



Ateliers d'octobre 2016 - Rencontres des Éditeurs Publics

Atelier 1 du 12 octobre 2016 : Stratégies pour les métadonnées de l'édition publique

Définition d'un outil de production commun de métadonnées (formulaires pour renseigner les standards des métadonnées de l'édition publique).

Participants :

Vincent BOULET : Bibliothèque nationale de France Chef du service des référentiels

François ROUYER-GAYETTE : Centre national du livre directeur-adjoint

Elsa PERSONNAZ : CEREQ responsable du pôle éditions

Laurent CATACH : CNRS chercheur associé

Jean-Baptiste BOYER : EHESS responsable du service des publications

Delphine SAGET : École nationale des Chartes secrétaire d'édition

Brankica RADIČ : Éditions de la Villette assistante d'édition

Anne LA RIVIÈRE : OCDE documentaliste

Anne-Camille BAZZOLI : ONISEP documentaliste

Saliha HAMZIC : ONISEP administratrice technique éditoriale

Christine CRESPIN : Presses universitaires de de la Méditerranée responsable

Huguette RIGOT : INALCO responsable des presses de l'INALCO

Bernard KENKLÉ : Réseau Canopé chef de projet

Amélie ROUFFIN : Réseau Canopé documentaliste

Yolande MANZANO : Réunion des musées nationaux responsable d'édition

Stéphane COTTIN : Secrétariat Général du Gouvernement chargé de mission développement des systèmes d'information et valorisation des ressources documentaires

Sophie BEAUCHAMP : DILA (projet de plateforme unique éditoriale)

Anne BIET-COLTELLONI : DILA rédactrice (projet de plateforme unique éditoriale)

Florence BLANC : DILA directrice de projet

Georges COLLET : DILA secrétaire d'édition (projet de plateforme unique éditoriale)

Brigitte GORCE : DILA éditrice (responsable des métadonnées du logiciel métier édition Libris)

Guilhem MESTRE : DILA chef de produit Stratégie

Isabelle MOCELLIN : DILA Responsable du département de l'Information régaliennne et économique (notamment, éditions des Journaux officiels)

Thomas SAINT-AUBIN : DILA responsable du pôle Stratégie.

L'objectif de cette rencontre est de pouvoir partager sur l'usage des métadonnées dans l'édition publique.

Un tour de table des différents participants est organisé.

On peut noter les enjeux suivants :

- Pour le centre national du livre, il s'agit d'avoir de la qualité dans la production des métadonnées.
- Pour l'école des Chartes, un nouveau format d'édition est en train d'être lancé, il s'agit de pouvoir harmoniser les métadonnées et d'avoir une interopérabilité des données. Nous collaborons avec la MRSH de Caen et leur pôle Document numérique¹.
- Pour l'EHESS, il s'agit d'élaborer le futur site de vente des éditions de l'EHESS ce qui oblige à une refonte du catalogue en plus du site institutionnel qui reste.
- Pour l'INALCO, il existe deux objectifs pour les métadonnées : avoir un index automatique, avoir un alignement multilingue. Nous travaillons avec la société Mondeca depuis un mois et demi et la première chose affaire consiste à créer un index à plat avec des données sous forme d'URI et d'avoir du datalink.
- Pour l'OCDE, la question des standards et de la qualité des métadonnées est primordiale.
- Pour l'ONISEP, une nouvelle plateforme éditoriale est en train d'être mise en place : quelles sont les métadonnées à utiliser ? Un autre projet concerne le site sur les formations qui est en Open Content : il y a les établissements avec des métadonnées². On fait appel à d'autres référentiels. On promeut l'ouverture avec des données ouvertes.
- Pour le Réseau Canopé, il y a une refonte du système d'information général dont le SI documentaire (qui a des normes différentes du SI général). Une des briques de production que ce soit pour le SI général et le SI documentaire concerne les métadonnées. Il s'agit aussi d'avoir une identification des référentiels.
- Pour le Secrétariat général du Gouvernement, il y a une nécessité de se mettre d'accord sur un référentiel commun. Le SGG a trente bases à gérer sans référentiel commun. Un référentiel existe à l'étranger au niveau européen avec celui de l'Office des publications de l'UE³. Il y a des bonnes pratiques à partager. Il s'agit d'améliorer l'information publique pour le BtoC et le AtoA (A : administration).
- Concernant FeniXX⁴, société habilitée à la numérisation, diffusion, distribution des œuvres indisponibles du XXIème siècle, il se pose aussi la question de la définition des métadonnées.

Pourquoi cet atelier ?

Il est la suite de ce que nous avons fait en juin dernier et s'intègre dans le mouvement d'Open Data et la description des données.

¹ http://www.unicaen.fr/recherche/mrsh/document_numerique

² <http://www.data.gouv.fr/fr/organizations/office-national-d-information-sur-les-enseignements-et-les-professions/>

³ <http://publications.europa.eu/mdr/>

⁴ <http://www.fenixx.fr/>

L'ouverture des métadonnées dans l'édition publique est notre sujet. Quelle gouvernance et quel(s) référentiel(s) ?

On peut grossièrement diviser les éditeurs publics en deux sous-groupes : les éditeurs universitaires et les éditeurs de l'État. Ils n'ont pas la même approche et il faudra donc traduire des descriptions différentes.

Il faut considérer la gestion des métadonnées sous la forme des « communs », i.e. avoir une gouvernance mutualisée de ce bien commun. Il s'agit donc de faire mieux ensemble que chacun séparé.

Ce chantier peut prendre la forme :

- d'une description des entités nommées,
- de la production de la métadonnée.

Quel pot commun peut-on définir ?

Au stade de la production de métadonnées, cela semble difficile car les éditeurs utilisent des logiciels différents.

Utiliser un dépôt OAI est lui possible permettant un alignement des données et un moissonnage des métadonnées en raccordant un maximum de plateformes d'édition publique. Les savoirs et les connaissances seraient ainsi plus largement référencés sur cet entrepôt commun.

On envisage de faire un hackathon des éditeurs publics avec « Urisation » et test d'entrepôts au printemps prochain.

Les résultats de celui-ci seraient présentés comme des éléments concrets pour les REP 2017.

Vincent BOULET, Bibliothèque nationale de France

L'identifiant international ISNI⁵ sert à la fois de norme pour les bibliothèques, les archives, les acteurs commerciaux, les gestionnaires de droits, les éditeurs de recherche.

Il y a cependant un noyau commun de données.

La BnF est contributrice d'ISNI par l'intermédiaire du VIAF (Fichier d'autorité international virtuel)⁶ qui transmet à l'ISNI. Elle siège à l'instance de gouvernance de l'ISNI et a un demi-siège avec la British Library.

La BnF essaye d'aligner les ISNI des auteurs qui sont différents dans les différentes entités qui les gèrent, via la mise en place d'une surcouche de métadonnées. L'alignement automatique est contrôlé par des humains, des bibliothécaires de la BnF. À la British Library, des vérifications par sondage sont effectués.

⁵ http://www.bnf.fr/fr/professionnels/isni_informer.html

⁶ <http://viaf.org/>

La BnF a lancé un nouveau projet avec la ligne ISNI – DLE pour lequel un test a été réalisé la semaine dernière avec Dilicom et Hachette. Les tuyaux sont branchés et disponibles pour monter en puissance.

Dilicom a besoin d'un ISNI très en amont dans leur système pour leurs auteurs et donc bien avant le dépôt légal.

Dilicom et Hachette envoient des données structurées en Onyx à la BnF qui les compare avec sa base ISNI et renvoie une réponse. S'il s'agit d'un nouvel auteur, la BnF envoie une prénotice d'autorité dans la base ISNI. Soit la demande est refusée et dans ce cas, la notice existe déjà. Soit la demande est acceptée et il s'agit alors d'un nouvel identifiant.

La BnF réutilise les données pour faire les notices d'autorité.

La BnF gère à la fois les auteurs et les collectivités.

Actuellement il peut très bien y avoir 2 ISNI différents (un par entité qui gère les auteurs) avec un lien entre les 2 ISNI, pour une même personne.

Le fichier VIAF est constitué des différentes autorités internationales qui gèrent des auteurs.

Avec l'ISNI, l'idée est de gérer une couche unique (surcouche) entre les différentes entités.

Le format utilisé est un format ONIX.

Le processus de création d'un identifiant ISNI d'un auteur dure environ 8 jours et est disponible sur ISNI.org.

Aujourd'hui les $\frac{3}{4}$ des fichiers Nom de personnes de la BnF ont un identifiant ISNI, il s'agit de 1,3 million d'auteurs et environ 20% des collectivités. Pour ces dernières, les pratiques sont très diverses. Les entités fusionnent, il y a des changements de noms.

De plus, il existe des pratiques différentes entre les bibliothèques et les archives.

Quels sont les cadres communs ?

L'ISNI doit mettre d'accord ses contributeurs qui eux créent des normes. L'ISNI n'impose pas de normes de travail mais a besoin de retrouver différents éléments. Par exemple, le nom de l'auteur.

La BnF est agence d'enregistrement et peut fournir les notices existantes qui sont disponibles sur data.bnf.fr au format RDF et XML.

A-t-on un alignement des données lors de la production ou après ?

On a un peu les deux. À la production des données, un contrôle qualité est effectué pour avoir une bonne structuration des données. Une fois les données structurées, on envoie à la BnF ou directement à ISNI.

La BnF réalise beaucoup d'alignement manuel. C'est un service qu'elle procure.

La liste des agences d'enregistrement des ISNI est sur le site de l'ISNI. Ce sont des intermédiaires qui font un travail d'alignement des ISNI.

La BnF, la British Librairie, d'autres bibliothèques nationales font ce travail (Corée, Pologne,...), mais il y a aussi des organismes privés comme Electre, Ringgold qui le font. Bowker s'est retiré en revanche d'ISNI.

Il existe une concurrence sur ce sujet entre les bibliothèques nationales qui sont regroupées dans l'association VIAF, hébergée par l'OCLC dans l'Ohio aux États-Unis. Le VIAF couvre non seulement l'alignement des noms de personnes, mais aussi ceux des collectivités et des œuvres d'art. En revanche il n'y a pas de données géographiques.

Le VIAF est aligné avec Wiki Data qui alimente le Wikipédia anglophone puis francophone.

L'OCDE ne travaille pas encore avec la BnF. Mais d'autres éditeurs sont intéressés.

Les Archives nationales sont un autre utilisateur d'ISNI (Sudoc aussi). Il s'effectue un chargement des ISNI depuis quelques temps ainsi que pour quelques archives départementales.

L'idée est d'établir un fichier national d'identités, un projet très en amont où seraient recensés les besoins en allant plus loin que l'ISNI. L'ISNI est l'alignement de l'existant.

Il faut donc avoir un travail de coproduction. Quelles données retrouve-t-on tout le temps ? Il faut identifier les niches, les productions fines de données.

Il s'agit donc de mettre en place une architecture des données.

La BnF est un gros producteur avec le dépôt légal et ses acquisitions mais pas de toutes les données. Il existe d'autres institutions avec des champs d'expertise propres.

Le but est d'avoir des économies d'échelle au départ pour les institutions publiques.

Il existe aussi d'autres projets comme celui de nommage NOMA, le projet IdRef de Sudoc⁷. Il faudrait essayer de croiser tout cela.

Quelle gouvernance adopter ? Elle doit être scientifique, administrative et technique.

Il faut définir un mode de gouvernance de ce projet qui pourrait se faire dans le cadre du service public de la donnée de la loi Lemaire.

L'idée serait de mener une réflexion sur la gouvernance en 20¹⁷ en y incluant la DILA, l'ABES, les Archives de France en plus de la BnF, d'élaborer un contrôle qualité de l'outil et de fournir comme livrable du projet un modèle de données unique. Le problème sera d'harmoniser les sources et de définir le périmètre des données. Le prérequis est d'avoir des données

⁷ <https://www.idref.fr/autorites/autorites.html>

ouvertes. Une des conclusions de juin était la problématique de la double saisie.

Le hackathon pourrait être un bon cas d'école.

Il permettrait de mettre en place un système d'expertise partagée où des institutions sont nommées responsables d'un sous-domaine et un modèle collaboratif à l'image de Wikidata.

Dans un premier temps, on peut partir des besoins communs, on peut aligner l'existant des métadonnées, cibler un premier noyau récupérable en données ouvertes qui permettrait de réaliser un commun, selon le modèle des communs Open Law.

L'idée est de constituer une couche supplémentaire. Chaque acteur pourrait se spécialiser sur certaines ressources ou type de données et échanger en temps réel, grâce à la surcouche de métadonnées, en mode collaboratif (récupérer les données dont on a besoin).

Ces communs pourraient être réutilisées ensuite par tous. Il faudra mettre l'accent sur l'information, qui produit le noyau : nom structuré, identifiant international,... Il s'agit bien d'un travail éditorial.

Brigitte GORCE, DILA

Une présentation de Libris, logiciel du catalogue des ouvrages de la DILA, permet de montrer la façon dont sont renseignées les métadonnées à la fois pour les éditeurs, les rédacteurs, les commerciaux de la DILA ainsi que le site de vente La Documentation française et les weblibraires.

Libris a été créé en 2003, une 3ème version en cours de développement est maintenant arrêtée en raison de la mise en place du projet « Editio », qui a pour objectif l'implémentation d'une plateforme unique de production éditoriale multi-support, et de diffusion vers les librairies en ligne et le futur site de la Documentation française. Celui-ci rassemblera notamment les contenus des sites ladocumentationfrancaise.fr et vie-publique.fr et des contenus issus de nos publications propres et pour les administrations.

Pour la mise en place de cette plateforme de production, il faut procéder à une analyse des champs de métadonnées utilisées qui proviennent de trois sources : le catalogue de la Documentation française (Libris), la Bibliothèque des rapports publics et des discours, mis en ligne sur le site vie-publique.fr (base Gipsi)⁸ et les contenus divers du site. Il s'agira de trouver un cœur, un noyau commun de métadonnées à ces contenus divers, trouver et établir

⁸ <http://www.vie-publique.fr/discours/>

des correspondances, s'appuyer sur les mêmes référentiels. Ce travail commence par une mise à plat des métadonnées utilisées.

Cette plateforme a pour objectif d'être utilisable pour les éditions produites en propre par la DILA et ambitionne de proposer ses services aux administrations pour les documents édités en compte de tiers.

La plateforme devra permettre - pour la production éditoriale des ouvrages, comme Libris aujourd'hui - de gérer des contenus à la granularité plus fine afin de faciliter d'une part une diffusion sur le nouveau site, et d'autre part des recompositions pour la constitution de nouveaux produits (ce qui impliquera des possibilités d'interrogation et par conséquent de recherche sur les contenus entrés).

Libris gère les formats numériques : PDF, ePub. Des liens d'association sont faits entre la version papier et les versions numériques, entre l'ancienne version papier et la nouvelle. Les ISSN et les ISBN sont bien gérés pour le papier mais certains ISSN pour les documents numériques manquent.

Par ailleurs des travaux ont été entrepris sous Onyx 3 sur les balises, les champs obligatoires afin de compléter Libris pour les champs qui manquaient (comme le poids des numériques, le public visé, les droits territoriaux).

Conclusion

Il faudrait réfléchir sur la couche de métadonnées qui permettrait le stockage «OAI » et moissonnage pour réutilisation en vue de la création de service

Exemple pour un professeur, une recherche sur la métadonnée « laïcité » pourrait permettre de croiser l'ensemble des productions disponibles sur le net et papier.

Les services pourraient se développer sur la base du « commun » de la donnée, disponible pour tous.

Ces communs pourraient être ouverts aux imprimeurs comme JOUVE.

Il faut recenser l'ensemble des intervenants possibles, en passant par le SNE.

Prochains point en avril. Possibilité d'organiser un hackathon avec les étudiants d'Épitech (EPITA) qui doivent réaliser un mémoire de 4/5 année.

La BNF pourrait également participer à l'enrichissement des données juridiques.