en faisant bien entendu partie, mais pas seulement.

Le web de données, et plus précisément le « *linked open data* » constitue un véritable enjeu pour l'évolution du paysage informationnel des bibliothèques, et tout particulièrement pour la récupération par les bibliothèques de données produites par les bibliothèques nationales[165]. Il permet d'envisager la mutualisation au niveau national ou international des opérations de qualification et d'enrichissement de données, ainsi que la maintenance des référentiels. La publication des bibliographies nationales sur le web de données ouvrirait par ailleurs vers la mise en place de services complètement nouveaux, où les bibliothèques n'auraient plus à dériver via Z 39.50/SRU ou FTP les notices des publications mais pourraient lier directement, sur le web, leurs données locales aux données bibliographiques et d'autorité publiées par les bibliothèques nationales, sans duplication des contenus. La British Library[166] expérimente ainsi la publication sur le web de données de sa bibliographie nationale.

Par ailleurs, les bibliothèques ont un rôle majeur à jouer pour assurer la qualité et la fiabilité des données publiées sur le web de données. Elles sont très attachées à publier des données riches, à haute valeur descriptive, alors que les autres grands acteurs sont davantage prêts à utiliser des formats appauvris comme schema.org ou OGP, limités à des usages très spécifiques de référencement par les moteurs de recherche ou de dissémination sur les réseaux sociaux. En apportant des jeux de données riches et de qualité auxquels les autres acteurs peuvent se lier, les bibliothèques constituent une base solide pour le web de données, et contribuent à aider les usagers à trouver plus facilement de l'information fiable dans la masse informationnelle du web.

Diffuser ses données sur le web de données reste toutefois encore complexe pour une bibliothèque de taille moyenne. L'évolution des outils et plus largement du paysage de l'information bibliographique en France répondra sans doute dans les années à venir à ces besoins. Mais les bénéfices du web de données ne tiennent pas seulement dans la diffusion des données mais aussi – voire surtout – dans la **réutilisation de données publiées par d'autres acteurs**, pour enrichir les données des catalogues et bibliothèques numériques, pour offrir à nos utilisateurs de nouvelles interfaces et de nouveaux services.

Grâce à l'utilisation de données ouvertes, on peut en effet créer plus simplement ce que l'on appelle des « *mash-up* », c'est-à-dire des applications utilisant et mêlant des données d'origines diverses pour fournir un nouveau service. Ce type d'application était déjà possible grâce aux API, mais il nécessitait de plus lourds développements pour s'adapter aux formats propriétaires de chacune des sources. On peut ainsi par exemple signaler un prototype réalisé pour la création de fiches sur les monuments historiques grâce à la réutilisation de données de sources multiples publiées en *open data* ou *linked open data* : bases de données du ministère de la Culture, données de la SNCF et de l'INSEE, Wikipédia, photographies de Wikimédia Commons, etc.[167] : cet outil permet une riche expérience de consultation, où l'on peut par exemple rechercher et afficher les photographies des monuments historiques situés à moins de 50 km de telle gare SNCF…

---

[165] Voir Gildas Illien, *Are you ready to dive in ? A case for Open data in national libraries*, IFLA 2012, http://conference.ifla.org/sites/default/files/files/papers/wlic2012/181-illien-en.pdf.

[166] http://www.bl.uk/bibliographic/datafree.html#lod.

[167] http://labs.antidot.net/search?afs:service=50005.

Ces réutilisations ouvrent des possibilités intéressantes pour les bibliothèques et leurs catalogues. Par exemple, data.bnf.fr enrichit les données des catalogues de la BnF grâce à l'ajout de vignettes d'illustration et de biographies des auteurs issues de Wikipédia, mais aussi de liens entre les thèmes RAMEAU et des référentiels d'autres bibliothèques (Bibliothèque du Congrès, Bibliothèque nationale d'Allemagne)[168].

Le monde du web de données est en expansion constante, et les bibliothèques s'y installent de façon durable. En l'espace de quelques années, on est passés de quelques projets expérimentaux à de nombreuses réalisations de grandes institutions. Le paysage de l'information bibliographique en est progressivement modifié, et les projets dans ce sens devraient se développer en même temps que les outils techniques. En attendant, les bibliothèques, numériques ou non, peuvent commencer à bénéficier des atouts du web de données, ne serait-ce qu'en diffusant de simples données embarquées pour améliorer leur référencement et leur dissémination, ou en profitant des données diffusées par d'autres acteurs pour enrichir leurs propres catalogues et les fonctionnalités offertes à leurs usagers.

CONCLUSION : « NO MATTER WHAT THE QUESTION IS, THE ANSWER IS METADATA ».

Cette phrase de Kara Van Malssen explicite de manière plaisante le rôle crucial des métadonnées dans la vie des bibliothèques. Quels que soient les objectifs d'une bibliothèque numérique, sa politique documentaire, son positionnement et sa politique de service, c'est en maîtrisant les métadonnées que les bibliothécaires acquièrent les outils de leur mise en œuvre. Interaction avec le public ? Métadonnées ! Interfaces visuelles et graphiques ? Métadonnées ! Interopérabilité et coopération avec d'autres bibliothèques ? Métadonnées. Préservation à long terme ? Métadonnées, encore.

Dans l'univers numérique, le document est devenu sa propre description, a acquis une nature ubiquitaire et fait l'objet d'accès à différents niveaux de granularité. Pour maîtriser ces nouveautés et organiser les services aux publics, les métadonnées sont indispensables et des centaines de formats ont été créés durant la dernière décennie. Pour évaluer ceux qui seront le plus utile à sa bibliothèque numérique, l'on peut se laisser guider par ses partenaires, si l'on a déjà un projet de coopération, ou, de façon plus générale, s'assurer que le format est interopérable et bénéficie d'une bonne documentation. Quant aux caractéristiques du format, il faut les évaluer en fonction de sa capacité à structurer l'information et à organiser l'information descriptive.

La gestion des métadonnées peut être complexe : outre les différents formats, il faut tenir compte des différentes sources de données qu'il faut pouvoir combiner ; certaines sont internes à la bibliothèque, mais certaines lui sont extérieures (ressources électroniques, etc.). Les informations fournies par les utilisateurs (commentaires, évaluation, indexation, etc.) ou par un enrichissement automatique (fouille de données) doivent également prendre leur place dans les catalogues pour offrir un meilleur service documentaire, qu'il s'agisse de s'inscrire dans une économie de la recommandation omniprésente sur le web, ou d'offrir de nouvelles fonctionnalités de recherche comme les interfaces visuelles.

---

[168] Voir ci-dessus l'encart « data.bnf.fr : un « pivot documentaire » pour le signalement des ressources de la BnF », p. ###.

Pauline Moirez et Stutzmann, Dominique. « Signaler les ressources numérisées : enrichissement, visibilité, dissémination ». In *Manuel de constitution de bibliothèques numériques*, édité par Isabelle Westeel et Thierry Claerr, Bibliothèques. Electre-Cercle de la Librairie, 2013, p. 115-171 (version auteur).

Sur le web, en effet, les bibliothèques ne sont pas seules. Pour être visibles, pour offrir à leurs publics des services qui répondent à leurs usages, elles doivent s'intégrer dans un vaste écosystème et interopérer avec des ressources, avec des métadonnées, qui leur sont extérieures. Plusieurs protocoles et formats répondent à différents besoins d'interopérabilité, depuis l'échange de données entre bibliothèques et la constitution de portails d'accès centralisés aux ressources numériques culturelles, jusqu'à l'ouverture la plus large sur les grands acteurs et usages du web, moteurs de recherche, réseaux sociaux, web mobile. Mais la diffusion comme la réutilisation des métadonnées sur le web repose *in fine* sur les mêmes principes de base : des métadonnées propres et de qualité, dans des formats normalisés, appuyées sur des identifiants uniques et stables et reliées à des référentiels partagés. Le web de données, encore expérimental il y a quelques années, ouvre à présent des perspectives inédites d'enrichissement des données de bibliothèques et de mutualisation de ces enrichissements, et place plus que jamais au cœur du web les métadonnées, et tout particulièrement les métadonnées de qualité produites par les bibliothèques.

Pauline Moirez et Stutzmann, Dominique. « Signaler les ressources numérisées : enrichissement, visibilité, dissémination ». In *Manuel de constitution de bibliothèques numériques*, édité par Isabelle Westeel et Thierry Claerr, Bibliothèques. Electre-Cercle de la Librairie, 2013, p. 115-171 (version auteur).

ANNEXE

5. ANNEXE : LE PROTOCOLE OAI-PMH (MICHEL FINGERHUT)

DE QUOI PARLE-T-ON ?

Le vocabulaire spécifique à tout domaine, principalement technique, permet des raccourcis utiles et d'alléger un style d'exposition, mais s'apparente parfois à un jargon incompréhensible aux personnes qui ne sont pas expertes du domaine en question (et qui, évidemment, le sont souvent d'autres domaines ayant leur propre langage).

On va donc aborder d'abord quelques expressions qui reviendront dans le cours du chapitre ; les définitions qui suivent ne visent pas à un formalisme rigoureux, mais plutôt à parler à l'intuition du lecteur. Elles sont donc volontairement simples (et donc parfois réductrices).

QU'EST-CE QU'UN PROTOCOLE DE COMMUNICATION ?

Un *protocole de communication informatique* est une spécification de l'ensemble des signaux ou de messages que peuvent échanger deux agents (ordinateurs, routeurs, logiciels…) pour accomplir une tâche ou un ensemble de tâches spécifiques : la façon dont ils sont codés, la logique des échanges et leur signification (quel signal peut venir en réponse à un autre signal)…

Ainsi, le protocole TCP/IP régit les communications entre ordinateurs sur l'Internet quel que soit le type de communication, tandis que les protocoles SMTP l'envoi, IMAP et POP la réception de courrier électroniques ; le protocole FTP le transfert de fichiers d'un ordinateur à l'autre ; le protocole HTTP celui de pages et de services sur le Web, etc.

Certains de ces protocoles nécessitent la présence d'autres protocoles sous-jacents. Ainsi, le Web (HTTP) fonctionne « au-dessus de » l'Internet, et donc de TCP/IP[169]. On retrouve cet empilement de protocoles dans d'autres domaines familiers : la téléphonie vocale, la télécopie ou l'ADSL nécessitent chacune l'utilisation du réseau téléphonique commuté et de son protocole de numérotation, d'envoi et de réception de signaux.

Le protocole OAI-PMH qui fait l'objet de ce chapitre utilise le Web comme infrastructure de communication : il envoie des requêtes par l'entremise du protocole HTTP, et reçoit des réponses sous une forme en général lisible aussi par un navigateur (HTML, XML, etc.).

Une conséquence utile de ce mode de fonctionnement est la possibilité de lancer ces requêtes à la main à l'aide d'URLs spécifiques, et de voir s'afficher le résultat à l'écran.

QU'EST-CE QU'UN CLIENT, UN SERVEUR ?

Pour faire simple, on appelle *client* tout agent qui initie une requête vers un agent distant qu'on dénomme alors *serveur*. Cette requête peut être unique, ou alors le début d'un échange plus long.

Ainsi : le logiciel de courrier électronique d'un internaute est le client de sa messagerie (en général située chez son fournisseur d'accès ou chez une tierce partie) ; le navigateur d'un internaute est le client du site qu'il va consulter – la requête consistant à demander à y voir des pages – et ce site en est le serveur[170].

Un même dispositif peut agir d'une part comme client, d'autre part comme serveur. C'est par exemple le cas des portails dont nous allons parler ici : d'un côté, ils vont, en tant que clients, chercher des métadonnées présentes dans des sites distants pour les indexer ; de l'autre, ils mettent à disposition, en tant que serveurs, ces informations collectées et éventuellement remises en forme à l'intention d'internautes ou d'autres portails.

INFORMATION, CONTENU, METADONNEES

Le terme *information* désigne ici tout message porteur de sens pour son producteur et son, ou ses consommateurs.

Par *contenu*, on désigne un « document » (texte, hypertexte, image fixe ou animée, enregistrement sonore…) qui ne soit pas uniquement descriptif d'un autre document.

Enfin, on appelle ici *métadonnée* tout document textuel[171] décrivant un contenu. Un tel document (qu'on appellera aussi indifféremment *notice*) peut servir à localiser le contenu qu'il décrit : pour un objet physique, en fournissant une identification du lieu (par exemple : une cote d'un livre sur les rayonnages d'une bibliothèque) ; pour un objet numérique, son adresse (URL) sur l'internet, qui permet donc d'y accéder.

Contenu et métadonnées sont donc tous des cas particuliers d'information. On évitera ici d'utiliser le terme *ressource* qui tend à désigner non pas uniquement des contenus (en général statiques), mais aussi des services.

MOTEUR DE RECHERCHE, PORTAIL, BIBLIOTHEQUE NUMERIQUE

On appelle communément *moteur de recherche* un dispositif qui indexe des contenus disponibles sur le Web (texte, image, son…) qu'il explore périodiquement, et qui fournit le moyen d'effectuer

---

[169] Mais aussi, entre autres, du DNS, protocole régissant les noms des ordinateurs sur l'internet – ou noms de domaine.

[170] Plus précisément, le logiciel du site qui fournit les pages Web (Apache, IIS, etc.).

[171] Certaines métadonnées – notamment dans le domaine du l'audio-visuel et du multimédia – ne sont pas forcément textuelles ou uniquement textuelles.

des recherches dans ses indexes. Les résultats de ces recherches fourniront les adresses de ces contenus là où le moteur les a trouvés, et qui permettent d'y accéder.

On appellera ici *portail* un dispositif qui indexe des métadonnées qu'il a récoltées sur le Web, et qui fournit, lui aussi, le moyen d'effectuer des recherches dans ses indexes. Les résultats seront en général une liste de métadonnées, qui comprendront ou non la localisation des contenus qu'elles décrivent. Europeana[172] est l'exemple d'un tel dispositif : il indexe les métadonnées d'un grand nombre de partenaires (dont toutes – c'est le choix d'Europeana – comprennent un lien vers un contenu numérique correspondant).

Europeana n'est pas une *bibliothèque numérique* comme le sont Gallica[173], HathiTrust[174] ou l'Internet Archive[175] : ces dernières indexent des contenus (qui se trouvent en général dans leurs propres bases et non pas ailleurs sur l'Internet, sauf exceptions), autant en texte intégral qu'à l'aide de métadonnées fournies avec ces contenus. Elles fournissent, bien évidemment, le moyen d'y effectuer des recherches, dont les résultats donneront directement l'accès aux contenus correspondants, voire même à la page et la ligne comprenant des termes inclus dans la recherche.

LES APPROCHES A LA RECHERCHE D'INFORMATIONS EN LIGNE

Si la prolifération des informations disponibles en ligne[176] a suscité, dès le début des années 1990, le développement de moteurs de recherche destinés à fournir un service centralisé, c'est bien avant le déploiement du Web sur l'Internet[177] que Z39.50, protocole d'échange interbibliothécaire, a commencé à être conçu[178] et a servi à réaliser des catalogues collectifs virtuels (qui sont en fait des portails, selon la terminologie que nous venons de voir).

Ces deux approches visent, l'une comme l'autre, à fournir un service centralisé de recherche d'informations dispersées[179]. Mais elles diffèrent sur plusieurs aspects fondamentaux :

WEB VISIBLE VS. INVISIBLE

Les moteurs de recherche ne font que suivre des liens hypertextuels présents dans la page qu'ils sont en train d'analyser et qui leur fournissent l'accès explicite à d'autres pages ; ils le font par l'entremise d'une seule requête – qui correspond au *clic* qu'effectue un être humain sur un lien hypertextuel – et ils ne peuvent donc accéder en général à des contenus qui nécessitent de remplir un formulaire de recherche et qui sont donc principalement stockées dans des bases de données[180] (constituant le Web profond, cf. note 176).

---

[172] Disponible à l'adresse <http://www.europeana.eu/>.

[173] Disponible à l'adresse <http://gallica.bnf.fr/>.

[174] Disponible à l'adresse <http://www.hathitrust.org/>.

[175] Disponible à l'adresse <http://www.archive.org/>.

[176] Au 1er novembre 2011, le Web indexable par les moteurs de recherche contenait plus de 12 milliards de pages, selon le site WorldWideWebSize.com. Ce chiffre est très inférieur au nombre de pages Web non indexables (constituant le Web dit profond ou invisible).

[177] Le Web est apparu en 1990 avec le développement, par Tim Berners-Lee, d'un serveur HTTP (le protocole d'échange qui constitue le Web), et d'un navigateur correspondant. Si l'Internet (identifié par le protocole d'échange TCP/IP) date de 1982, l'existence de réseaux qui l'ont précédé et qui y ont fusionné remonte aux années 1960.

[178] Cf. *The ANSI/NISO Z39.50 Protocol: Information Retrieval in the Information Infrastructure*, disponible à l'adresse http://old.cni.org/pub/niso/docs/z39.50-brochure/50.brochure.part05.html (vérifié le 2/11/2011). Sous l'influence prédominante des protocoles du Web, Z39.50 a évolué (ZING, SRU/SRW…).

[179] À l'instar de la Bibliothèque numérique européenne.

[180] Ce n'est pas toujours le cas : certaines pages sont par exemple protégées par des codes (appelés « captcha ») destinées à éviter les *spams*. À l'inverse, certaines pages provenant de bases de données peuvent

Les protocoles à l'instar de Z39.50 sont destinés à permettre l'interrogation programmée et à distance de bases de données, et donc d'accéder à ces contenus « cachés » pour les moteurs de recherche : ils possèdent en général un vocabulaire fort complexe de requêtes pour ce faire.

INDEXATION SELECTIVE OU NON

La requête servant aux moteurs de recherche à récupérer un document disponible sur le Web est inconditionnelle : ils ne peuvent y rajouter des conditions ayant trait, par exemple, au contenu de la page (son auteur, son sujet…). Ce n'est qu'après avoir reçu le document qu'ils peuvent l'analyser – avec un taux de réussite plus ou moins bon – et choisir de l'indexer ou non : il est donc complexe de réaliser un moteur de recherche thématique (dont les références concerneraient un sujet particulier). Par contre, Z39.50 peut interroger les bases distantes (constituées de métadonnées) selon des indexes matière, par exemple.

STRUCTURATION DES INFORMATIONS

Les informations auxquelles accèdent les moteurs de recherche – les contenus – ne sont pas forcément décrites ou structurées[181], et nécessitent des analyses parfois très pointues pour permettre leur indexation, procédé qui cause la production implicite de métadonnées plus ou moins détaillées et précises pour ces contenus.

À l'inverse, Z39.50 permet d'accéder à des informations structurées ; mais il s'agit en général de métadonnées, par exemple de notices bibliothéconomiques d'un catalogue. Un service utilisant ce type de protocole d'interrogation à distance pour créer un portail doit faire face à la possible différence des schémas de structuration et de description des contenus correspondants d'une base distante à l'autre.

FRAICHEUR DES INFORMATIONS

Les moteurs de recherche constituent un index des contenus qu'ils ont visités par le passé. Les réponses qu'ils fournissent aux recherches qui y sont effectuées se font en généralement rapidement, puisqu'ils n'ont qu'à consulter leur propre base de données mais peuvent concerner des contenus qui ne sont plus à jour, voire qui ont disparu, selon que le moteur les aura visités récemment ou non.

À l'inverse, une interrogation lancée sur un service utilisant Z39.50 va se propager en temps réel vers toutes les bases cibles de ce service : ceci a l'avantage de fournir des réponses à jour, mais ce mécanisme peut se heurter à des problèmes de connexion, temporaires ou non, à certaines bases distantes avec pour conséquence une lenteur, voire une absence de réponse de ces cibles.

LA MONTEE IRRESISTIBLE DU PROTOCOLE OAI

Le protocole OAI-PMH (ou, plus communément OAI[182]) a émergé à la fin des années 1990 – donc après l'invention du Web – afin de fournir, au départ, des moyens simples de localiser et

---

être accessibles directement, pour peu que le lien qui mène vers eux comprenne la requête correspondante de façon explicite.

[181] Il est possible d'inclure des métadonnées dans des contenus, qu'ils soient textuels (certains champs de l'entête d'une page codée en HTML) ou non (par exemple, les champs ID3 présents dans les fichiers sonores codés en MP3 et fournissant des informations à propos de l'enregistrement).

[182] Acronyme de *Open Archive Initiative Protocol for Metadata Harvesting*. À ne pas confondre avec OAI-ORE (*Open Archive Initiative Object Reuse and Exchange*), destiné à fournir une méthode de description de données numériques composites.

d'accéder à des publications d'articles scientifiques qui commençaient à s'y multiplier[183]. Ces articles, disponibles en ligne, pouvaient déjà y être indexés par des moteurs de recherche, mais ce genre d'indexation, résultant d'une analyse en texte intégral, ne peut assurer d'identifier de façon précise les références bibliographiques de ce type de textes (auteur, co-auteurs, titre, date de publication, lieu…), nécessaires à la communication savante.

Ce protocole, résolument simple, définit comment un site Web peut exposer des métadonnées et permettre leur récupération par tout client intéressé. Il a ainsi permis de réaliser ArXiv[184], un catalogue de plus de 700.000 articles en ligne dans les domaines de la physique, des mathématiques, de l'informatique, de la biologie, de la finance et des statistiques[185]. HAL, l'archive ouverte pluridisciplinaire mise en place par le CNRS[186], y verse automatiquement ses référence.

Le succès de ce protocole se mesure au fait que ses usages se sont trouvé être bien plus généraux que la réalisation de catalogues de publications scientifiques : le portail Europeana (cf. note 172) qui comprend plus de 10 millions de notices concernant des contenus culturels numériques disponibles dans des bibliothèques, des archives, des fonds audiovisuels et des musées, est constitué à l'aide de ce protocole. Ces organismes ont dû donc mettre en place, de leur côté, ce protocole pour permettre à Europeana – et à tout autre « client » – de récupérer ainsi les métadonnées présentes dans leurs catalogues ou leurs bases de données.

Une autre dimension de ce succès est son emploi pour fédérer des métadonnées concernant d'autres informations que celles concernant des contenus numériques ou physiques : événementielles, par exemple, à l'instar du Portail de la musique contemporaine[187].

Enfin, tout site client récupérant des métadonnées de cette façon peut lui-même les mettre à disposition, ce qui permet de créer des réseaux – hiérarchiques ou non – de clients et de serveurs OAI. Ainsi, le moteur Collections du ministère de la culture et de la communication[188], qui a constitué sa base à partir des métadonnées de nombre d'organismes culturels français, les fournit de cette façon à Europeana.

On ajoutera que ce protocole, utilisé communément pour créer des portails, peut en fait aussi servir à créer des bibliothèques numériques à part entière. Car si le protocole OAI ne spécifie que la récupération de notices, rien n'empêche, techniquement, un portail basé sur OAI pour la récupération des notices de recopier ensuite les contenus référencés par ces notices pour les indexer aussi. Il faudrait pour cela que le portail OAI, après avoir récupéré et indexé des métadonnées, récupère et indexe les contenus décrits par ces métadonnées si elles en fournissent l'adresse[189]. Les barrières à ce type de développement sont principalement juridiques[190].

---

[183] Il s'agit par exemple de pré-publication, permettant à des chercheurs de diffuser rapidement et de façon informelle les résultats de leur travail en attente de leur publication dans un périodique papier (où parfois le délai entre l'envoi de l'article et sa parution se mesure en mois, voire en années). Il peut s'agir aussi d'auto-publication concernant des textes qu'un auteur ne publie qu'électroniquement.

[184] Disponible à l'adresse <http://fr.arxiv.org>.

[185] Ce n'est pas une coïncidence : le protocole a été co-conçu et développé par Carl Lagoze de l'université Cornell aux Etats-Unis (avec Herbert Van de Sompel, du laboratoire de Los Alamos), et arXiv.org a été créé et est hébergé à la bibliothèque de cette université.

[186] Disponible à l'adresse <http://hal.archives-ouvertes.fr>.

[187] Disponible à l'adresse <http://www.musiquecontemporaine.fr/>.

[188] Disponible à l'adresse <http://www.culture.fr/fr/sections/collections/moteur_collections>.

[189] On avait proposé une méthode alternative pour ce faire déjà en mars 2005, cf. le texte disponible à l'adresse <http://blog.le-miklos.eu/?page_id=278>.

[190] Les droits de propriété intellectuelle, et notamment le droit de reproduction (ce à quoi s'apparenterait la recopie par le portail – même uniquement aux fins d'indexations – d'un contenu mis en ligne ailleurs).

Pauline Moirez et Stutzmann, Dominique. « Signaler les ressources numérisées : enrichissement, visibilité, dissémination ». In *Manuel de constitution de bibliothèques numériques*, édité par Isabelle Westeel et Thierry Claerr, Bibliothèques. Electre-Cercle de la Librairie, 2013, p. 115-171 (version auteur).

ENTREPOT, MOISSON ET LEURS PRINCIPALES CARACTERISTIQUES

On appelle *entrepôt* l'ensemble des métadonnées qu'un serveur met à disposition par l'entremise du protocole OAI-PMH[191]. La récupération de ces métadonnées par un client s'appelle *moisson* ou *moissonnage*.

En voici leurs principales caractéristiques.

UNICITE DES NOTICES

Toute notice présente dans un entrepôt doit posséder un identifiant unique à cet entrepôt, et de préférence, unique à l'organisme, voire unique sur l'Internet, mais ce n'est pas une condition imposée par le protocole OAI.

Cet identifiant doit lui être associé durant toute sa vie (voire même après son éventuelle suppression) : cette condition assure que tout changement (correction, précision…) apporté ultérieurement à cette métadonnée – y compris sa suppression – puisse se répercuter correctement chez tous les sites qui la moissonnent périodiquement. En outre, si ces derniers sont à leur tour moissonnés, et qu'une notice parvienne ainsi par plusieurs chemins détournés dans un même portail, ce dernier pourra effectuer un dédoublonnage (sous réserve que les moissonneurs intermédiaires aient veillé à préserver l'identifiant d'origine) et n'en garder qu'une copie[192].

La syntaxe des identifiants doit se conformer à celles des URI[193], très générale. OAI recommande l'utilisation d'une forme plus spécifique. Pour faire simple, voici quelques exemples qui s'y conforment :

oai:lcoa1.loc.gov:loc.music/sm1819.360010
oai:bnf.fr:catalogue/ark:/12148/cb390007325
oai:dcmi.ischool.washington.edu:article/13
oai:ircam.fr:programmation:5

Voici les principales caractéristiques de ce format :

- il est composé d'une chaîne de caractères sans espaces, ponctuée par le deux-points ;
- la première composante en est le mot oai, suivie du nom de domaine ou d'un sous-domaine de l'organisme détenteur de l'entrepôt (loc.gov est la Bibliothèque du Congrès) ;
- les composantes suivantes varient au choix de l'organisme, et utilisent souvent la barre inclinée (ou le deux-points) pour structurer cette partie de l'identifiant ou dénoter l'organisation hiérarchique de ces métadonnées en interne chez eux.

LE FORMAT DES METADONNEES DANS UN ENTREPOT

[191] Ou, plus précisément, le point d'accès – sous forme d'URL – du serveur OAI permettant de récupérer ces notices.

[192] Et ce même si cette notice est arrivée dans des formats différents (par exemple, en Dublin Core par un chemin, en MODS par un autre).

[193] Acronyme de *Uniform Resource Identifier*, schéma très général servant à identifier des ressources en réseau. Un cas particulier en est l'URL (Uniform Resource Locator), qui fournit en fait l'adresse de ressources sur le Web. Un autre en est l'adresse électronique d'une personne.

```
<record>
  - <header>
      <identifier>oai:bnf.fr:catalogue/ark:/12148/cb410000954</identifier>
      <datestamp>2010-06-09</datestamp>
      <setSpec>catalogue:collections:musique</setSpec>
      <setSpec>catalogue:edition:musique</setSpec>
      <setSpec>catalogue:partitions</setSpec>
    </header>
    <metadata>
      <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
      http://www.openarchives.org/OAI/2.0/oai_dc.xsd"
      xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
      xmlns:dc="http://purl.org/dc/elements/1.1/"
      xmlns:oai_dc="http://www.openarchives.org/OAI/2.0/oai_dc/">
        <dc:identifier>http://catalogue.bnf.fr/ark:/12148/cb410000954/description</dc:identifier>
        <dc:title>Suspens : [instrumental] / musique, Maïdi Roth, Franck Pilant & Nathalie
          Pilant</dc:title>
        <dc:creator>Roth, Maïdi (1970?-....). Compositeur</dc:creator>
        <dc:creator>Pilant, Franck. Compositeur</dc:creator>
        <dc:creator>Pilant, Nathalie. Compositeur</dc:creator>
        <dc:publisher>Universal music publishing (Paris)</dc:publisher>
        <dc:date>2000</dc:date>
        <dc:description>Cotage : U. M. P. 3966</dc:description>
        <dc:type xml:lang="fre">partition musicale</dc:type>
        <dc:type xml:lang="eng">score</dc:type>
        <dc:type xml:lang="eng">text</dc:type>
        <dc:rights xml:lang="fre">Catalogue en ligne de la Bibliothèque nationale de France</dc:rights>
        <dc:rights xml:lang="eng">French National Library online Catalog</dc:rights>
      </oai_dc:dc>
    </metadata>
</record>
```

Un entrepôt OAI doit réagir à toute requête qui lui provient en renvoyant sa réponse codée en XML. La réponse pouvant être fort longue (par exemple lors de la moisson d'un nombre important de notices), elle peut être fournie graduellement, par paquets de 50 ou 100 notices par exemple, accompagnés d'un *jeton de continuation* (*resumption token*, en anglais) permettant au moissonneur de demander le paquet suivant.

Les notices fournies en réponse comprennent deux parties :

1. *Un entête*, qui contient, entre autres, l'identifiant unique de la notice ainsi que la date de sa mise à jour, information primordiale pour le bon fonctionnement des moissons. Dans la Figure 1, il s'agit de la partie entre les balises <header>…</header>.
2. *Le corps*, qui comprend les métadonnées elles-mêmes. Ces métadonnées doivent structurées selon un ou plusieurs formats précis et destinés à permettre l'identification aisée et univoque des champs constituant ces métadonnées : auteur, titre, date, etc. Elles se trouvent entre les balises <medatada>…</metadata>.

Le protocole OAI impose l'obligation d'utiliser au moins le format Dublin Core[194] pour coder les métadonnées. Un entrepôt peut être conçu pour proposer *aussi* des formats alternatifs (tels que MODS[195], EAD[196] ou EDM[197]), plus précis et/ou plus expressifs pour les besoins de la

[194] Disponible à l'adresse <http://dublincore.org/>.
[195] Format dérivé de MARC, disponible à l'adresse <http://www.loc.gov/standards/mods/>.

structuration fine de l'information. S'il le fait, une même métadonnée peut donc être récupérée

```
<record>
    <header>
        <identifier>oai:bnf.fr:catalogue/ark:/12148/cb410000954</identifier>
        <datestamp>2010-06-09</datestamp>
        <setSpec>catalogue:collections:musique</setSpec>
        <setSpec>catalogue:edition:musique</setSpec>
        <setSpec>catalogue:partitions</setSpec>
    </header>
    <metadata>
        <telap:record xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:dc="http://purl.org/dc/
        xmlns:telap="http://catalogue.bnf.fr/namespaces/TEL_ApplicationProfile"
        xmlns:tel="http://krait.kb.nl/coop/tel/handbook/telterms.html" xmlns:mods="http://www.loc.gov/mod
        xmlns:dcterms="http://purl.org/dc/terms/">
            <dc:title>Suspens : [instrumental]</dc:title>
            <dc:contributor tel:role="Composer">Roth, Maïdi (1970?-....)</dc:contributor>
            <mods:typeOfResource>notated music</mods:typeOfResource>
            <dc:type xsi:type="dcterms:DCMIType">Text</dc:type>
            <dcterms:extent>[3] p.</dcterms:extent>
            <dc:identifier>cotage : U. M. P. 3966</dc:identifier>
            <dc:publisher>Universal music publishing (Paris)</dc:publisher>
            <dcterms:issued>2000</dcterms:issued>
            <mods:location>Bibliothèque nationale de France</mods:location>
            <mods:location>751131015</mods:location>
            <tel:recordURL xsi:type="tel:URL">http://catalogue.bnf.fr/ark:/12148/cb410000954</tel:recordURL>
        </telap:record>
    </metadata>
</record>
```

sous des formats distincts, au choix du moissonneur.

Dans la Figure 1, on peut voir que la balise <dc:creator>, destinée à indiquer les « créateurs » de la ressource décrite ici, ne permet de structurer (informatiquement parlant) l'information : le prénom, le nom et la fonction de la personne sont indiquée dans cette même balise. Par contre, un format tel que MODS permet de signaler dans une balise distincte la fonction d'une personne, et de l'associer de façon unique à son nom (indiqué dans une autre balise), ce qui en facilitera l'indexation automatique.

Dans la Figure 2, où la même notice est présentée au format requis par TEL[198], on voit que l'indication du nom et du rôle se fait autrement, en utilisant toujours une balise Dublin Core – en l'occurrence <contributor> mais en l'enrichissant d'un attribut provenant de TEL et appelé role avec comme valeur composer[199]. On peut apercevoir aussi quelques balises MODS qui s'y rajoutent.

Un même agent peut, de son côté, moissonner un entrepôt dans un format, et un autre différemment : c'est le cas du Portail de la musique contemporaine, qui moissonne ses partenaires en MODS, mais doit utiliser Dublin Core pour moissonner la Bibliothèque nationale de France, qui présente ses entrepôts uniquement dans ce format. Dans la Figure 1, on peut constater que

---

[196] Format destiné principalement aux métadonnées d'archives, disponible à l'adresse <http://www.loc.gov/ead/>.

[197] Le modèle sémantique d'Europeana, disponible à l'adresse <http://group.europeana.eu/web/europeana-project/technicaldocuments/>.

[198] Destiné à la moisson par la Bibliothèque européenne, disponible à l'adresse <http://www.theeuropeanlibrary.org/>.

[199] Ce terme fait sans doute partie d'une liste de valeurs contrôlées définies par TEL, et qui assure que tous les rôles qui lui seront indiqués par ses partenaires européens le seront de façon uniforme et sans ambiguïté.

les métadonnées sont effectivement présentées dans ce format (le oai_dc qui suit la balise <metadata> signifie Dublin Core).

MOISSON INCREMENTALE

La seule façon qu'un site a de maintenir à jour les métadonnées qu'il a moissonnées, c'est de les remoissonner périodiquement ; plus souvent elle se fera, plus les informations seront à jour[200].

Or comme les entrepôts peuvent contenir un nombre très important de notices – c'est le cas par exemple pour ceux de la Bibliothèque nationale de France, de l'INA, d'Europeana… – un mécanisme permet au moissonneur de spécifier à l'entrepôt qu'il ne souhaite récupérer que les notices créées ou modifiées depuis une certaine date (logiquement : celle de son précédent passage).

Ainsi, si la valeur d'un champ a été corrigée ou un champ rajouté à une ancienne notice, sa date de modification (présente d'ailleurs dans son entête, cf. la balise <datestamp> dans la Figure 1) reflétera ce changement. En conséquence, cette notice sera présentée de nouveau à la moisson, pour peu que la date de début mentionnée dans la requête précède la date de création ou de modification de la notice.

Le protocole OAI permet – sans l'imposer – de signaler aussi la suppression d'une notice : elle sera remplacée dans l'entrepôt par une sorte de notice fantôme[201] dont la date de modification est en fait celle de la suppression, sorte d'avis de décès dont le moissonneur pourra se servir pour effacer les métadonnées correspondantes dans sa propre base.[202]

MOISSON SELECTIVE

Un entrepôt peut choisir de marquer ses notices d'une ou plusieurs étiquettes destinées à indiquer une caractéristique particulière, par exemple le type de support des contenus qu'elles décrivent (texte, audio, vidéo…), leur genre (prose, poésie, fiction…), la période (renaissance, baroque…), leur provenance (fonds). Ces étiquettes – des chaînes de caractères arbitraires – sont enregistrées dans les entêtes des notices.

On appelle toutes les notices portant une de ces étiquettes un *ensemble* ou *lot* (ou *set*, en anglais). Une même notice, pouvant porter plusieurs étiquettes, peut donc appartenir à plusieurs lots.

Quand bien même ces caractéristiques peuvent déjà être précisées dans le corps de la notice, ce marquage a une fonction tout à fait différente : il permet de ne moissonner qu'un lot particulier et non pas tout l'entrepôt, sans avoir à préciser autre chose que l'identifiant du lot (qui est identique à l'étiquette en question).

Exemple : l'entrepôt OAI du catalogue de la Bibliothèque nationale de France comprend un nombre considérable de lots. En voici les identifiants de quelques-uns d'entre eux :

- catalogue
- catalogue:archives
- catalogue:archives:articles

---

[200] Un entrepôt n'a pas, actuellement, le moyen – du moins via le protocole OAI – d'envoyer une « alerte » à un moissonneur pour l'informer de la présence de nouvelles notices ou de notices modifiées depuis son dernier passage.

[201] Indiquée par le statut *deleted* dans l'entête de la notice.

[202] Dans le cas où un entrepôt n'effectue pas ce type de signalement, ses divers moissonneurs peuvent se débarrasser des copies de notices disparues en remoissonnant tout l'entrepôt en question après en avoir effacé l'ancienne copie de chez eux, ou en lui demandant la liste des identifiants de toutes les notices courantes dans l'entrepôt pour comparer à ceux des notices qu'ils ont moissonnées jusqu'ici, et ainsi pouvoir effacer, une à une, celles qui sont « en surplus » (et donc qui ont été supprimées dans l'entrepôt).

- catalogue:archives:contrats
- catalogue:archives:correspondances
- …
- catalogue:av *(« av » pour « audiovisuel »)*
- catalogue:av:bruitages
- catalogue:av:documentselectroniques
- …
- catalogue:edition:musique
- …
- catalogue:fonds:muscont *(« muscont » pour « musique contemporaine »)*
- …

On voit là sans besoin d'explication particulière que ces lots représentent une structuration hiérarchique des notices de la BnF. Le lot catalogue:archives comprendra *toutes* les notices de manuscrits modernes et d'archive, tandis que le lot catalogue:archives:contrats n'en comprendra que celles des comptes et contrats.

Enfin, on précisera qu'un site n'a pas l'obligation de moissonner un lot particulier d'un entrepôt ; s'il n'en mentionne aucun lors de sa requête de moisson, il obtiendra toutes les notices de l'entrepôt (sous réserve d'autres contraintes, telles les dates de dernière modification).

LES REQUETES

Les six requêtes qu'un moissonneur émet vers un entrepôt sont constituées d'un *verbe* et, selon le cas, de paramètres servant à restreindre le champ de la requête.

Ces requêtes peuvent aussi être envoyées « à la main », par l'entremise d'un navigateur ; elles comprennent l'adresse de base de l'entrepôt, suivi du verbe et éventuellement de ses paramètres.

Par exemple, dans la requête :

http://catoai.bnf.fr/oai2/OAIHandler?verb=ListRecords
&set=catalogue:edition:musique&metadataPrefix=oai_dc

(qu'il faut imaginer tenir sur une ligne et où l'on a souligné les mots réservés) :

- http://catoai…/OAIHandler est l'adresse de base de l'entrepôt ;
- ? indique que tout ce qui suit n'est plus une adresse, mais des informations destinées à l'entrepôt ;
- verb= sert à préciser le verbe (en l'occurrence : ListRecords) ;
- & sert à séparer les paramètres qui suivent ;
- set= précise le nom du lot que l'on souhaite moissonner ;
- metadataPrefix= indique le format auquel on souhaite le moissonner (en l'occurrence : Dublin Core).

Voici les verbes en question.

REQUETES DE PRISE DE CONNAISSANCE DE L'ENTREPOT

Les requêtes suivantes sont en général utilisées une seule fois par un moissonneur, au cours de son paramétrage initial pour accéder pour la première fois cet entrepôt. Elles lui permettent de récupérer des informations de base sur l'entrepôt, son identité et sa structure, et ne seront plus utilisées lors des moissons périodiques.

IDENTIFY

Cette requête, destinée à obtenir l'« identité » de l'entrepôt, n'admet aucun paramètre supplémentaire.

Dans la réponse, le moissonneur y trouvera, entre autres :

- le nom (en langage humain) de l'entrepôt ;
- la date de la création (ou modification) la plus ancienne dans cet entrepôt ; il ne lui servira donc à rien de demander des notices antérieures à cette date ;
- la précision sur le traitement de notices effacées (cf. §0) ;
- l'adresse de courriel du gestionnaire de l'entrepôt.

Exemple d'utilisation de cette requête (pour voir le résultat, la copier dans son navigateur) :

http://catoai.bnf.fr/oai2/OAIHandler?verb=Identify

LISTMETADATAFORMATS

Ce verbe sert à obtenir la liste des formats sous lesquels les métadonnées sont fournies par l'entrepôt.

En principe, cette liste doit comprendre au moins le format Dublin Core (puisqu'il est imposé par le protocole OAI, cf. §0), codé, dans la réponse à cette requête, ainsi :

<metadataPrefix>oai_dc</metadataPrefix>.

Exemple d'utilisation de cette requête :

http://catoai.bnf.fr/oai2/OAIHandler?verb=
ListMetadataFormats

LISTSETS

Cette requête sert à obtenir la liste de tous les lots que l'entrepôt contient : leurs identifiants nécessaires à leur moisson, et leur nom en clair.

Exemple : http://catoai.bnf.fr/oai2/OAIHandler?verb=ListSets

REQUETES DE MOISSON

Les trois dernières requêtes sont destinées à obtenir une ou plusieurs notices présentes dans l'entrepôt.

LISTRECORDS

C'est le verbe le plus commun, celui destiné à obtenir les notices présentes dans l'entrepôt.

Il comprend un paramètre obligatoire, celui indiquant lequel des formats que propose l'entrepôt (cf. § 0) est celui où le moissonneur souhaite les recevoir. Il est nécessaire de le préciser, même si l'entrepôt ne propose qu'un seul format.

La forme la plus simple de cette requête est donc la suivante (pour le cas où l'on précise comme format Dublin Core) :

http://...?verb=ListRecords&metadataPrefix=oai_dc

Les autres paramètres, facultatifs ceux-ci, sont :

- set=…, le nom d'un lot qu'on souhaite moissonner (s'il n'est pas fourni, c'est tout l'entrepôt qui sera moissonné) ;
- from=… et/ou until=… , les dates la plus ancienne et la plus récente de modification de notices en cas de moisson incrémentale (cf. §0).
- resumptionToken=…, le jeton de continuation d'une moisson en cours (cf. §0), destiné à obtenir un paquet de notices supplémentaire ; s'il est précisé, les autres paramètres (de moisson sélective et/ou incrémentale) ne peuvent l'être, puisqu'il a cours lors d'une moisson en cours, pour laquelle ces paramètres ont éventuellement été précisés à son début.

Exemple : la requête suivante indique vouloir récupérer les notices de la BnF présentes dans le lot catalogue:edition:musique depuis le 5 novembre 2011, au format Dublin Core :

http://catoai.bnf.fr/oai2/OAIHandler?verb=ListRecords
&set=catalogue:edition:musique&metadataPrefix=oai_dc
&from=2011-11-05

La réponse à ce verbe consiste en une page en XML constituée d'une liste des métadonnées correspondantes (à l'exemple de celles reproduites ici en illustration), les unes à la suite des autres, dans un « enrobage » comprenant lui-même :

- un entête, qui précise la date et l'heure précises de réponse à la requête et quelle était la requête et ses paramètres ;
- un pied-de-page, précisant le nombre de notices qui ont été renvoyées et éventuellement un jeton de continuation, si toutes les notices n'ont pu figurer, vu leur nombre, dans cette page.

LISTIDENTIFIERS

Cette requête est identique à la précédente en ce qui concerne ses paramètres.

C'est la réponse qui en diffère : elle ne comprendra que les *entêtes* des notices, sans leur corps (les métadonnées).

Cet usage peut servir, par exemple, à mettre d'équerre un portail et l'entrepôt qu'il moissonne périodiquement, au cas où ce dernier n'indique pas qu'une notice est supprimée (cf. note 202).

GETRECORD

Cette requête sert à obtenir une notice unique. Les deux paramètres, obligatoires, qui doivent y figurer, sont :

- Identifier=…, l'identifiant de la notice en question ;
- metadataPrefix=…, le format auquel on souhaite l'obtenir.

POURQUOI ET COMMENT METTRE EN PLACE UN ENTREPOT OAI

Pour résumer : OAI est une solution logicielle à la fédération d'informations structurées. Un entrepôt sert à extraire des notices à partir d'une base de données (catalogue ou autres), à les mettre en forme (en Dublin Core et éventuellement en d'autres formats), à les attribuer à des lots distincts selon le cas, et à répondre aux requêtes d'un moissonneur. Ce type de dispositif est bien plus léger à mettre en œuvre que, par exemple, le protocole Z39.50.

En contrepartie, le moteur de recherche ainsi proposé au niveau d'un moissonneur n'effectue pas une recherche en temps réel dans les bases distantes concernées, mais dans une copie locale de leurs métadonnées, qu'il moissonne périodiquement. Ce service est plus précis – autant quant au choix des notices qu'il indexe et aux types de recherche qu'il y propose – qu'un moteur de recherche généraliste, mais se distingue de la recherche en texte intégral dans les contenus que ce dernier effectue.

Il est toutefois possible, comme on l'a vu plus haut (§0), d'enrichir un portail de façon à ce qu'il devienne une bibliothèque numérique à part entière.

La mise en œuvre d'un entrepôt peut se faire, en principe, de deux façons différentes :

1. par l'utilisation d'un module interne au logiciel (ou progiciel) gérant la base de données ; certains logiciels bibliothéconomiques et plus généralement, systèmes de gestion de bibliothèques numériques, en proposent ;

2. par la réalisation d'un logiciel externe[203], qui obtient les notices de la base en question de l'une de deux façons :
    a. par l'entremise de requêtes qu'il enverra selon que de besoin à la base pour en extraire les notices, ce qui lui permettra de les traiter à la volée ; ceci suppose que le logiciel gérant la base propose un volant de requêtes (appelé API en anglais – *application programming interface*) ;
    b. à la suite d'exports des notices effectués manuellement (ou de façon programmée) à l'aide d'une commande proposée par le logiciel gérant la base ; une fois ces exports finis, les notices pourront être traitées.

Dans l'un ou l'autre cas, la tâche la plus complexe est celle qui consiste à effectuer le paramétrage de la solution, et notamment à établir la correspondance entre le format des notices dans la base source (le catalogue, en général) et le, ou les, formats cibles. On en a vu un exemple concernant la façon d'exprimer un nom et un rôle (cf. §0) : les formats cibles étaient moins riches que ceux de la source.

Mais il se peut que ce soit l'inverse qui se produise : le format cible (c'est le cas pour EDM, cf. note 197###) peut être plus exigeant sur la syntaxe, voire la sémantique, des métadonnées, ce qui nécessite parfois de revisiter la base source pour l'enrichir ou mieux la structurer.

[203] Dans ce deuxième cas de figure, la réalisation du logiciel externe lui-même n'est pas particulièrement difficile : il suffit d'utiliser un logiciel libre à l'instar de phpOAI2, disponible à l'adresse <http://physnet.uni-oldenburg.de/oai/>.