# XML et les métadonnées

## Irruption dans la sphère bibliographique

« *Rien ne se perd, rien ne se crée, tout se transforme* » (Antoine Lavoisier)

**Si on prend pour angle d'observation l'essor fulgurant des grandes bibliothèques numériques, comme Gallica, Europeana ou Trove** – pour ne citer qu'elles –, on perçoit d'emblée que la voie d'un renouveau fondamental du signalement bibliographique est ouverte. Afin de garantir l'interopérabilité, non plus seulement entre les architectures techniques, mais aussi entre les formats de données, l'adoption de nouvelles règles et de bonnes pratiques communes s'est avérée indispensable. Clé de réussite pour une véritable intégration de l'ensemble des ressources numériques au sein des catalogues de bibliothèque, c'est sous les auspices non plus du catalogage, mais du *Metadata Management* que se profile cette (r)évolution.

### Focus sur la courte histoire des métadonnées

C'est tôt dans l'histoire du web, lors de la conférence WWW Geneva 94[1] au cours de laquelle fut par ailleurs annoncée la création du W3C, que la notion de métadonnées (en anglais : *Metadata*) est proposée par l'inventeur du web, Tim Berners-Lee. Données signifiantes destinées à faciliter l'accès au contenu informationnel d'une ressource numérique, les métadonnées sont en effet au cœur de l'architecture du web – et, de fait, au cœur du métier de bibliothécaire !
L'ISO les caractérise comme des « *données qui définissent et décrivent d'autres données* » et on peut, à ce titre, considérer les données bibliographiques au format ISBD ou leur traduction aux formats MARC comme des « *métadonnées (externes) qui s'ignorent* » !
Mais qu'il s'agisse de notices décrivant un document (métadonnées externes), de marqueurs introduits directement dans un fichier informatique (métadonnées encapsulées), ou de schémas prédéfinis pour intégrer tout à la fois le document lui-même et les jeux de métadonnées qui lui sont associés (métadonnées englobantes), les métadonnées ont un rôle commun : elles permettent le déploiement d'un système d'information apte à répondre à l'ensemble des besoins inhérents aux ressources numériques (identification, description, administration, gestion technique, des droits, de l'archivage pérenne…).

Autre dénominateur commun : le langage XML, grâce auquel les métadonnées ont trouvé un moyen d'expression privilégié, à tel point que l'une de leurs caractéristiques est justement d'être « encodées en XML ». En effet, ce langage extensible, standardisé par le W3C, répond aux besoins d'échanges de données entre systèmes d'informations hétérogènes. On dit communément d'XML qu'il est un langage pivot, par définition à l'aise dans la jungle des formats : indépendant car non propriétaire, il répond parfaitement à la nécessaire interopérabilité entre les matériels, les logiciels, les structures de données, les interfaces. Simple fichier texte décrivant des données en séparant la présentation et le contenu, destiné à être interprété par une machine (en anglais : *Machine Readable*), tout en restant lisible par un être humain, XML, de par l'extensibilité de sa structure, a l'avantage de pouvoir gérer l'arborescence des données et la notion d'héritage. Ainsi, en s'insérant dans des structures prédéfinies, qu'il s'agisse des DTD (*Document Type Definition*), vocabulaires définissant les modèles, ou des schémas, structures destinées à englober l'ensemble des jeux de métadonnées et contenus associés, les métadonnées écrites en XML peuvent rester conformes aux normes et formats « métiers », garants d'une certaine cohérence en termes de normalisation nationale et internationale, tout en adaptant leur structure pour définir les différents niveaux de granularité d'information et permettre leur portabilité.

Concernant les formats propres aux notices bibliographiques, le développement d'une logique reposant sur les métadonnées a mis très tôt en évidence le manque flagrant d'un modèle conceptuel auquel se référer : ce constat est à la base des « Spécifications fonctionnelles des notices bibliographiques » (en anglais : *Functional Requirements for Bibliographic Records*, FRBR), modélisation des informations contenues dans les notices bibliographiques, développées par un groupe d'experts de l'IFLA de 1991 à 1997, approuvées en 1997 par le Comité permanent de la Section de catalogage de l'IFLA, et publiées en 1998. Complété par un modèle analogue pour les données d'autorité (FRAD[2]), FRBR constitue l'ossature des nouveaux codes de catalogage, au premier rang desquels RDA (Ressources : description et accès), qui se positionne d'ores et déjà comme le futur code international.

### La mutualisation pour une plus grande dissémination

Pour définir les schémas de métadonnées indispensables au développement des bibliothèques numériques et à l'adaptation des catalogues, la Bibliothèque du Congrès a joué un rôle moteur en élaborant, dès 1999, le schéma MARCXML, pour permettre la conversion du format MARC21 (et consorts) en XML, puis les schémas MODS[3] et MADS[4], dédié aux autorités, et enfin, le schéma METS[5], format d'implémentation permettant de combiner différents jeux de métadonnées descriptives, techniques, de droits et de préservation. Avec METS, sur le modèle duquel sont encodées par exemple les données de la bibliothèque numérique Gallica, c'est l'ensemble des besoins de la vie du document numérisé qui est couvert.

Au niveau national, l'implémentation des schémas de métadonnées a également été mise en œuvre depuis quelques années, notamment pour les outils de signalement mutualisés gérés par l'ABES : on citera par exemple l'EAD (*Encoded Archival Description*) ou TEF[6], utilisés respectivement pour encoder les données patrimoniales dans le cadre de Calames (Catalogue en ligne pour les archives et manuscrits de l'enseignement supérieur), et celles des thèses électroniques *via* STAR (Signalement des thèses électroniques, archivage et recherche), outil sur lequel repose en partie « theses.fr », brique majeure pour la valorisation des thèses françaises.

Indépendamment de l'évolution des formats proprement bibliographiques, le Dublin Core est devenu « l'autre » schéma de métadonnées incontournable pour la gestion des ressources électroniques. Commandité par OCLC et NCSA (*National Center for Supercomputing Applications*), le Dublin Core tire son nom du groupe de travail qui s'est réuni en 1995, à Dublin (Ohio). Il s'agissait pour ce groupe d'experts issus d'univers professionnels divers, de définir un tronc commun élémentaire, utilisable pour la description de toutes ressources numériques. Schéma de métadonnées dites encapsulées, le Dublin Core comporte 15 éléments répartis au sein de 3 domaines : le *contenu* (titre, sujet, description, source, langue, relation, couverture) ; la *propriété intellectuelle* (créateur, éditeur, contributeur, droits) ; la *matérialisation* (date, type, format, identifiant). Par sa simplicité, le Dublin Core est apparu d'emblée comme le schéma le plus efficace pour structurer les entrepôts de métadonnées communs aux archives ouvertes, aux musées, aux archives audiovisuelles, au patrimoine écrit numérisé, ressources (jusqu'alors) décrites dans des formats spécifiques propres à chacune des institutions et, par définition, difficilement interopérables.
Le Dublin Core s'est en outre imposé comme le moyen le plus efficace dans le cadre notamment des programmes du mouvement *Open Access Initiative*, au travers par exemple du protocole OAI-PMH (*Open Archive Initiative - Protocol for Metadata Harvesting*). Le principe est simple : il s'agit de mettre à disposition les métadonnées concernées, en les exposant au sein d'entrepôts afin de les rendre récupérables par des moissonneurs. C'est de cette façon que les grandes bibliothèques numériques s'alimentent en données extérieures, multipliant les points d'accès (dissémination) et garantissant ainsi une plus grande diffusion. À titre d'exemple, nous citerons Gallica (qui moissonne les métadonnées exposées par une quinzaine de bibliothèques patrimoniales françaises), Europeana (qui moissonne les données des bibliothèques nationales européennes), mais aussi Dart Europe[7] (point d'accès vers les thèses européennes disponibles en texte intégral qui moissonne, par exemple, les données de TEL[8] et de STAR, entre autres dépôts institutionnels dédiés aux thèses).

***Mutualiser pour mieux disséminer.***
Phot. Barcoseb sur Flickr (CC BY-NC-ND 2.0)

À l'ère du *Metadata Management*, les opérateurs nationaux et internationaux (agences bibliographiques, bibliothèques nationales, consortia) se situent donc, de fait, en tant que « hub de métadonnées », élément phare du nouveau projet d'établissement de l'ABES.

*Christine Fleury*
Conservatrice à l'ABES
fleury@abes.fr

[1] Extrait de la conférence: http://www.w3.org/Talks/WWW94Tim/
[2] FRAD : *Functional Requirements for Authority Data*, 2009.
[3] MODS : *Metadata Object Description Schema*, publié en 2003 ; il s'agit d'une simplification de MARC 21.
[4] MADS : *Metadata Authority Description Schema*, publié en 2004.
[5] METS : *Metadata Encoding and Transmission Standard*, publié en 2001.
[6] TEF : Thèses électroniques françaises - Recommandation issue des travaux de l'AFNOR (CG46 /CN357/GE5) : http://www.abes.fr/abes/documents/tef/
[7] Dart Europe : http://www.dart-europe.eu
[8] TEL : Thèses en ligne - http://tel.archives-ouvertes.fr/