Mémoire de diplôme de conservateur / janvier 2010

Diplôme de conservateur de bibliothèque

La qualité des métadonnées en sciences humaines et sociales dans l'édition électronique ouverte

François-Xavier BOFFY

Sous la direction de Benoît Epron
Directeur des études - ENSSIB

école nationale supérieure des sciences de l'information et des bibliothèques

Remerciements

Je remercie Benoît Epron, qui a toujours été présent dès que je faisais appel à ses avis et dont chaque remarque m'a été précieuse.

Ce travail doit beaucoup à Marin Dacos, dont les conseils m'ont guidé tant sur le plan théorique qu'au niveau pratique.

Ma gratitude va également à Delphine Cavallo, Nicolas Barts, Jean-Baptiste Bertrand et toute l'équipe du Cléo pour leur attention constructive.

Merci à Anne Durand pour son amicale aide logistique lors de mon séjour à Marseille.

Merci à Emma Bester pour son travail de lien entre bibliothèques et édition électronique ouverte.

Je ne peux enfin manquer de remercier pour leur patience et leur compréhension Irène, Tristan, Vincent et la toute petite Ariane.

Résumé :

Ce mémoire s'attache à analyser l'utilisation des standards de métadonnées dans l'édition électronique ouverte, dans le domaine des sciences humaines et sociales. Une typologie de ces standards est proposée, ainsi que des pistes pour améliorer la qualité des métadonnées dans l'édition scientifique. Le rôle des bibliothèques dans cette amélioration est ainsi souligné.

Descripteurs :
Métadonnées ; Normes ; Sciences humaines – Recherche ; Edition

Abstract :

This report analyses the use of metadata standards in open access publications about humanities. A possible classification of these standards is explained, with advices about how metadata quality could be improved in academic publication. The place of libraries in this improvement is stressed at the same time.

Keywords :
Metadata; Standards; Humanities – Academic research; Publishing processes

Droits d'auteurs

Table des matières

Introduction

Il est difficile, au moment de décrire le projet et la problématique qui sous-tendent ce travail, de ne pas le situer dans le contexte de sa réalisation. L'ENSSIB, en tant qu'institut de formation des personnels de bibliothèque et des spécialistes en sciences de l'information, doit tenir particulièrement compte des évolutions dans la façon de décrire les documents. Or depuis plusieurs années le mot de « métadonnée » est mis en avant lorsqu'il s'agit d'évoquer la description des données. C'est ainsi que la notion est devenue incontournable, par exemple, lorsqu'il s'agit d'esquisser le futur des catalogues[1].

A quelques jours d'intervalle, dans l'amphithéâtre d'une école formant les cadres de bibliothèque, on peut ainsi désormais entendre souligner par deux fois le rôle majeur des métadonnées dans l'avenir de la galaxie « Gutenberg », celle du livre, des revues et de la presse. Lors d'une journée d'étude intitulée *Horizon 2019 – Bibliothèques en prospective*, tout d'abord, c'est un éditeur du secteur « traditionnel » (même si son point de vue sur l'avenir du livre ne l'est pas), François Gèze (La Découverte), qui a souligné l'importance de produire des métadonnées de qualité, quel que soit le rôle dans la chaîne éditoriale, éditeur, diffuseur ou institution mettant les ressources à disposition[2]. Puis c'est le responsable du secteur patrimonial de la bibliothèque Sainte-Geneviève, Yann Sordet, qui a mis en valeur le travail de signalement du patrimoine par le biais de métadonnées de qualité, sans lesquelles les documents disparaissent ou ne prennent pas leur place dans un ensemble organique et historique.

Mais parallèlement un autre secteur, contigu à celui des bibliothèques, a lui aussi été profondément affecté par l'augmentation conjointe des accès et des informations. Le domaine de l'édition scientifique, dont les bibliothèques sont souvent les premières « clientes », a vu émerger rapidement la notion de métadonnée au point de la rendre semble-t-il incontournable dans la compréhension du fonctionnement de ce secteur éditorial. C'est pourquoi les questions que peuvent se poser les éditeurs de publications de recherche peuvent alimenter avec intérêt la réflexion actuelle du monde des bibliothèques à ce sujet.

Le souhait du Centre pour L'édition Electronique Ouverte (CLEO), éditeur de Revues.org, des carnets de recherches d'Hypothèses et de l'agenda des sciences sociales Calenda, de clarifier les potentialités des nombreux standards de métadonnées dans son domaine (l'édition scientifique en sciences humaines et sociales) rencontre donc le besoin, pour les conservateurs de bibliothèque, de mieux maîtriser les concepts et les processus à l'oeuvre dans la documentation électronique.

[1] (Wenz 2009)
[2] (Gèze)

Sur quoi peut-on fonder des préconisations en matière de standards de métadonnées ? Des préconisations « appliquées » ne sont pas séparables d'une analyse des processus de génération et d'échange de métadonnées. Se poser la question de ce qui fait la qualité d'un ensemble de métadonnées, en faisant le lien avec la qualité des métadonnées elles-mêmes, nous permettrait d'esquisser des orientations en matière de standards de métadonnées.

Il a semblé utile, dans cet objectif, de rappeler ou préciser succinctement la notion de métadonnée, puis de remettre en perspective le rôle de ces données particulières dans l'économie générale de l'information scientifique et de l'édition électronique. Le qualificatif d' « ouvert » doit également être mieux cerné car il sera particulièrement intéressant de mettre en relief son impact sur les préconisations présentes et à venir.

Il est logique pour un non-spécialiste d'asseoir ses recommandations sur des observations. La démarche adoptée sera donc d'observer les choix et les pratiques de quelques acteurs de la chaîne des métadonnées, producteurs, plates-formes de publication ou outils informatiques de médiation, mais aussi d'observer avec plus d'attention les nuances qu'apportent les types de documents édités en sciences humaines et sociales.

Sans décrire l'ensemble des standards, on pourra donc en faire une description organisée, qui rende compte aussi bien de leurs caractéristiques « absolues » que de leur impact réel dans le fonctionnement du domaine d'application étudié. Ce sera d'ailleurs l'occasion d'une première mention du rôle des bibliothèques dans la diffusion des standards de métadonnées.

A partir de cette analyse des pratiques en matière de métadonnées chez différents intervenants, il sera surtout possible de poser des jalons pour augmenter ou garantir la qualité de ces métadonnées dans le domaine des revues ou monographies en sciences humaines et sociales. Il sera inévitable de s'appuyer sur une analyse conceptuelle partiellement inspirée de la riche littérature sur le sujet. Mais des références précises à des cartographies de métadonnées éclaireront avantageusement notre propos et seront cohérentes avec le positionnement particulier des « digital humanities », entre réflexion et applications.

Le rôle des bibliothèques dans la création, la diffusion et la consommation des métadonnées pourra enfin être souligné. Ce sera donc l'occasion de mettre en valeur leur propre rôle dans l'augmentation de la qualité des métadonnées.

Contexte

LES METADONNEES

Le substantif de « métadonnée » est, à l'échelle historique, très récent. L'étymologie, parfois évoquée pour rendre compte du sens correspondant, est en fait une étymologie reconstruite puisque le préfixe grec « meta » était à l'origine le signe de la postériorité et non de la supériorité hiérarchique (la Métaphysique d'Aristote était le groupe des écrits suivant sa Physique, et non les écrits traitant des principes premiers du monde physique[3]).

Le « meta » au sens de « principiel » est donc récent; mais l'ensemble « métadonnée » semble profondément lié à notre ère technologique. On en trouve trace en 1969, mais le terme ne se diffuse (lentement) qu'à partir des années 1980. Une définition simple, satisfaisante pour notre propos, a été alors donnée par James Martin : les métadonnées sont, d'après lui, des « données sur les données » (« data about data »)[4].

Deux remarques s'imposent à partir de cette définition :

- « métadonnées » est un terme très générique recouvrant des réalités extrêmement nombreuses et diverses, donc parler des métadonnées en général tient au mieux un défi philosophique, au pire de la mission impossible.
- L'usage des métadonnées remonte à une date bien antérieure à l'invention du mot. Les codes postaux, les indications de page dans un codex, les registres paroissiaux ou les dictionnaires relèvent tous d'un processus de méta-information, même si on restreint celui-ci à la *structuration* de données sur des données.

Cela étant précisé, la métadonnée est actuellement et presque exclusivement pensée dans un contexte d'échanges numériques de données numériques ou analogiques. C'est aussi le cadre dans lequel nous travaillerons ici, puisque le domaine où nous étudierons la qualité des métadonnées est celui de l'information scientifique et de l'édition en sciences humaines et sociales.

De façon générale, on peut dire que la principale et première métadonnée manipulée n'est plus le titre ou l'auteur, mais l'adresse URL du document. Celle-ci est en effet une donnée, respectant une grammaire précise, et fournissant au moins une information sur le document. Cela n'est pas forcément visible par l'utilisateur; on garde la plupart du temps un attachement aux balises que sont l'auteur et/ou le titre, et si on demande à un être humain de proposer des métadonnées pour un document il y a de fortes chances pour qu'il commence par ces deux éléments (le sujet est aussi une question très présente à l'esprit lorsqu'il s'agit de décrire les documents).

[3] Voir les détails sur le Trésor de la Langue Française (http://atilf.atilf.fr/tlf.htm).
[4] (James Martin 1982, 127)

Une mesure permet de se rendre compte de la difficulté croissante à embrasser dans une seule étude tous les aspects des métadonnées : au 10/11/2009, le nombre de réponses à la requête « metadata » était environ de 34 300 000 dans Google recherche web (dont l'expérience nous a appris que bien qu'incomplet il était plus précis que Google Recherche de Livres). Au 10/12/2009, soit un mois plus tard, ce nombre de réponses était passé à 36 200 000, soit une augmentation de 5,5% en 1 mois.

Afin de circonscrire les questions sur les métadonnées, il est courant de les aborder sur le plan d'un « profil d'application », c'est à dire dans le cadre d'un champ disciplinaire et technique donné. Faire des règles de bonne conduite qui soient communes à la fois à la présentation des constructions aéronautiques, à la description des phénomènes météorologiques et à l'articulation des démonstrations mathématiques est possible en théorie, totalement vain en pratique. Un échange de métadonnées est une situation particulière de communication, or il n'y a pas de message qui puisse s'affranchir de son contexte d'énonciation. Nous y reviendrons, la solution d'une articulation des multiples « profils d'application » par le biais d'une démarche d'interopérabilité est certainement la meilleure, mais implique de prendre en compte de très nombreux paramètres, encore trop souvent négligés.

Si à première vue la recherche scientifique peut constituer un champ d'application pour une structure de métadonnées riche, les « branches » de la science sont elles-mêmes des sujets suffisamment vastes pour qu'existent des ensembles de métadonnées spécifiques pour chacune.

Les sciences dites « dures » et les technologies sont assez richement pourvues en travaux poussant très loin la réflexion et l'analyse. Les sciences humaines et sociales (SHS) se sont globalement intéressées plus tardivement à la question des métadonnées, à quelques rares exceptions près (on peut penser à la TEI, Text Encoding Initiative, mise en place à partir de 1983)[5]. Toutefois parallèlement à la révolution des pratiques de la recherche d'information et de la gestion bibliothéconomique, l'écriture des SHS se fait maintenant parfois aussi avec une prise en compte, dès l'origine, de la question des métadonnées.

C'est même l'expression d'un besoin dans le domaine qui a présidé à l'élaboration de ce mémoire d'étude. Depuis quelques mois, la notion de « digital humanities » se transpose dans l'épistémologie francophone pour désigner ce pan de la connaissance qui mêle nouvelles technologies et « humanités », sciences humaines et sociales. Milad Doueihi, dans *La Grande Conversion numérique*, décrit ainsi de façon très opératoire les ponts qui se jettent entre les professionnels de l'histoire, de la sociologie, de la philosophie, et le travail des ingénieurs de la connaissance, pour répondre aux besoins des nouveaux usagers qu'il nomme « numériciens »[6]. De la même façon, Corinne Welger-Barbosa emploie l'expression de « corpus instrumenté », très liée à la question des outils de manipulation du savoir que sont les métadonnées[7]. Cela souligne encore l'enjeu qu'elles constituent dans la circulation de l'information, et plus particulièrement dans l'édition (voir à ce sujet, entre autres, l'article de Guy Beaudry[8]).

[5] http://www.tei-c.org/index.xml.
[6] (Doueihi 2008)
[7] Expression tirée de (Equipe des rédacteurs de Calenda 2009)
[8] (Beaudry)

STANDARD, NORME, FORMAT

La norme a un sens précis qu'il n'est pas permis de négliger : sa validation se fait selon des processus de certification aux plans nationaux (AFNOR) ou internationaux (ISO) et qui lui donnent sa valeur. Mais à propos des métadonnées le fait de ne prendre en compte que les normes reviendrait à se couper de la plus grande part des éléments du domaine. C'est pourquoi nous parlerons à dessein de standards, au sens d'usages si bien établis et si répandus qu'ils se comportent de facto comme des normes. On pourra aussi étendre cette notion de standard aux règles cohérentes qui constituent une référence d'organisation et / ou de fonctionnement des objets numériques, mais n'ayant pour l'heure qu'une utilisation réduite.

Ainsi qu'on le verra, les métadonnées peuvent être proposées dans un fichier à part, ou dans un fichier mêlant données et métadonnées. Dans les deux cas on fera donc référence à des formats de fichiers. Là où la distinction est parfois difficile à faire, c'est que la plupart des formats correspondent de fait à une définition d'un ensemble de métadonnées. Par exemple, on peut parler du format BibTeX ou d'un fichier en BibTeX, d'un fichier Endnote (avec une extension « .enl ») ou d'une notice encodée avec le standard Endnote. Format peut donc être utilisé pour désigner ceux des ensembles de métadonnées qui ont une structuration et / ou une expression propre (voir *infra* la typologie proposée pour les standards de métadonnées).

« SETS » ET SCHEMAS, GRAMMAIRES ET VOCABULAIRES

Pour désigner un ensemble de « cases » dans lesquelles sont placées les métadonnées, l'anglais propose le très pratique « set », dont la traduction par « ensemble » est trop imprécise. « Schéma » est un équivalent encore meilleur, car il renvoie aussi à l'une des deux expressions privilégiée des règles de métadonnées pour les fichiers XML (on peut en effet les exprimer avec une DTD ou un XSD, XML Schema Description).

En revanche, le français propose le couple « grammaire » et « vocabulaire » pour décrire métaphoriquement mais justement le fonctionnement des standards de métadonnées : les standards peuvent définir tantôt des éléments à utiliser pour décrire les données « primaires », des « vocabulaires », tantôt des méthodes d'expression des descriptions, des « grammaires ».

Le terme « tag », très répandu, est utilisé dans deux acceptions, c'est pourquoi dans un souci de clarté nous l'utiliserons avec précaution. Il signifie en effet parfois étiquette apposée par les producteurs de documents (web 1.0) ou les utilisateurs de documents (web 2.0). On préfère volontiers parler dans le monde des bibliothèques de « mots-clés », d' « indexation » et « folksonomies », qui ne sont rien d'autre que des métadonnées descriptives du contenu. L'autre sens de tag est plus récent et moins rigoureux : il s'agit du nom des champs utilisés par un standard de métadonnées. Le « coverage » est par exemple, dans cette acception, l'un des 15 « tags » du Dublin Core non-qualifié (DCMI 2004). Le terme de « label » est synonyme de « tag » en ce sens, et plus avantageusement évocateur.

LE ROLE DES METADONNEES DANS L'INFORMATION ET L'EDITION

Il faut bien prendre conscience que la qualité des métadonnées n'est pas l'alpha et l'oméga de toute réussite dans le domaine de l'information en ligne. Google Books est par exemple souvent et à juste titre critiqué pour la qualité de ses métadonnées[9]. Son succès, qui semble aller au-delà de l'attrait de la nouveauté des premiers mois, montre que cette initiative, en se positionnant comme « moteur de recherche » et non comme « bibliothèque »[10], n'a pas besoin de description fine du contenu telle que le permettent par exemple les formats de la famille MARC. Evidemment, le poids économique et technique de Google n'est pas étranger à ce succès « par-delà » la qualité des métadonnées. C'est néanmoins l'occasion de rappeler qu'une offre éditoriale rencontre son public non seulement grâce à une démarche technique adaptée mais aussi en fonction du contexte social.

Dans le domaine de l'édition ouverte, que nous définirons plus précisément par la suite, la donne est certainement différente, dans le sens où la moindre imprécision dans le référencement d'un ouvrage, d'un article, peut se traduire par un manque à gagner que l'échelle, généralement plus modeste, n'autorise pas. L'autre champ d'exploitation des métadonnées que nous aurons à examiner est celui des publications scientifiques en SHS. Ici plus encore, la précision et la permanence de la référence sont nécessaires pour garantir une juste évaluation des scientifiques. En outre la « trouvabilité » des documents, pour reprendre un terme introduit récemment par Morville, implique une description efficace et (ou *mais*?) juste du contenu.

Cette importance s'est accrue avec le glissement d'un web statique à un web dynamique : les pages, les documents sont de plus en plus constituées de « flux » et non plus d'ensembles figés de données. C'est la conséquence logique du développement du phénomène de « computing in the clouds », c'est à dire du transfert sur des machines et réseaux distants des opérations et applications fonctionnant auparavant localement. Chaque flux doit être manipulé avec rapidité et sans ambiguïté pour intervenir au bon moment et pour le bon utilisateur. Il doit donc être pourvu d'identifiants uniques et de points d'accès.

Ce secteur éditorial est directement dans le type de questionnement sur les métadonnées décrit par Catherine Morel-Pair[11] :

> *La création de métadonnées de qualité portant sur les ressources, actions et acteurs a un coût qui s'ajoute à celui de la maintenance des ressources, même si elle y participe grandement. Une large intervention humaine majore évidemment ce coût ; il est intéressant d'automatiser le processus au maximum, dans les différentes étapes correspondant au cycle de vie des ressources. A la création, quels éléments descriptifs peuvent (et doivent) être apportés par l'auteur, sachant qu'il est souvent souhaitable d'intégrer une vision documentaliste complémentaire dans le work-flow ; quels éléments*

[9] (Nunberg 2009)
[10] Dans (Nunberg), voir la réponse de Jon Orwant.
[11] Dans (Morel-Pair 2007)

techniques et administratifs peuvent être extraits ? Pour des ressources acquises, comment réutiliser les descriptions existantes ? Quels éléments pourront être générés automatiquement ensuite au moment de la diffusion, de l'archivage, des migrations ?

L'EDITION « OUVERTE »

Même si l'expression se rencontre couramment, il est difficile d'opposer une édition qui serait entièrement « ouverte » à une édition qui serait absolument « fermée ». En effet la frontière n'est pas radicale, et jusqu'à récemment l'existence d'un modèle économique d'édition ouverte était un postulat, une forme de pari. Un pari semblable à celui de Pascal ? A la différence de l'existence de Dieu, certains, dont Revues.org, ont pu démontrer l'existence d'une viabilité de la « voie dorée » du libre accès aux documents.

Pour mémoire, ce travail s'intéresse essentiellement à une seule des branches de l' « open access », la « golden road » (voie dorée), celle qui garantit un libre accès par *publication* gratuite des informations scientifiques. Evidemment, la « green road » (voie verte), celle qui garantit un libre accès par *mise à disposition* gratuite des informations scientifiques, n'est jamais très loin des esprits des intervenants de ce secteur particulier de l'édition, très lié à la recherche et à l'enseignement supérieur.

Il serait tentant de définir l'édition ouverte par défaut, en énumérant ce qu'elle n'est pas. Elle ne serait pas:

- édition de textes et documents du domaine public ou libres de droits puisque ce qui est proposé est sous droits et porte des mentions de responsabilité
- édition de textes et documents payants puisque ceux-ci sont très majoritairement disponibles gratuitement
- édition à compte d'auteur puisque dans la majorité des cas ce ne sont pas les auteurs qui paient le droit d'être publiés, mais les revues; il y a en outre une politique éditoriale, et parfois même des processus de « peer reviewing » (validation du caractère scientifique du document par les pairs) qui sont antinomiques du processus de publication à compte d'auteur
- édition sans financement, non professionnelle, puisque ce ne serait pas un secteur éditorial; l'édition de chaque revue ou chaque livre a un coût qui n'est pas pris en charge par le seul éditeur
- édition institutionnelle, puisque sa politique éditoriale ne pourrait être « ouverte » dans ce cas
- édition traditionnelle incluse dans le circuit du papier car ses stratégies d'accès reposent sur la consultation à distance

L'édition électronique ouverte a un modèle économique particulier, qui repose en France sur 3 sources principales différentes :

- En tant qu'éditeur de revues ou de monographies, elle participe (non sans réticences de la part de certains intervenants de l'évaluation scientifique) à la validation de la recherche des auteurs. Mais comme la consultation n'est pas payante, les revues et / ou auteurs paient donc la plupart du temps pour être publiés (cette règle et son application varient suivant les éditeurs).

- Certaines plates-formes d'édition proposent une offre technique complémentaire, facturée aux producteurs de données qui souhaitent en bénéficier. Le concept de « fremium » (mélange de « free », libre, et « premium », caractéristique des offres de services supplémentaires) contracte en un mot la double stratégie éditoriale permettant de financer partiellement le gratuit avec des services payants.
- En tant que pionnières en matière de nouvelles technologies de l'information et d'infrastructures de la recherche, certaines plates-formes bénéficient du soutien des établissements de tutelle de l'enseignement supérieur ou des grands instituts de recherche (comme le CNRS *via* TGE ADONIS pour Revues.org)

Les métadonnées ont une importance toute particulière pour le secteur « ouvert » de l'édition électronique, puisque celui-ci repose sur l'identification précise des articles et monographies consultées et, corrélativement, sur la visibilité maximale des documents pour l'utilisateur à la recherche d'information. Qu'il s'agisse des sciences humaines et sociales ajoute un facteur supplémentaire : leur valorisation économique passe en effet par l'accès d'un public élargi, d'amateurs et de curieux, ne disposant pas toujours d'instruments de recherche très raffinés (à la différence des entreprises par exemple, qui offrent un débouché constant pour les conclusions des sciences dites « dures »).

C'est peut-être dans l'édition électronique ouverte que les concepts de « citabilité » et de « trouvabilité » (en anglais « findability ») ont le plus de poids. Mais par-delà ces caractéristiques techniques, il y a aussi et surtout un esprit qui préside aux choix stratégiques des éditeurs « ouverts ».

Marlène Delhaye[12] nous donne sans doute la meilleure description actuelle de l'édition ouverte en jouant sur la transitivité de l'adjectif :

> *« Il est désormais question d'une édition électronique ouverte :*
> - *sur ce qu'elle va devenir (ouverte à une réflexion commune)*
> - *qui s'appuie sur des logiciels en open source*
> - *à la lecture et à l'écriture*
> - *à tous les acteurs de la chaîne du livre (y compris auteurs et lecteurs). »*

Ces différentes bases notionnelles étant posées, il nous est possible d'étudier les usages des différents acteurs de l'édition électronique, de façon à (re)définir la qualité des métadonnées dans la publication de la recherche en sciences humaines et sociales.

[12] Elle cite les propos de Marin Dacos lors de la table-ronde de clôture de l'Université d'été du CLEO, dans son billet disponible à l'adresse http://marlenescorner.blogspirit.com/archive/2009/10/25/ue-cleo-table-ronde-2-2.html

Analyse : des usages aux conceptions

LES USAGES DES PRODUCTEURS DE METADONNEES

Les institutions autres que Revues.org

Nature

Nature est la revue qu'on pourrait dire de référence, « visible » jusque dans les médias traditionnels. Ce n'est pourtant plus une revue, mais une plate-forme complète de publication, avec de nombreux titres, des produits documentaires (newsletters, blogs,...), un outil de gestion bibliographique intégré (Connotea) et même, depuis quelques mois, des revues en libre accès[13]. La production de métadonnées de Nature Publishing Group (NPG) est très importante en volume.

Pour la stabilité et l'image de marque de *Nature*, le fait de proposer des standards éprouvés semble logique. En contrepartie, les moyens financiers à disposition peuvent permettre de pallier les défauts des formats insuffisants seuls ou d'assumer des développements coûteux. Mais concrètement *Nature* ne prend pas beaucoup de risque en matière de formats de métadonnées : les métadonnées principales des articles sont placées dans les balises « meta » du code HTML, avec du Dublin Core non-qualifié et du PAM (sigle à tiroir qui signifie PRISM Agregator Message, PRISM signifiant lui-même Publishing Requirements for Industry Standard Metadata; voir glossaire des sigles en annexe). Pour la gestion des PDF, *Nature* s'appuie, là encore sans grande surprise, sur XMP.

Un quatrième standard est envisagé, celui de Google Scholar, mais il n'a pas encore été implémenté par *Nature*, par manque d'information précise et stable. Ce point est important pour caractériser l'attitude de la plate-forme, qui anticipe sur l'adoption probablement large de Google Scholar, mais ne s'engage que lorsqu'un certain nombre de garanties sont réunies.

Le RSS (ce sigle a plusieurs développements possibles, voir glossaire des sigles) est une autre recette éprouvée que met en pratique *Nature*, avec une attention du détail puisque sont disponibles des fils pour les numéros, pour les articles, mais aussi pour recevoir les informations sur la plate-forme, pour récupérer les nouvelles concernant directement les bibliothécaires, ou enfin pour agréger tous les flux afin de faciliter la syndication dans

[13] http://www.nature.com/.

un outil d'agrégation (du genre de Google Reader ou Netvibes). Un fichier OPML[14] complète cette offre.

Par ailleurs, une métadonnée particulière d'identification est mise en avant par *Nature*, le DOI (Digital Object Identifier). La mention particulière qui en est faite, à comparer à l'indication sans emphase de l'ISSN électronique par exemple, montre le rôle du DOI dans l'édition électronique actuelle et à venir. L'identification pérenne, vers laquelle tend l'initiative du DOI de CrossRef, est en concordance avec l'image que Nature Publishing Group souhaite sensiblement suggérer aux utilisateurs de son portail, celle de données scientifiques validées, certifiées, sûres.

L'exemple de *Nature* nous montre incidemment que la qualité des métadonnées se mesure selon leur **stabilité**.

JSTOR

La démarche de JSTOR[15] est de préserver les revues et publications périodiques sur le long terme grâce à la numérisation, alors que beaucoup sont menacées de destruction par l'usure prématurée liée aux conditions de stockage et à la composition du papier. Mais il ne s'agit pas seulement de conserver, l'objectif étant aussi de valoriser et de publier les fonds numérisés, en direction des chercheurs notamment.

Cette institution propose aux utilisateurs enregistrés des fichiers en PDF, créés au cours d'un circuit qui génère parallèlement les métadonnées qui seront exposées (de façon passive) ou exportées (à la demande des utilisateurs). Les informations sur le fichier peuvent être affichées sous forme de page web (en HTML); il est possible également de récupérer la référence bibliographique dans un logiciel de gestion bibliographique, de façon plus ou moins aisée. L'adresse de la page du document, qui est un contenu de métadonnée au même titre que les autres, peut être envoyée par mail.

Comme chez la plupart des grands acteurs de la publication scientifique, l'export vers RefWorks peut se faire à l'aide d'un simple bouton générant un script. Le format BibTeX est pris en compte au travers d'une manipulation simple mais à la longue sans doute fastidieuse (génération de texte dans ce format, à sauvegarder pour le conserver). Un fichier RIS est enfin disponible pour une utilisation de Endnote, Procite ou de la plupart des « Reference Managers ».

L'avantage de JSTOR pour notre propos est que la démarche, institutionnelle, est assez détaillée, y compris en ce qui concerne les processus internes de fabrication de l'information et de la « méta-information ». On apprend ainsi que les métadonnées sont avant toute chose identifiées par un bibliothécaire, qui rédige un document de cadrage de la récolte de ces métadonnées « constatées ». Celles-ci sont donc ensuite compilées au moment de la numérisation, puis vérifiées à la fois par des automates et des agents humains.

[14] Outline Processor Markup Language
[15] http://www.jstor.org/.

Le standard des métadonnées qui accompagne les documents pris en charge n'est pas courant, puisqu'il s'agit du format NLM adapté spécifiquement par JSTOR pour pouvoir prendre en charge, dans un futur probablement proche, les références croisées (citations de, citations par, identifiant unique). Sans surprise, le document lui-même est contenu dans un fichier XML.

Il semble ressortir de l'analyse de JSTOR que les métadonnées sont d'autant plus précises qu'elles sont créées et encadrées très **en amont** dans la chaîne de production éditoriale, et que cette précision ne peut être garantie que quand un travail a posteriori en donne l'assurance. Cela permet au portail de JSTOR de proposer une recherche à facettes, dont l'efficacité est hautement tributaire de l'exactitude des renseignements portés dans chacune des facettes, c'est-à-dire des champs de métadonnées.

En revanche, il est permis de s'interroger sur l'absence de mention d'une politique de conservation des métadonnées liées aux documents, qui font eux-mêmes l'objet d'un traitement à des fins de conservation. On peut penser qu'un interpréteur de XML sera disponible durant les décennies à venir, mais les formats de métadonnées seront-ils reconnus, compris ? Ne faut-il pas envisager une conversion, une adaptation, ou tout au moins s'assurer que le standard employé est suffisamment bien documenté pour préparer des traitements automatisés ?

Si comme on peut le penser une réflexion est menée sur ce domaine, les signes de cette réflexion sur la préservation à long terme des métadonnées sont invisibles. JSTOR montre que la qualité des métadonnées est affectée par le moment et la durée de leur création.

OCLC

OCLC[16] est l'organisme qui gère le volume de métadonnées numériques (a priori) le plus important au monde, au travers de WorldCAT. Et comme une réponse à la taille considérable de cette base, OCLC propose un très grand nombre de services associés, avec quasiment un format de métadonnées spécifique à chacun – mais tous ne sont pas de type bibliographique, caractéristique d'une production éditoriale ou même en relation avec le sujet qui nous occupe –. Décrire par le menu tous les fonctionnements récents, présents ou pressentis d'OCLC pourrait faire l'objet d'un rapport à part : on pourra se limiter à quelques traits particuliers.

Ce qui peut frapper un novice entrant sur le portail d'OCLC, c'est le nombre de formats et de méthodes spécialement utilisés ou « portés » par cet organisme. C'est le cas de VDX (Virtual Document eXchange) qui encadre le prêt entre bibliothèques, de Zportal, de VIAF (en partenariat avec de grandes bibliothèques nationales) ou de l'OCLC MARC (une version légèrement modifiée de MARC). C'est sans doute l'effet de la position dominante de WoldCat et de l'institution OCLC, qui permet à cette dernière de tenter d'imposer ses choix particuliers.

L'orientation en matière de métadonnées est témoin d'une évolution d'ensemble des bibliothèques. Les formats en usage à OCLC sont d'une part issus de la galaxie MARC

[16] http://www.oclc.org/fr/fr/default.htm .

(historiquement lié aux bibliothèques), d'autre part issus des langages à balises engendrés par le monde du World Wide Web, comme le Dublin Core et RDF. De la même manière que les formats de la Bibliothèque du Congrès, les schémas utilisés et développés par OCLC ont du prendre en compte la diffusion des usages d'Internet. Ainsi les notices bibliographiques peuvent être récoltées ou éditées, via la plate-forme *Connexion*, en MARC21 ou en Dublin Core (qualifié ou non), et dans des fichiers HTML, RDF ou XML. Enfin OCLC, parmi les nombreux services proposés, permet d'enrichir des notices d'éditeurs en ONIX par des autorités validées par le monde des bibliothèques.

Marginalement on peut noter que l'un des principaux reproches faits à Google Book Search concerne une mauvaise indexation matière, et le fait que Google n'ait pas passé de contrat avec OCLC pour ce processus d'amélioration des métadonnées a certainement autant une raison stratégique qu'économique. Quoi qu'il en soit le contrôle du vocabulaire qui qualifie les documents est sans doute le point fort des métadonnées produites par OCLC. La logique sous-jacente est qu'une métadonnée exprimée dans un vocabulaire contrôlé est plus facile à transmettre et à adapter par les outils à disposition des utilisateurs. Cette même logique est à l'oeuvre dans l'initiative d'un livre blanc des échanges de métadonnées dans la chaîne éditoriale, portée conjointement par OCLC et le NISO (agence de normalisation des Etats-Unis). L'intéressant rapport de Judy Luther met l'accent justement sur une rationalisation fondée sur une utilisation des vocabulaires contrôlés partagés par ONIX et MARC[17].

Des catalogueurs récupérant fréquemment les notices de WorldCat ont observé que la volonté de dédoublonner les notices est certes bienvenue, mais qu'elle restait au niveau des intentions pour de nombreux documents, OCLC étant il est vrai tributaire des renseignements donnés par les bibliothèques membres du réseau. Sur ce point, les métadonnées gagneraient en qualité si elles respectaient le principe du « one to one » (un document, un groupe de métadonnées) en s'appuyant sur un principe d'identification unique et pérenne plutôt que sur des comparaisons terme à terme, plus lentes.

L'exemple d'OCLC permet de mettre en avant l'importance de l'**uniformité** d'expression dans les métadonnées.

Postgenomic

Les blogs scientifiques ont pris un rôle croissant dans la publication et l'élaboration de la recherche depuis plusieurs années. Pour de nombreux chercheurs, comme l'indiquent Pierre Mounier et Alexandre Serres[18], ils font partie du travail de recherche ; leur reconnaissance comme instrument « officiel » d'échange scientifique, au même titre que les revues, les monographies ou les bases de donnée, passe néanmoins par une pérennisation, une citabilité et une trouvabilité accrues, donc par une mise en oeuvre attentive des métadonnées.

[17] (Luther 2009, vol. 1,)
[18] (Serres et CLEO 2009)

Postgenomic[19], comme plate-forme de blogs scientifiques / de recherche, a été préférée pour cette analyse à d'autres, dont Science blog[20], car cette jeune structure s'est lancée dans un travail intéressant sur les API (voir glossaire des sigles). On peut à ce propos remarquer la plus grande propension des petites structures à l'innovation, à l'expérimentation, ce qui est une façon sans doute de compenser les déséquilibres existants en termes de notoriété et de visibilité.

Postgenomic récolte des billets en HTML, y compris lorsqu'ils incluent des scripts en JAVA ou en d'autres langages. Les billets sont typés par l'insertion de balises complémentaires, et peuvent faire l'objet d'un suivi par le biais de fils RSS. Ce qui est plus spécifique à cette plate-forme d'édition est la mise en place de plusieurs API. Elles permettent une interrogation distante des métadonnées des documents, avec la gestion de requêtes en javascript et l'export de fils ATOM ou de fichier JSON. Une autre API permet de transférer les métadonnées dans les plates-formes de gestion bibliographique de *Nature*, Connotea, ou dans celle de Pubmed.

Cet usage pionnier de l'ingénierie documentaire afin de valoriser non seulement les données mais aussi les métadonnées rejoint une tendance générale à la dissémination des informations, d'autant plus aisée que chaque information est précisément tracée, identifiée. Nous ne pouvons pas savoir si le « cloud computing » passera comme une mode, ainsi que certains l'affirment. Mais il est probable qu'entre un standard rigide, proposant des informations en consultation simple, et un standard très adaptable, offrant un large choix de modes de consultation, c'est le dernier qui aura le plus de chances de s'imposer dans les usages.

L'exemple de Postgenomic nous montre que la qualité des métadonnées se mesure pour partie à l'aune de leur **fluidité** (c'est à dire la possibilité qu'ont des usagers de les manipuler, de les décomposer, de les recomposer en fonction de leurs besoins spécifiques)**.**

Le cas de Revues.org

Parmi les créateurs de métadonnées, nous pouvons porter une attention toute particulière à Revues.org[21], pour trois raisons :

- cette plate-forme est identifiée et se revendique comme un éditeur « ouvert » en sciences humaines et sociales, ce qui implique (on l'a vu plus haut) un rôle particulier des métadonnées;
- elle a engagé depuis plusieurs mois et poursuit une réflexion sur les standards de métadonnées qu'elle utilise;
- et son développement rapide, son attention particulière aux innovations technologiques comme son ouverture à une échelle internationale mettent en tension les méthodes et les fondements antérieurs de sa stratégie concernant les métadonnées.

[19] http://postgenomic.com/
[20] http://www.scienceblog.com/cms/index.php
[21] http://www.revues.org/

Pierre Mounier a récemment donné une description assez complète de ce qu'est Revues.org lors d'une journée d'études s'étant tenue le 11 décembre à Paris[22].

Parmi les publications de Revues.org, 2 revues, 1 collection de livres, 2 blogs de recherche et la plate-forme d'annonces professionnelles Calenda ont été étudiées. Le directeur du CLEO, Marin Dacos, a guidé le choix de cet échantillonnage pour permettre la prise en compte de la plus grande variété possible dans le temps imparti.

La *REMI*

La *REMI* est la *Revue Européenne des Migrations Internationales*[23]. Pour cette revue, nous pourrons nous appuyer sur l'exemple très précis d'un article récent de Françoise Bourdarias, actuellement en accès restreint *via* Cairn (Bourdarias Françoise, 2009, « Migrants chinois au Mali : une pluralité de mondes sociaux », *Revue européenne des migrations internationales*. Adresse : http://remi.revues.org/index4876.html [Accédé : 25 Décembre 2009].).

La *Revue Européenne des Migrations Internationales* a comme son titre l'indique une diffusion à vocation internationale. C'est pourquoi certains champs comportent des contenus en trois langues, français, anglais, espagnol. Les traductions constituent une question importante posée aux schémas de métadonnées, et même la solution des ontologies, envisagée pour les années à venir par les pionniers du web, risque de ne pas écarter tous les problèmes que posent les langues dans l'expression des métadonnées.

La première difficulté tient à la correspondance entre un ensemble de métadonnées et le document qu'elles décrivent. En théorie, si un standard est bien fait et le créateur de métadonnées rigoureux, il doit y avoir une description pour un document (les métadonnées de type annotations ou fichiers connexes mises à part). Mais une notice identique dans une autre langue, tout aussi rigoureuse, peut être prise à tort comme renvoyant à un autre document. On pourrait penser que les standards destinés à l'exploitation automatique et non pas humaine évitent cet écueil, toutefois un standard tel que MARC comporte encore des « Couv. ill. en coul. » dans certains champs de métadonnées, ou des formats de date en usage dans tel ou tel pays. Certains langages à balises permettent de spécifier la langue d'expression du contenu de la balise; malheureusement cette multiplication des champs rend lourde l'opération de génération des métadonnées, et il est alors presque obligatoire d'enrichir les métadonnées après la création du document, ce qui éloigne du temps optimal de description du contenu.

La seconde difficulté vient des nuances de sens, toujours difficiles à restituer en traduction, parfois porteuses de malentendus. Ceux-ci rendent risquée l'option, intéressante au demeurant, d'une expression des données et des métadonnées en fonction de la langue de l'utilisateur, les autres langues et la langue d'origine en particulier étant masquées.

Ce n'est pas le choix fait par Revues.org, qui a mis en place un système de boutons permettant de passer d'une langue à l'autre pour les résumés, avec donc un champ par

[22] (Mounier 2009)
[23] http://remi.revues.org/index.html

langue au niveau de la publication. Il faut toutefois noter que lors d'un import des métadonnées dans un logiciel de gestion bibliographique tel que Zotero, l'effort du multilinguisme perd une part de son intérêt car par commodité ce genre de logiciel ne prend qu'un champ (« résumé » par exemple), et au mieux fusionne les champs contenant des versions en différentes langues.

Cet aspect mis à part, lorsqu'on s'intéresse aux schémas utilisés par Revues.org pour la *REMI*, on rencontre dans un article des métadonnées en :

- HTML, avec la seule balise « title » dans la partie « meta » de la page.
- Dublin Core « légèrement qualifié », c'est à dire avec 12 des 15 champs du Dublin Core (DC) et un champ du Dublin Core Terms (DC Terms), « relation.isPartOf », servant à signaler l'institution hébergeant le document. On rencontre 2 champs « identifier », l'un pour l'adresse URL et l'autre pour l'ISSN, et 4 champs « contributor », avec 2 contenus identiques car la directrice de la publication et la directrice du numéro sont une seule et même personne (Marie-Antoinette Hily). La date est exprimée selon le standard du W3C (W3CDTF). La langue est exprimée selon les RFC3066[24] (même si pour l'article de Françoise Bourdarias daté du 1er juin 2009, le principal contenu disponible, le résumé, l'est en fait en 3 langues, on conçoit que le document principal, de langue unique, soit caractérisé par une seule mention de langue).
- Google Scholar (labels de métadonnées commençant par « citation_ »)

Signalons qu'en complément,

- des flux RSS sont disponibles dans les cadres fixes du site de la revue, permettant de suivre soit la série des articles soit la série des numéros de la revue;
- un dépôt OAI expose les métadonnées de la revue en Dublin Core et en METS.

L'évocation de METS et du RSS permet de souligner, pour tout format de métadonnées, la question de la granularité : sur quoi doivent porter les métadonnées d'une revue pour garantir un rapport qualité / coût optimal ? Sur la revue dans son ensemble, sur chaque numéro, sur chaque article, voire sur chaque paragraphe ? Les standards légers (RSS, DOI) peuvent décrire des objets réduits, mais le « bruit » (nombre excessif de références) serait très dommageable à l'institution qui en abuse. Les formats plus complets et plus complexes ne peuvent pas être raisonnablement appliqués à une échelle moindre que l'article, même si la disparition progressive de la notion de page dans l'édition électronique donne une valeur nouvelle au paragraphe.

A ce sujet, dans une plate-forme éditoriale aussi importante que celle de Revues.org, où le suivi qualité de toutes les séries demanderait en théorie beaucoup de temps, les métadonnées générées automatiquement peuvent se révéler inadaptées au contexte. La numérotation des paragraphes est intéressante (même si la citation visant un paragraphe particulier est loin d'être répandue), mais ne convient pas, par exemple, à des documents du type signets, alternant intitulé du site et adresse. La granularité des données sur lesquelles portent les métadonnées ne peut donc, comme nous le montre l'analyse de la *REMI*, se penser qu'en rapport avec la possibilité d'assurer la qualité de ces métadonnées. La qualité d'un standard est donc relative à l'échelle à laquelle on l'applique, ce qu'on pourrait appeler le caractère **proportionné** du standard.

[24] http://www.ietf.org/rfc/rfc3066.txt

Dans le travail sur les métadonnées de la *REMI* comme sur celles d'autres revues en ligne, on peut en revanche deviner l'application scrupuleuse de quelques bonnes pratiques, en particulier cette mention régulière des schémas d'encodage pour encadrer le contenu des métadonnées. L'ensemble donne d'ailleurs l'impression d'une attention portée à la justesse des informations sur le document, comme en témoigne cette mention de l'auteur avec l'indication de la partie correspondant au nom de famille.

Mais on peut remarquer, dans l'exemple très éclairant de l'article de Françoise Bourdarias, que l'étiquette de citation, en texte simplement stylé, ne correspond pas aux informations proposées par les métadonnées (texte daté du 1er janvier 2012 au lieu du 1er juin 2009, plus vraisemblable). De façon générale, on peut donc rencontrer des données descriptives fautives encadrées par des métadonnées de description correctes. Puisque le processus de création des métadonnées semble pour cette revue précis et correct, il est regrettable que celles-ci ne soient pas réutilisées pour la génération du texte lui-même, d'autant plus lorsque cette génération se fait avec plus de risques d'erreurs que pour les métadonnées (génération automatique par Lodel).

L'exemple de la *REMI* nous montre que la qualité des métadonnées a moins d'effet positif sans une utilité directe pour la qualité du document. C'est ce qu'on pourrait appeler le critère de la **réciprocité**. Ce dernier est néanmoins délicat à manier car il ne faudrait pas en déduire que les caractéristiques des métadonnées peuvent donner les caractéristiques des données. Ainsi les métadonnées d'un billet de présentation (http://remi.revues.org/index4075.html) même aussi détaillées que celles d'un article académique, ne font pas de ce billet un article scientifique. Cette réciprocité n'est un critère de qualité que lorsque l'utilisation des métadonnées peut se faire au bénéfice de la valorisation et la qualité du document.

Au moment de la rédaction de ce mémoire, une chaîne de production se mettait en place pour fournir à Cairn, dans le cadre d'un partenariat visant à une continuité de diffusion des revues avec barrière mobile, un balisage du texte en TEI. Mais les résultats n'en étant pas encore directement visibles, nous ne pouvons apprécier l'impact de ce balisage sur la qualité générale des métadonnées. Tout au plus peut-on préjuger d'une meilleure qualification des différents contenus des documents (noms, citations, traductions, résumés, etc.) au vu des exemples précédents référencés sur le site de la Text Encoding Initiative[25].

La revue *Cybergeo*

Si les usages en matière de métadonnées ont une certaine constance d'une revue à l'autre sur la plate-forme, on peut toutefois relever d'éclairantes nuances. La première différence visible dans *Cybergeo*[26] est le fait que les articles portent un numéro unique dans la revue. Cela semble un détail, mais dans la perspective d'une utilisation d'identifiants uniques et surtout des démarches de vérification de la qualité, c'est un pas important déjà effectué.

[25] http://www.tei-c.org/Activities/Projects/index.xml
[26] http://www.cybergeo.eu/index.html

La revue propose un guide du contributeur en 2 langues, mais les métadonnées de la page indiquent dans les deux cas un document en français. Il n'y a donc pas de cohérence entre les métadonnées et les données, du fait d'une erreur humaine. Cette question de l'entrée manuelle des métadonnées par l'auteur (pour les nombreuses publications en sciences humaines ayant des moyens réduits) ou par un secrétaire de rédaction (pour les revues plus structurées) peut être approfondie dans le cas de *Cybergeo*.

Revues.org utilise pour les revues, cahiers, livres et bibliographies une chaîne de publication reposant sur l'outil en *open source* Lodel. Quand on observe le module d'administration on peut être impressionné par les possibilités d'adaptation du modèle éditorial. La contrepartie de cette versatilité est que les formulaires de renseignement des métadonnées sont peu contraignants.

La figure suivante montre par exemple que les champs d'expression de date ou de pagination ne suivent pas une structure particulière : c'est ainsi qu'on pourrait rencontrer, dans une même revue pour les mêmes champs, les dates « 09/22/2009 », « 22 septembre 2009 » ou « sept. 2009 ».

Figure 1 : Formulaire "métadonnées" dans Lodel

La figure ci-dessus montre aussi que si le comité de rédaction ne fait pas de démarche pour la préciser, la « licence portant sur le document » reste vide ou, dans le meilleur des cas, reste celle choisie par défaut par l'éditeur.

En sciences humaines et sociales, la question de l'indexation matière a une particulière acuité, on est donc en droit de s'interroger sur la façon dont les auteurs peuvent décrire le contenu de leur document. Les mots-clés sont, c'est un point positif, choisis dans une liste indicative; mais une fois de plus la comparaison entre les différentes langues éclaire d'un jour particulier les problèmes liés à ces termes d'indexation. Les mots-clés ne sont pas toujours les mêmes en français et en anglais, comme le montre l'exemple de l'article sur « La lutte des places à Chamonix », visible ci-dessous. On remarque que le problème ne vient pas d'une traduction difficile d'une langue à l'autre, mais d'une simple omission en anglais, volontaire ou non, puisqu'il s'agit du nom de la ville de Chamonix (les deux mentions de « place » montrent que ce n'est pas en raison d'une homographie).

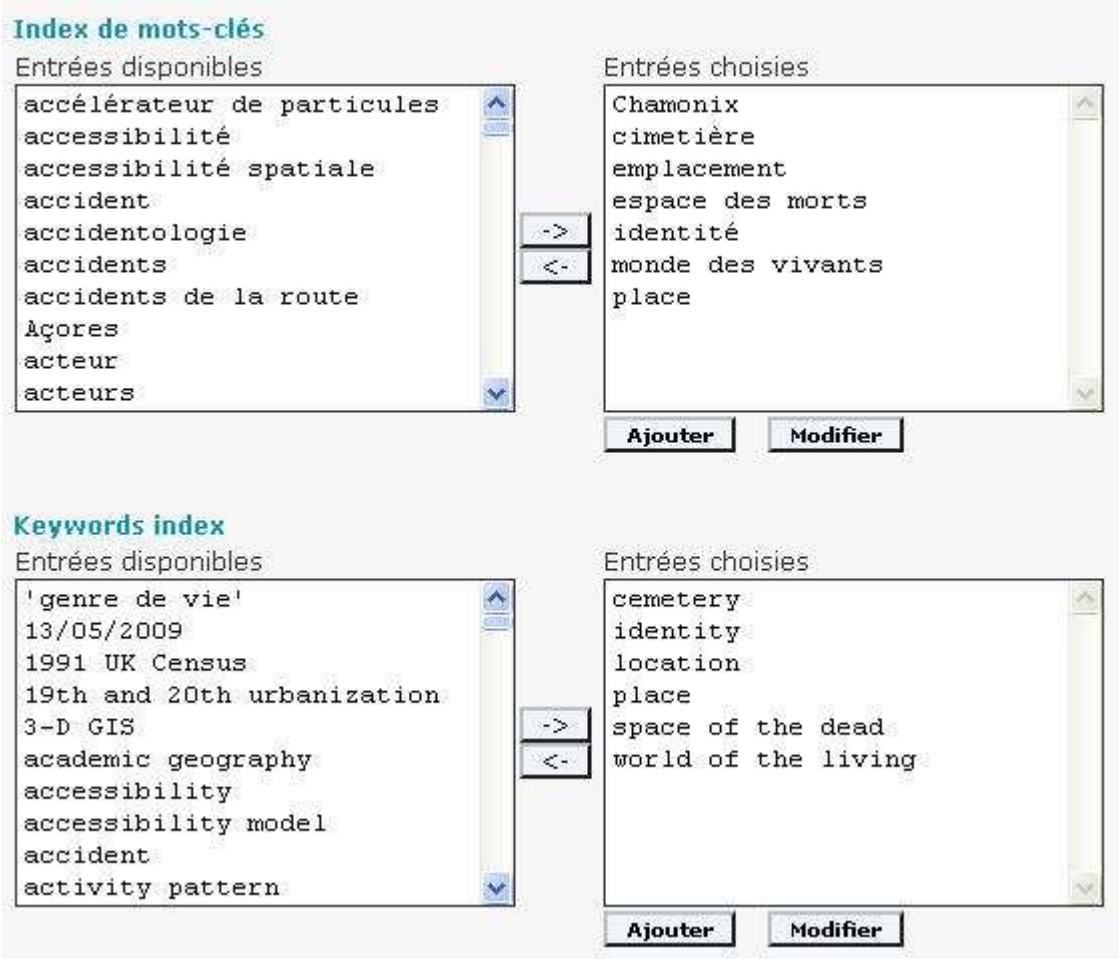

Figure 2 : Formulaire "mots-clés" dans l'espace Lodel de la revue Cybergeo

Les cadres de création de métadonnées sont autant de potentialités d'information sur les documents. L'idée des index est précieuse pour valoriser les « facettes » des documents noyés dans le flux d'une revue ou d'une plate-forme de publication telle que Revues.org. Mais il est difficile d'imaginer une institution ou un utilisateur tirer profit des index chronologiques et thématiques de *Cybergeo* étant donnée leur pauvreté:

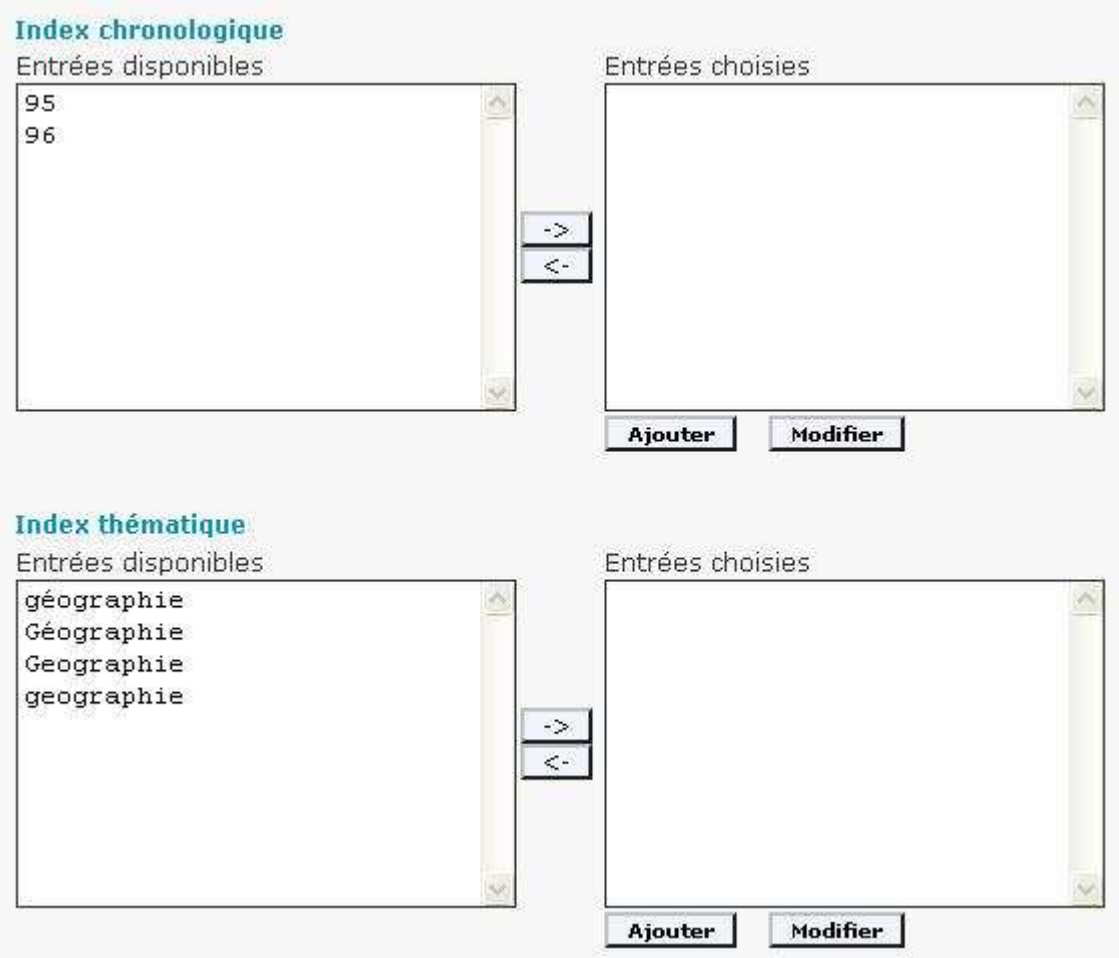

Figure 3 : Formulaire "index" dans l'espace Lodel de la revue Cybergeo

Cette observation permet de formuler 2 hypothèses non exclusives : la direction de la revue a pu imaginer à tort que ces types d'index feraient partie des besoins des auteurs pour décrire finement leurs documents; les auteurs ont pu ne pas savoir comment enrichir ces index en cohérence avec leur propre travail et les orientations de la revue. Quoi qu'il en soit l'uniformisation de la richesse (et donc aussi de la qualité) des métadonnées passe sans doute par un dialogue régulier entre ceux qui déterminent la politique de métadonnées et ceux qui alimentent le contenu de ces « petites cases » (pour reprendre l'expression mise en exergue par Gautier Poupeau dans le titre de son blog[27]).

De l'exemple de *Cybergeo* on peut tirer la constatation suivante : la qualité d'un standard de métadonnées augmente lorsque celui-ci **structure** suffisamment les initiatives humaines pour que les traitements automatiques restent probants.

Si on poursuit l'analyse des métadonnées de cette revue de géographie, on ne peut manquer de s'intéresser au traitement des images (cartes, photographies, modèles). Les règles édictées par le comité de rédaction de la revue imposent une charte de nommage simple (nom de l'auteur de l'article, type d'image, numéro – cf. http://www.cybergeo.eu/index5021.html). Les images sont présentées sur le site de la revue avec 1 seul élément obligatoire, le titre, et 2 éléments facultatifs, la source et la

[27] http://www.lespetitescases.net/

légende. Le fait que l'indication de la source soit seulement facultative n'est pas forcément une bonne chose, dans le double contexte de la diffusion sur Internet et de la publication scientifique : les données et les fichiers circulent très vite et facilement sur le web, et il est préférable de ne pas laisser de document sans indication d'autorité. Et ce d'autant plus que la rigueur scientifique impose de préciser autant que possible l'origine des informations, même visuelles, qui sont utilisées. Une bonne pratique à ce sujet est observable dans l'article de Sébastien Hardy « La vulnérabilité de l'approvisionnement en eau dans l'agglomération pacénienne »[28] : même si les cartes et photos sont l'oeuvre de l'auteur de l'article, elles portent une mention de responsabilité individuelle.

Le dossier « Vulnérabilités urbaines au sud » dont est issu ce dernier article permet de constater l'intérêt de COinS pour une revue, tant qu'un service OpenURL répond aux besoins de traitement par lots de l'utilisateur : il est possible d'exploiter en une opération toutes les références bibliographiques du dossier (par exemple les récupérer dans Zotero). Le standard RSS en revanche a plutôt été négligé par la revue, puisque le seul fil de syndication disponible est celui des 10 derniers documents publiés par la revue (il n'aurait pas été coûteux ou gênant d'étendre à la série complète des articles, puisque les agrégateurs de flux actuels évitent plutôt bien les doublons). Nous ne tirerons pas forcément de ces dernières observations de règle générale sur la qualité des métadonnées mais elles participeront à l'appréhension de l'usage relatif des standards dans l'édition électronique ouverte.

La collection des livres de l'IFPO

Première initiative d'ampleur de Revues.org dans le domaine des monographies, la collection des livres de l'IFPO (Institut Français du Proche-Orient)[29] permet de s'intéresser au fort développement actuel des e-books. L'expression anglophone « e-book » étant ambiguë, pour plus de clarté, nous utiliserons la distinction faite par Alain Pierrot, lors de sa conférence sur la convergence numérique dans l'édition à l'occasion de l'Université d'été du CLEO[30], entre livrel (document correspondant aux critères traditionnels du livre, mais disponible et lisible sur support numérique; le fait qu'un document, composé en majorité de texte, fasse plus de 300 kilo-octets est de plus en plus perçu comme un critère suffisant pour le désigner comme livre électronique) et liseuse (matériel dédié à la lecture des documents numériques, dont en particulier mais pas exclusivement des livrels).

Mais pour l'heure les livres disponibles sur Revues.org le sont en XHTML ou en PDF, car ces formats sont assez aisés à obtenir à partir d'un texte balisé tel que Lodel en produit nativement et directement. On peut percevoir 2 avantages, du point de vue de la qualité des métadonnées, à ce choix fait au détriment d'un format plus spécifiquement orienté vers la publication de monographies, tel que le propose ONIX. Premièrement, le travail fait sur les autres types de documents, puisqu'il porte *grosso modo* sur les mêmes standards (famille des langages à balise du web et format PDF), bénéficie aussi aux collections de livres. Deuxièmement, le XHTML est une porte d'entrée vers le standard EPUB, dont la partie « contenu » (le fichier OPS) est directement inspirée; par le biais

[28] http://www.cybergeo.eu/index22270.html
[29] http://ifpo.revues.org/
[30] http://www.digitalhumanities.cnrs.fr/wikis/edelec-shs/index.php?title=La_convergence_num%C3%A9rique_dans_l%E2%80%99%C3%A9dition

du format MOBI devient même également accessible, moyennant des ajustements, la publication sur les liseuses Kindle d'Amazon (voir à ce sujet la note technique d'Adobe[31]).

Nous apparaît alors l'un des critères possibles de la qualité d'un standard de métadonnées, sa **potentialité**, c'est à dire son ouverture à des conversions ultérieures propices à une diffusion efficace de l'information scientifique. Ce critère est très lié au contexte technologique, aux initiatives ponctuelles et aux besoins émergents en matière de métadonnées, ce qui prouve encore s'il en était besoin la nécessité d'une veille (de préférence partagée) sur les expériences de « crosswalks ».

Dans le cadre général de la recherche en sciences humaines et sociales, où le livre a eu et a toujours un rôle fort face aux publications sous forme de revues (principalement pour des raisons historiques, économiques et épistémologiques), il est frappant de constater que les livres publiés par Revues.org adoptent une forme de présentation extrêmement proche de celle des revues. C'est lié aux 2 formats de fichiers employés, qui sont communs aux deux types de publication, au fait que de nombreux livres de niveau recherche sont des recueils d'articles, et à une stratégie de valorisation globale, indifférente au type de document. On pourrait presque dire que la principale différence réside dans les champs utilisés dans les standards de métadonnées, adaptés à la description soit d'un article de revue soit d'un chapitre de livre (le fait de se placer à ce niveau hiérarchique du document est d'ailleurs en soi une preuve de la proximité des approches).

L'ISBN, la fréquence de contributions techniques particulières (illustrateur, traducteur...) ou l'existence d'une image en page de garde sont quelques-unes des rares spécificités des livres électroniques. En revanche, sur le plan de la plate-forme de publication, il est fréquent (et c'est le cas pour les collections numérique de livres de l'IFPO) d'ajouter un échelon. Alors que pour les revues on applique des métadonnées « riches » aux niveaux de l'article et du numéro de la revue, dans le secteur des livres proposés par Revues.org, ce sont le chapitre, le livre et la collection qui peuvent être qualifiés (on laisse de côté les sections dans un numéro et les parties dans un livre, qui sont en général de simples guides de la lecture peu caractérisés). L'éditeur (au sens de l'anglais *publisher*), le directeur de la collection, le directeur de l'ouvrage collectif et l'auteur du chapitre sont des mentions de responsabilité qu'il est par exemple mal aisé d'indiquer à tous les niveaux de granularité... ce qu'il faudrait pourtant faire ! Ce n'est pas tant à cause de l'opportunité de traiter des documents par lots, comme le permettrait par exemple le duo COinS / Zotero, que du fait d'un usage croissant de la bibliométrie pour l'évaluation des sciences humaines et sociales, d'où la nécessité d'identifier les mentions de responsabilité à toutes les échelles de consultation.

Au travers d'un processus **d'héritage** du niveau supérieur vers un niveau inférieur, qui complète automatiquement mais ne fige pas les métadonnées, il est envisageable d'augmenter l'exhaustivité, donc la qualité, des métadonnées. C'est une voie empruntée par Revues.org, comme en témoigne son utilisation de METS ou l'application automatique du nom de l'institution éditrice, mais on peut supposer possible de poursuivre plus loin cette démarche.

[31] (Adobe 2009)

Le blog *Homosexus*

La plate-forme de blogs scientifiques en sciences humaines et sociales de Revues.org s'appelle Hypothèses. Le fait de regrouper les blogs thématiques autour d'une discipline permet de faire jouer les rebonds d'un blog à l'autre, de la même façon que certains utilisateurs explorent les différentes revues signalées dans les résultats d'une requête fédérée. Hypothèses repose sur le logiciel Wordpress, ce qui garantit une certaine forme d'unité de traitement des données et des métadonnées.

A propos de ces dernières, il n'existe pas de norme spécifiquement conçue pour décrire les documents produits par les blogs. On parle de publication discontinue, irrégulière, et il est vrai que par exemple une date n'est pas toujours suffisante pour identifier un billet de blog (plusieurs articles pouvant être édités le même jour). Certains articles sont pourtant aussi insérés dans une série sur un thème, et devraient comporter une mention de cette relation : le blog d'*Homosexus*[32], « compagnon » de la rédaction d'un « livre sans titre » (Pierre Mounier parle de « séminaire virtuel » dans la notice d'utilisation d'Hypothèses), est justement un très bon exemple de cette structure globale dans laquelle s'insèrent les documents, mais que seuls les utilisateurs humains savent interpréter directement. Le fait qu'il soit facile d'accéder aux billets précédents ou suivants n'est pas une réponse adaptée, car ce fil thématique peut être suivi par intermittence. Nous pourrons revenir sur ce sujet après avoir étudié les standards utilisés par le blog *Homosexus*.

L'une des méthodes les plus connues des blogueurs, même amateurs, pour la dissémination des blogs est celle des fils RSS. C'est aussi celle qui est la plus encadrée par l'usage et par les recommandations techniques. C'est donc sans surprise que le blog *Homosexus* propose un fil RSS à ceux qui souhaitent agréger le flux des articles sans forcément vouloir revenir dans leur cadre de présentation originel. ATOM semble d'autant moins privilégié par les tendances actuelles que le RSS 1.0 s'appuie sur le RDF, standard en vogue dans les développements en cours (le RSS 2.0, plus évolué, n'a pas cette caractéristique). Dans le cas d'*Homosexus*, le fil mis à disposition est celui des articles et non des commentaires, soulignant la volonté de l'auteur de diffuser des hypothèses scientifiques plus que d'engager un dialogue à bâtons rompus.

En tant que page XHTML un article de blog comporte des métadonnées dans les balises « head », à commencer par le titre. On peut rappeler que les robots des moteurs de recherche prennent assez peu en compte les métadonnées de cette partie « head » d'une page web, à l'exception notable du titre (balise « title ») qui a toujours un rôle dans la mesure de pertinence par ces moteurs. Les autres standards de métadonnées exploités le sont sur des ancres particulières dans le texte, par exemple au travers de l'utilisation d'un widget (cadre dynamique dans la page permettant le lancement d'une application tierce), mais pas de façon globale pour un document.

Ainsi les blogs de la plate-forme Hypothèses utilisent le standard COinS, ce qui rend très aisée l'utilisation par un outil de gestion bibliographique tel que Zotero, d'autant plus que le marquage par mots-clés est assez naturellement employé pour catégoriser les billets publiés. Il est donc facile de récolter tous les billets d'une catégorie qui intéresse le lecteur ou le chercheur.

[32] http://homosexus.hypotheses.org/

Les autres grands standards de description sont absents, même si en théorie ils pourraient être mis en application : un enrichissement par du Dublin Core non-qualifié, généré automatiquement, serait envisageable dans un contexte de publication régulière et scientifique. Toutefois le contraste entre l'exemple d'*Homosexus*, très structuré et méthodique, et d'autres blogs de recherche, nous permet de mesurer la difficulté d'ajuster des pratiques de métadonnées pour des formes certes à vocation scientifique, mais au statut documentaire encore trop variable. Il semble que la masse critique des documents soit trop rarement atteinte sur les blogs pour justifier de mettre en place une exigeante politique globale en matière de métadonnées : on imagine difficilement la constitution d'une notice bibliographique complète par l'auteur pour son billet d'humeur. C'est un autre témoignage en faveur de métadonnées d'une complexité **proportionnée** à l'objet qu'elles documentent.

Le blog Culture et politique arabe

Le blog *Culture et politique arabes* (*CPA*)[33] mélange du texte avec de nombreux documents multimédias. On est donc en droit de se demander quels standards sont utilisés, et de quelle façon. Les images sont insérées aisément grâce à la grammaire du HTML, mais les informations sur ces images sont assez limitées. On retrouve ce qui est automatiquement signalé par les logiciels de création de blog, à savoir le format et la résolution; mais il manque très souvent l'indication de l'origine de ces images, de leurs conditions de réutilisation, de leur auteur ou même une légende qui leur donne sens. Nous pouvons nous appuyer sur le cas de l'article « No (Coranic) Logo ! »[34] pour relever d'emblée une image sans identification autre que son nom, formé par le mot « priere » et le format (198x300). Et encore faut-il user de méthodes indirectes pour avoir ces informations complémentaires (en accédant aux propriétés de l'image ou au code source de la page).

Les vidéos sont parfois pointées en tant que lien, parfois insérées dans des widgets. L'avantage des solutions « encapsulées » (de « embedded » en anglais), telles que les widgets, est que les mentions de responsabilité, si importantes dans le cadre de la publication en ligne, sont du ressort de la structure d'origine. Les liens en général, ceux des vidéos en particulier, gagneraient en visibilité et en clarté si des métadonnées autres que la seule adresse étaient indiquées. Si les séquences filmées ou sonores sont hébergées directement par l'institution éditrice, la nécessité d'une meilleure identification impliquera d'améliorer le balisage de ces données non-textuelles. Dans le cas étudié, un widget permet une identification restreinte du sens de la vidéo, ce qui contraste aussi avec l'image insérée sans titre ni légende.

A partir du mode d'insertion de cet objet, on peut identifier une forme de méta-information encore peu employée dans l'édition scientifique en général, en sciences humaines et sociales en particulier : les scripts ne sont pas des données statiques mais des outils de modification des données de l'objet. Ils permettent par exemple au lecteur du blog *CPA* de passer la vidéo en mode plein écran. C'est en somme une part des métadonnées (ici le format de visualisation) qui peut s'adapter aux besoins de

[33] http://cpa.hypotheses.org/
[34] (Gonzalez-Quijano 2009)

consultation. C'est cette même conception qui préside à la prolifération des API dans la diffusion de l'information. La technicité de la question rend le fonctionnement de ces outils délicat à retranscrire, mais on peut dire qu'un standard de métadonnées a un intérêt supplémentaire lorsqu'à partir de sa conception et grâce à cette conception, des utilisations variées et dynamiques des métadonnées sont possibles. On pourrait parler du caractère **dynamique** d'un format de métadonnées. Il est évident que cette capacité de mouvement est d'autant plus aisée à obtenir que la structure de métadonnées est simple.

Pour revenir au cas précis du blog *CPA*, on peut remarquer que la fréquence des publications (le rythme annoncé est d'un article par semaine) rend difficile l'attention des rédacteurs aux manipulations de texte. Il est peu visible pour les lecteurs que le texte en français qu'ils consultent est déclaré par une balise « span » comme étant de langue anglaise britannique. Ce phénomène se rencontrait aussi marginalement dans le blog *Homosexus* mais c'est ici plus prégnant. Ce genre de discordance peut toutefois avoir des conséquences sur les traitements automatiques futurs. En imaginant qu'un programme récupère automatiquement tout le texte en français du blog *CPA*, on devine qu'une partie du contenu serait à tort écartée au détriment des besoins du producteur de métadonnées comme de l'usager.

Dans un univers où les assemblages de documents hétérogènes sont systématiques, et donc les métadonnées éparses réunies, il apparaît important que les métadonnées d'un document composite puissent être à la fois héritées (d'une structure d'ensemble) et **choisies** (par le créateur du document). C'est particulièrement vrai pour les champs qui affectent le sens escompté et les mentions de responsabilité : rien de plus déroutant que de conserver une indexation matière dans un contexte qui l'invalide (tel un exemple de formule mathématique employé en linguistique), ou de déclarer la propriété d'un éditeur dont les droits d'exploitation ont été cédés.

Calenda

Calenda[35] est une initiative originale, en tout cas dans l'univers de la recherche en sciences humaines et sociales, et correspond bien à l'esprit qui anime les éditeurs de l'édition électronique « ouverte ». Il s'agit en effet de diffuser le plus largement et le plus finement possible les informations sur les rencontres, séminaires, conférences, colloques professionnels, donc finalement de favoriser la diffusion et la vitalité de la recherche grâce aux technologies de l'information. Parler de métadonnées pour des rendez-vous peut surprendre au premier abord, et pourtant dans un but de classement ou de présentation ces données structurées sont précieuses.

Pour l'instant les événements publiés le sont avec 5 champs Dublin Core (identifiant titre, date, type, description et créateur), mais à terme le CLEO envisage de coder chaque rendez-vous avec une trentaine de champs Dublin Core, dont certains DC Terms (éléments du Dublin Core qualifié) qu'il était logique d'employer : DC.coverage.temporal pour la période historique traitée par l'événement, DC.coverage.spatial pour la zone géographique étudiée. On voit en pratique l'avantage du caractère répétable du Dublin Core (qui génère certes aussi des inconvénients), puisque la répétition des champs « description » et « contributor » permettraient de faire

[35] http://calenda.revues.org/index.html

passer différentes informations complémentaires et utiles : le lieu du rendez-vous, la date du rendez-vous (et non la date de la notice sur l'événement, indiquée dans le champ « date »), la personne ayant mis en ligne l'annonce, les sources de l'information ayant servi à la publication de l'annonce....

Il est envisagé de puiser les indications de lieu et de temps dans les vocabulaires contrôlés tels que le TGN (thésaurus des noms géographiques du Getty) ou le schéma de date du W3C (le W3C-Date and Times Format, qui est une norme ISO 8601). Et c'est en effet une règle générale en matière de métadonnées, dès lors qu'on a pu identifier une formalisation plus précise, et dont l'interprétation se fait de façon transparente pour l'utilisateur, il est préférable d'en tirer profit.

Le parti pris par l'équipe de Calenda est de considérer l'annonce comme un document en soi, en relation avec un événement extérieur. Les métadonnées de la page, les balises Dublin Core envisagées s'appliquent donc à l'annonce. On est bien ici dans le cadre d'une plate-forme d'édition scientifique, dont le coeur de métier est la publication de documents. Ce n'est pas un problème dans la mesure où, en relation avec cette annonce, l'utilisateur peut disposer des métadonnées de l'événement lui-même.

Or justement, des balises du format hCalendar (version codée en html de iCalendar) sont intégrées dans les pages des annonces, permettant entre autres d'exporter un fichier .ics (iCal), d'ajouter l'événement à Google Calendar ou Yahoo! Calendar. Leur principe, commun aux différents standards de Microformats, est de rendre quelques éléments ciblés d'information détectables par des applications dédiées, offrant ainsi un éventail d'utilisations pour ces métadonnées. L'attention a donc été portée au fait que les formats d'export de Calenda véhiculent les métadonnées de l'événement lui-même et non celles de l'annonce.

Le dernier standard de métadonnées mis en avant pour la valorisation des informations de Calenda est le RSS, avec une très fine adaptation aux besoins des utilisateurs. En effet les flux disponibles ciblent presque tous les niveaux hiérarchiques de la plate-forme. On peut (rarement) souhaiter recevoir toutes les annonces sans distinction, ou on peut préférer recevoir les annonces de toute la catégorie Sociétés, ou de la sous-catégorie Sociologie. Ce n'est pas le cas mais on aurait pu imaginer un fil RSS de suivi de la subdivision Sociologie du travail.

La question du suivi d'une annonce en particulier n'est pas si accessoire qu'elle le paraît. En effet, il est arrivé qu'un événement soit annulé, comme le cycle de conférences « Destins du structuralisme » de Eduardo Viceiros de Castro, prévu au musée du Quai Branly début 2007[36]. Ne faudrait-il pas en informer les personnes intéressées par cette annonce en particulier, par exemple grâce à un fil RSS dédié ? Les milliers de fils RSS seraient certainement difficiles à gérer par les serveurs de Revues.org; il serait toutefois utile de distinguer, autrement que par les dates et des éléments de plein texte, les événements prévus, les événements passés, les événements qui n'auront pas lieu, les événements qui n'ont pas eu lieu. Sur ce genre de métadonnées le Dublin Core n'est pas totalement satisfaisant, car il lui manque une prise en compte fine du cycle de vie et des états du document; on pourrait néanmoins faire bon usage, dans ce contexte, des balises « extent » et « modified » pour repérer respectivement les notices n'étant plus valides et les notices modifiées depuis leur création.

[36] http://calenda.revues.org/nouvelle7729.html

A la suite de cette courte analyse des métadonnées utilisées par et pour Calenda, on peut mesurer l'importance de pouvoir rendre compte avec précision des évolutions des objets qui sont liés à un ensemble d'éléments descripteurs. On pourrait dire que dans le domaine des standards l'existence d'un **cycle de vie** des métadonnées est un critère de qualité. De plus, dès lors que des vocabulaires contrôlés sont disponibles, le fait de s'appuyer sur leur **précision** augmente l'intérêt des métadonnées en valorisant leur « **réutilisabilité** ».

LES PRATIQUES DES UTILISATEURS DE METADONNEES

Après avoir étudié les usages des créateurs de métadonnées, il semble intéressant de faire un rapide parcours des pratiques de certains acteurs qui utilisent les métadonnées, les traitent, avant une mise à disposition des utilisateurs finaux. Nous nous pencherons donc sur un outil de gestion bibliographique (Zotero), une service de reconstitution des tables des matières (ticTOCs), et une fondation gérant entre autres un dispositif d'identification pérenne (CrossRef, qui a la charge du DOI).

Zotero

Dans l'univers des outils de gestion bibliographique, après Endnote et Refworks, semble venu l'époque de Zotero[37]. La raison de cette popularité croissante, par rapport à un autre logiciel tel que Jabbref, est justement qu'il ne s'agit pas d'un logiciel mais d'un module de navigateur (pour l'instant, réservé au seul Mozilla Firefox). C'est tout-à-fait significatif d'un changement dans les pratiques de recherche, où l'utilisation d'internet est devenue essentielle. Puisque l'outil principal du chercheur, qu'il soit en sciences humaines et sociales ou en sciences dites « dures », est devenu le navigateur web, celui-ci devient très naturellement et avantageusement son « porte-document », le « classeur » où il range les articles et documents dont il a besoin.

De façon liminaire à cette étude de Zotero, il faut préciser que ce module bibliographique est en pleine évolution, avec une version « beta » (2.xx) de plus en plus éloignée des caractéristiques de la version stable actuelle (1.xx). Il est donc difficile de décrire son fonctionnement avec autant de certitude que d'autres logiciels sur lesquels nous disposons de plus de recul. Par exemple les champs de métadonnées pris en charge pour un article de revue ne sont pas les mêmes entre la version stable et la version en développement; mais quels champs resteront finalement dans la prochaine étape du développement ? Nous serons donc prudents concernant les points techniques caractéristiques de Zotero, tout en tirant des indications précieuses sur les tendances actuelles chez les standards de métadonnées privilégiés par les logiciels bibliographiques.

En tant qu'outil de gestion de bibliographie, Zotero est principalement un « récupérateur » de métadonnées. Il utilise à cette fin des interpréteurs de métadonnées (appelés « translators ») capables de lire les formats suivants :

[37] http://www.zotero.org/

- MODS
- MAB2
- MARC
- RDF
- Refer / BiblX
- RIS
- BibTeX

Avec raison ce module bibliographique prend en charge les formats d'export des principaux logiciels déjà existants, Refer, RIS (pour Endnote et Procite) et BibTeX. Il conserve l'ouverture aux formats de notices de bibliothèques que sont MARC et MODS. Et il s'ouvre au standard de base du « web de données », le RDF (sans préciser qu'il s'agit en fait du RDF XML). MAB2, le format bibliographique de la bibliothèque nationale d'Allemagne, est beaucoup plus rare (et on se demande donc si dans les développements ultérieurs de Zotero ce format ne sera pas abandonné, au profit d'un import indirect *via* MABxml par exemple). Pour tous les autres types de documents, la capture soit directe (PDF, HTML, PHP) soit en fichier attaché (.doc, .odt, .txt, .xls, etc.) garantit leur disponibilité même hors connexion. Ce fonctionnement semble un bon compromis entre deux écueils, celui de l'insuffisante diversité des sources des documents récupérés automatiquement et celui d'une prise en compte trop coûteuse des formats complexes, mineurs ou peu utilisés.

Zotero permet également de choisir un résolveur de lien privilégié lorsque l'utilisateur rencontre des élements en OpenURL, et de récupérer notamment les métadonnées des documents signalés grâce à COinS, ce qui est une façon aussi de promouvoir l'utilisation de ces standards.

De l'exemple de Zotero, nous pouvons déduire que la qualité d'un standard de métadonnée tient pour partie à son statut de **pivot** pour des échanges entre des sources et des utilisateurs hétérogènes. Un standard qui sera très utilisé par une communauté mais qui restera hermétique pour les outils propres à une autre communauté n'a pas une qualité aussi grande que celui qui sera disponible pour des conversions, des « crosswalks » intéressants. Dans le cas de la recherche en sciences humaines et sociales, dont Jenny Fry (appliquant aux champs disciplinaires la mesure du « scatter » de Bates, par exemple lors de sa présentation à la journée d'étude sur la diversité des pratiques numériques[38]) a montré qu'elle tirait un grand avantage des initiatives transdisciplinaires, l'importance de cette ouverture aux pratiques de métadonnées afférents à divers réservoirs d'information apparaît d'autant plus. Mieux vaut donc gérer correctement quelques standards fondamentaux que de chercher à adhérer à l'éventail des usages d'un public ciblé, ces usages étant de toute façon amenés à changer à plus ou moins longue échéance.

L'autre aspect qui mérite notre attention dans la politique de Zotero en matière de standards de métadonnées est que, sur le mode de tous les outils en *open source*, les utilisateurs (ou plutôt les communautés d'utilisateurs) ont toute possibilité pour proposer des extensions et des compléments qui leur semblent opportuns. Dans un dialogue dont la vitalité est perceptible dans le forum de développement de Zotero, les besoins et les propositions sont examinés assez attentivement. Certains formats de métadonnées sont

[38] (Fry 2009)

certes parfois mal appréhendés par le code de Zotero (on pense par exemple à MODS, loin d'être rigoureusement traité), mais la relative clarté des scripts permet aux développeurs de soumettre des améliorations, si la nécessité se fait jour. Mais du côté des métadonnées également, la clarté du point de vue des humains joue positivement sur les réglages des manipulations : MARC ou le format .enl de Endnote sont par exemple beaucoup plus difficiles à aborder pour les professionnels des traitements automatisés que le Dublin Core et ses labels explicites. En résumé, un standard de métadonnées gagne en qualité lorsqu'il est **interprétable** à la fois par les machines et par les humains, puisque comme le rappelle le rapport de Jennifer Schaffner, « les métadonnées sont l'interface »[39].

ticTOCs

Soutenu en particulier par le JISC (Joint Information Systems Committee), ticTOCs[40] est un service dédié aux chercheurs. Ce service vise à uniformiser et collecter les métadonnées issues des tables des matières de différentes revues de niveau recherche. L'objectif est de proposer aux utilisateurs, en lieu et place de flux d'articles lorsqu'ils sont disponibles, des agrégats par numéro de revue consultables avec des agrégateurs RSS traditionnels.

Une grande partie du travail sous-jacent a été exposé lors d'une conférence de Santiago Chumbe et Roderick Macleod qui s'est tenue à Leuven[41]. Dans la chaîne de diffusion des informations scientifiques ticTOCs ajoute à la fois une étape et un service : la qualité d'utilisation des métadonnées est-elle améliorée ou dégradée par cet intermédiaire supplémentaire ? Il faudrait sans doute consulter un large panel de chercheurs pour le dire avec certitude (certains y ayant recours, d'autres non). Néanmoins l'exclusivité de RefWorks dans l'exploitation bibliographique des tables des matières constituées par ticTOCs s'apparente à une forte limitation. Surtout lorsqu'on constate, comme nous l'avons dit plus haut, une très rapide et logique progression de l'usage de Zotero, impliquant des opérations de traitement de données supplémentaires (export, conversion puis import) pour ceux qui ont recours régulièrement à cet outil.

Dans la réflexion globale des instigateurs de ce service l'accent a été mis sur la nécessité d'une complémentarité entre les enrichissements automatisés et les validations par des agents humains. Et avec raison car aucun processus d'une telle ampleur ne pouvant être parfait, il reste des irrégularités dans le traitement des sources différentes. A un moment donné par exemple, la *REMI*, présente dans les flux agrégés par ticTOCs, était signalée sous le titre *Articles – Revue européenne des migrations internationales* car le fil RSS d'origine, portant sur les articles, soulignait la différence avec le fil RSS des numéros. C'est la conséquence du risque pris par le service ticTOCs avec le choix de récolter des métadonnées très actuelles car artificiellement créées (à la différence de Zetoc, qui avec un objectif similaire demande aux revues de fournir des fichiers déjà structurés en tables des matières et fait donc porter sur elles l'effort d'ingénierie informatique).

[39] (Schaffner 2009)
[40] http://www.tictocs.ac.uk/
[41] (Chumbe et Macleod 2009)

D'autres pistes de traitement de métadonnées sont esquissées par ce service du JISC, comme la création d'un fil de suivi à partir de l'ISSN de la revue, une récolte ciblée sur un champ disciplinaire à partir du code Dewey correspondant, ou un processus d'alertes sur les nouvelles tables des matières. Mais s'il est permis un jugement rapide, on soulignera les incohérences et lacunes de ces initiatives :

- Rares sont les usagers qui ont connaissance de l'ISSN d'une revue scientifique sans pouvoir accéder au site Internet de cette revue, et réciproquement rares sont les revues académiques en ligne ne proposant pas encore d'option de suivi (RSS ou ATOM).
- S'appuyer sur le code Dewey pour balayer exhaustivement un champ disciplinaire, étant donnée la diversité des codes possibles, d'une situation à l'autre, pour un même document, semblera aussi insatisfaisant que le fait de transporter de l'eau entre ses mains.
- Alors que la diffusion sur Internet et la « voie dorée » du libre accès en particulier favorisent la « désintermédiation » (c'est-à-dire la levée progressive des barrières entre les producteurs et les utilisateurs de l'information), ticTOCs envisage de signaler à l'utilisateur l'apparition de nouvelles tables des matières générées à partir des indications de mise à jour envoyées par les fils RSS des revues... ce qui s'apparente à une inutile « intermédiation ».

Pourtant malgré ces errements du développement ticTOCs a une utilité dans la chaîne de transmission des métadonnées, surtout si on se place dans le cadre des bibliothèques universitaires, car il peut s'interfacer avec les outils de gestion des revues électroniques (tels EZProxy, WAM ou LibX) pour transformer presque instantanément une mention « pauvre » de la revue en vue détaillée sur le dernier numéro. La richesse de cette vue détaillée reste tributaire de celle des métadonnées fournies par les revues dans les fils RSS, mais c'est l'occasion de percevoir justement l'importance de la qualité des métadonnées dès la création des documents. La pertinence du standard de métadonnées choisi ne peut se mesurer que lorsqu'on considère les vecteurs de diffusion dans leur ensemble et dans leur totalité.

On voit au travers du fonctionnement de ticTOCs que les métadonnées ne sont plus seulement des « étiquettes » favorisant le classement du document, ou des guides de transmission, mais des éléments de base de développements automatisés de services et d'information. La réussite de ces raffinements tient parfois à des facteurs totalement extérieurs à la technologie mais qu'il faut bien prendre en compte, comme les partenaires contractuels de l'institution, les limites temporelles et financières ou les conditions d'utilisation de tel ou tel logiciel. Un standard de métadonnées favorise la qualité lorsqu'à toutes les étapes de la vie du document il facilite (ou ne gêne pas) la valorisation de l'information. A l'inverse, le coût est double lorsque la qualité des métadonnées de départ ne peut parvenir jusqu'à l'utilisation finale : celle-ci se fait sur des métadonnées insuffisantes, et le travail de création de métadonnées pertinentes est perdu. On pourrait parler de **permanence** dans la qualité des métadonnées.

CrossRef

Etant donné que ticTOCs est soutenu par la fondation CrossRef[42], et que ce service utilise le DOI pour améliorer l'identification des documents traités, il est logique de s'intéresser plus précisément à ce Digital Object Identifier et à son institution de tutelle. Le nom de CrossRef vient de « cross reference », en français « référence croisée ». Le principe des références croisées est, pour faire simple, celui d'une interaction entre les documents qui citent et les documents qui sont cités. Si un document en cite un autre, ce dernier en est affecté en retour. Ce fonctionnement rétroactif nécessite une identification pérenne et unique des documents, malgré une multitude d'avatars possibles.

CrossRef a mis au point et développé depuis 2000 un système, le DOI, qui prend une place prépondérante dans le domaine de l'édition électronique. On l'a dit plus haut, l'URL est devenue la métadonnée la plus utilisée. Or comme le soulignait Marin Dacos dans sa conférence sur « La citabilité »[43], une part importante des URL des documents change à très court terme : d'après Google 10% des pages changent d'adresse chaque mois, tandis que la fondation DOI évalue à 1/6 la part des pages qui évoluent tous les 6 mois. Même si ces évaluations ont aussi un rôle publicitaire pour ces deux entités, la mutabilité des documents sur Internet est indéniable et justifie par exemple qu'on signale systématiquement dans une bibliographie la date du dernier accès à la page. Cela faisait dire à Tim Berners-Lee que les URL « sympas » ne changeaient pas[44] : un besoin d'identification pérenne apparaît donc, à plus forte raison dans le domaine de la publication scientifique. Si l'adresse d'un article est erronée, il perd l'attention de tous ceux qui n'iront pas plus loin que le message indiquant que le document n'existe plus ou a été déplacé (la célèbre « erreur 404 »). Il perd aussi un potentiel de citation dont on connaît le prix du point de vue de l'évaluation de la recherche. Le DOI est une forme de réponse à ce besoin de pérenniser les URL, car en principe les éditeurs et la fondation CrossRef s'assurent que l'adresse au format DOI reste valide.

Les amendes perçues par CrossRef sont assez dissuasives pour imposer ce que le bon sens recommande, à savoir la mise à jour aussi fréquente que nécessaire du lien entre le document et son adresse pérenne. Le fait, pour les éditeurs, de cotiser en amont pour le service, et d'engager des frais à hauteur d'un euro par DOI n'incite pas non plus à négliger ce partenariat. Malgré ce coût induit pour les éditeurs, ceux-ci trouvent leur compte dans l'existence de ce résolveur de lien particulier, qui garanti une mesure précise de tous les liens qui amènent à leurs publications. Ces intérêts conjugués, le fonctionnement déjà bien rôdé du DOI et la masse critique déjà atteinte par cet identifiant pérenne expliquent que tous les acteurs de la chaîne de publication scientifique tiennent actuellement compte de l'existence des DOI (Zotero propose par exemple d'importer une notice bibliographique sur la foi du simple DOI).

Mais l'action de CrossRef ne se limite pas à la redirection depuis un DOI vers le document, et il est instructif de voir vers quels usages de métadonnées se tournent ses développements. Le « laboratoire » de CrossRef[45] répertorie les expériences en cours. La prise en compte des InChi ne relève pas des sciences humaines et sociales mais de la chimie, c'est pourquoi nous ne ferons pas d'analyse particulière sur leur prise en compte par CrossRef. En revanche, on peut noter que CrossRef propose son propre résolveur de liens OpenURL, ce qui montre la complémentarité entre les références croisées et

[42] http://www.crossref.org/
[43] (Dacos 2009)
[44] Traduit par Karl Dubost à l'adresse http://www.la-grange.net/w3c/Style/URI
[45] http://labs.crossref.org/index.html

l'enrichissement documentaire fondé sur des métadonnées (CrossRef le dit explicitement d'ailleurs dans sa présentation rapide[46]). Sur un principe similaire, CrossRef tente de mettre en place un agencement, pour les blogs de WordPress par exemple, entre COinS et le DOI, de manière à combiner la rapidité du premier standard avec la stabilité du second[47].

Faisant la même constatation que les fondateurs de Zotero, CrossRef tente aussi de faciliter l'accès au document et à ses métadonnées par le biais d'un module de navigateur, outil essentiel dans les démarches de recherche actuelles. Le but de cette expérience (reprenant les bases du module Ubiquity, disponible à l'adresse http://labs.mozilla.com/projects/ubiquity/) est de détecter les mentions de DOI dans un document et d'en proposer une vue détaillée. Ces deux modules tirent leur fonctionnement d'un moteur de recherche de métadonnées appelé CrossRef Metadata Search. Ils montrent que le DOI est considéré par ses propres inventeurs comme une métadonnée utile à la recherche, mais comme insuffisamment explicite pour répondre à tous les besoins des intervenants humains.

Pour le fonctionnement du DOI comme pour les autres initiatives de CrossRef, il est frappant de voir avec quelle clarté les limites connues sont déclarées. Sur ce standard décrivant une unique métadonnée, l'institution qui l'a mise en place documente la plupart des cas d'utilisation possibles, les rapports envisageables avec d'autres standards et les dysfonctionnements connus avec leurs raisons. Ce n'est pas spécifique au DOI, mais un standard bien **documenté** est susceptible d'être utilisé plus efficacement par tous ceux qui y ont recours.

SYNTHESE DES OBSERVATIONS

Sans prétendre à une parfaite exhaustivité, nous avons au cours des analyses précédentes identifié des critères de qualité pour un standard de métadonnées, et l'impact correspondant sur la qualité des métadonnées. La mention « d'origine de l'observation » ne signifie pas que le critère ne se rencontre que dans ce cadre, mais que l'analyse de la situation a amené une réflexion particulière sur ce critère : parfois même c'est l'absence de ce critère qui a amené à en percevoir l'existence.

Critère de qualité du standard	Impact sur la qualité des métadonnées	Origine de l'observation
Stabilité	Conservation de la qualité d'origine au cours de l'évolution historique globale	*Nature*
Présence en amont	Authenticité des métadonnées par rapport au document et étalon d'origine de la qualité	JSTOR

[46] (Crossref 2009)
[47] https://sourceforge.net/projects/crossref-cite/

Critère de qualité du standard	Impact sur la qualité des métadonnées	Origine de l'observation
Uniformité d'expression	Possibilité de comparaisons plus justes entre les métadonnées	OCLC Calenda (CLEO)
Fluidité	Extraction ciblée des métédonnées utiles dans des contextes appropriés	Postgenomic
Proportionnalité	Adaptation de la finesse des métadonnées aux données d'origine	*REMI* (CLEO) *Homosexus* (CLEO)
Réciprocité	L'augmentation de la qualité des métadonnées valorise les données	*REMI* (CLEO)
Caractère structurant	Qualité exploitée par les automates	*Cybergeo* (CLEO)
Potentialité	Possibilité de conversions conservant la qualité d'origine	Livres de l'IFPO (CLEO)
Fonctionnement par héritage	Rapidité de création de certaines métadonnées	Livres de l'IFPO (CLEO)
Caractère dynamique	Possibilités d'exploitations multiples renforçant la qualité d'origine	*CPA* (CLEO)
Fonctionnement par choix	Processus de vérification diminuant les pertes de qualité	*CPA* (CLEO)
Fonctionnement par réutilisation	Economie de moyens pour une qualité égale	Calenda (CLEO)
Prise en compte du cycle de vie	Conservation de la qualité d'origine au cours de la vie du document	Calenda (CLEO)
Statut de pivot	Possibilité de conversions conservant la qualité d'origine	Zotero
Interprétabilité	Meilleure conservation de la qualité d'origine lors des tranformations grâce à une meilleure compréhension humaine de cette qualité	Zotero
Permanence	Pleine expression de la qualité des métadonnées au long de la vie du document	ticTOCs
Caractère documenté	Conservation de la qualité d'origine plus aisée lors des transformations	CrossRef

L'ensemble des critères de qualité d'un standard sont rarement, voire ne sont jamais, réunis pour un standard donné. Certains de ces critères entrent en effet en opposition mutuelle, amenant les créateurs et utilisateurs de métadonnées à choisir les meilleurs compromis (typiquement, le fonctionnement par héritage et le fonctionnement par choix ne peuvent exister au même moment pour les mêmes métadonnées).

Frédéric Martin, faisant référence au *Guide des bonnes pratiques en OAI* de la National Science Digital Library, rappelait en 2007 douze critères essentiels à la qualité des métadonnées en vue d'une meilleure interopérabilité des descriptions documentaires[48]. Ces critères peuvent être ici mis en rapport avec ce qu'on peut attendre d'un standard de métadonnées de qualité :

Critère de qualité des métadonnées	**Explication développée**	**Correspondance au niveau d'un standard**
Complétude	Sélectionner et utiliser de la façon la plus complète possible un format de description. L'exemple donné par Frédéric Martin serait de se fixer un nombre minimum d'éléments.	Caractère structurant, fonctionnement par héritage
Pertinence	La syntaxe du format sélectionné doit être précise et les informations décrites correctes.	Présence en amont
Provenance	Il est important de mentionner les informations relatives au producteur des métadonnées ainsi que son expertise.	Réciprocité, fonctionnement par choix
Conformité aux attentes	Le traitement des métadonnées (éléments, vocabulaires) doit répondre aux attentes d'une communauté.	Caractère dynamique, fluidité
Cohérence	La normalisation est capitale dans le traitement des métadonnées. Le codage des données doit être réalisé rigoureusement de façon à rendre leur gestion plus linéaire et aisée.	Uniformité d'expression
Cycle de vie	Il ne faut pas négliger que la ressource décrite peut évoluer. Il faut donc anticiper lors du traitement des métadonnées quelles pourraient être les futures modifications.	Prise en compte du cycle de vie

[48] (Frédéric Martin 2007)

Critère de qualité des métadonnées	Explication développée	Correspondance au niveau d'un standard
Accessibilité	Il faut s'assurer de la véracité du lien entre les métadonnées et les ressources décrites qui leur sont relatives. Il faut également que ce lien soit accessible à la communauté cible.	Stabilité, statut de pivot
Granularité	Il faut une seule notice par document, que celui-ci soit numérique ou physique.	Proportionnalité
Indépendance	Les notices doivent être compréhensibles en dehors de leur contexte local pour éviter le recours systématique à des informations extérieures.	Caractère structurant, interprétabilité, caractère documenté
Auto-documentation	Il est important que la notice soit auto-suffisante, excluant toute référence locale.	Caractère documenté
Normalisation du vocabulaire	La normalisation du vocabulaire permet l'interopérabilité entre des notices venant de sources différentes.	Fonctionnement par réutilisation
Constance	Les décisions prises sur les différents points clefs (vocabulaire, traitement des métadonnées) ne doivent pas changer pour conserver la cohésion et la cohérence de l'ensemble des notices.	Permanence

Parfois indirects, les liens entre qualité des standards de métadonnées et qualité des métadonnées elles-mêmes semblent incontestables. On peut également mesurer la multitude des facteurs jouant sur ce qu'on peut appeler la qualité d'un standard. Mais on ne peut qu'être d'accord avec Sylvie Dalbin lorsqu'elle insiste sur la « qualité du modèle initial » dans l'appréciation générale de la qualité des métadonnées[49].

Toutefois le tort des analyses précédentes est de mettre sur un même plan des standards très différents, ne régulant pas les usages sur un plan identique ou ne portant pas sur les mêmes types de données. Cela ne permet pas d'avoir une vue plus globale du paysage des standards, condition d'un meilleur repérage en synchronie et en diachronie.

[49] (Dalbin 2008a)

TYPOLOGIE DES SCHEMAS DOMINANTS DANS L'EDITION SCIENTIFIQUE

Il y a sans doute autant de standards de métadonnées que d'utilisations possibles : une page du site de l'IFLA[50] en recensait en 2005 plusieurs dizaines, et il en existe de nombreux autres dans des secteurs industriels, dans les administrations, dans certaines grandes entreprises.... Même dans un champ qu'on pourrait qualifier de réduit, l'édition électronique ouverte en sciences sociales, les standards utilisables sont donc très nombreux.

Sans qu'on puisse ou veuille en faire un inventaire complet, il est intéressant de caractériser les différences entre les standards les plus utilisés, les plus étudiés ou les plus intéressants dans le domaine d'application choisi. Beaucoup de synthèses sur les métadonnées, à commencer par l'article de Sylvie Dalbin « Métadonnées et normalisation »[51], s'appuient sur une différence entre métadonnées de description, administratives, d'expression des droits.... Bien qu'appropriées pour la plupart des standards de métadonnées, ces comparaisons rencontrent des limites quand il s'agit de prendre en compte les processus plus marginaux ou plus globaux de méta-information, comme les standards d'échanges ou les formats de fichiers. On pourrait ajouter que toutes les métadonnées sont descriptives, mais la description ne porte pas sur le même aspect de l'objet (description des conditions de formation des métadonnées, description du contenu, description des conditions d'accès au contenu, description des autres descriptions).

Pour la réalisation de cette typologie, on préférera s'inspirer des principes qui ont régi la création des FRBR (distinguant Oeuvre, Expression, Manifestation et Item[52]) : nous différencierons donc les standards de conception des métadonnées, les standards d'organisation des métadonnées et les standards d'expression des métadonnées.

Standards de conception

Certains standards proposent une approche conceptuelle de l'information concernant les objets de connaissance (autrement dit, une approche conceptuelle des métadonnées). On appellera donc standards de conception les préconisations à propos de l'appréhension des données (granularité, prise en compte de données statiques ou dynamiques...) et des liens les unissant à des métadonnées (liens uniques ou multiples, champs interactifs on non, métadonnées incluses ou externes...).

Sur les choix opérés par les standards entre différents types de lien unissant données et métadonnées, nous pouvons citer en exemple la distinction formulée par Romain Wenz[53]:

[50] http://archive.ifla.org/II/metadata.htm
[51] (Dalbin 2008b)
[52] cf. http://www.bnf.fr/pages/infopro/normes/pdf/FRBR.pdf
[53] (Wenz 2009), p. 61.

> *« Selon la façon dont elles lui sont intégrées, on distingue les métadonnées « encapsulées » (intégrées par balises dans le corps du document), « englobantes » (caractérisant l'ensemble et en général placées dans l'en-tête du code source), et « externes » (fournies dans un fichier séparé du document). »*

Du fait de la globalité de leurs préconisations, ces standards sont apparentés à des familles dont les membres permettraient aux standards de conception de s'adapter aux besoins fonctionnels.

C'est le cas de la famille des langages à balises. Dans cet ensemble, le XML s'est imposé en raison même de sa conception, très adaptée aux échanges de données indépendants d'un contexte difficile à connaître à l'avance. Alors que les humains sont aptes à appréhender un document mêlant structure et contenu, les machines traitent avec plus de facilité les documents qui séparent les deux, et c'est ce que propose le format XML. Son avantage est d'autoriser néanmoins des balises « explicites », compréhensibles par des humains (on voit là un intéressant compromis entre deux critères de qualité des standards, évoqués plus haut). Par ailleurs, le XML permet à tout utilisateur de définir ses propres standards d'organisation et ses propres standards d'expression (grâce aux DTD et aux XSD), ce qui a affranchi la création de métadonnées des bornes imposées par les normes.

Mais XML n'est qu'un élément de l'ensemble des langages à balises, dont relèvent SGML, HTML ou XHTML. On le comprend, dans le contexte de la publication électronique les langages à balises proposent la plupart du temps des métadonnées encapsulées, même si par exemple le HTML ou la TEI conservent une certaine importance à l'en-tête du fichier. Le XHTML, qui est en fait du HTML à la syntaxe rigoureuse, permet un contrôle bien meilleur des données et de leur présentation. Il est donc très utilisé, ce d'autant plus que les transformations automatiques s'y appliquent avec beaucoup moins de problèmes.

Le cas de RDF est exemplaire de la distinction que nous esquissons entre standards de conception et standards d'organisation, au sens où l'une des applications les plus logiques de ce standard se fait avec un format XML (RDF-XML), mais en tant que *framework* il pose des principes qui peuvent s'appliquer à de toutes autres méthodes de méta-information (voir à ce sujet le guide très illustré du W3C[54]).

Le RDF, comme l'indique son acronyme développé, est en effet un « framework », un modèle de création de métadonnées. Les triplets définis sont réutilisables à volonté, en dehors de toute démarche spécifique d'échange par lots de métadonnées. C'est une modification importante, avec de nombreuses implications techniques, mais sur le plan de l'organisation et de l'expression des métadonnées, le fait de passer des doublets (champ / contenu) aux triplets (sujet / prédicat / objet) n'est peut-être pas un bouleversement aussi fort qu'on le dit. L'établissement des règles de fonctionnement, des règles d'interprétation des prédicats en particulier, ne se génère pas spontanément mais reste extérieur aux données et aux métadonnées elles-mêmes, selon une logique déjà identifiée dans un autre contexte par Kurt Gödel (cf. à ce propos l'article de l'encyclopédie Encarta sur « Les théorèmes de Gödel »).

[54] (W3C 2004)

La bibliothèque du Congrès des Etats-Unis, avec ses derniers formats (MODS, METS, MADS) a proposé une organisation des formats de métadonnées les uns par rapports aux autres, sur la base du XML et des espaces de nom. Si elle peut être déduite et si elle a fait l'objet de commentaires éclairés, la logique à l'oeuvre derrière ces formats n'a pas été pourvue d'une documentation conceptuelle. METS a pu être décrit comme un « méta-format », car il décrit les formats de métadonnées comme des objets numériques (comme le souligne Catherine Morel-Pair[55]); cela reste néanmoins dans le cadre des emboîtements propres aux langages à balises, appliqué avec une grande habileté comme on le verra.

XMP, autre standard fondé sur les balises du RDF/XML, est très lié à l'utilisation massive des fichiers PDF[56]. Son principe diffère toutefois légèrement des standards similaires. Afin de répondre au besoin d'enrichir en métadonnées un fichier numérique donné, le XMP englobe ledit fichier et y insère sous forme de « paquets » les métadonnées souhaitées. Cela permet notamment les enrichissements a posteriori, et le traitement automatisé pour des types de documents dont les pratiques de métadonnées ne sont pas suffisamment structurées.

Open Archival Information System (OAIS), en revanche, se place bien au niveau du standard de conception, car les métadonnées y sont considérées comme des éléments dynamiques, inclus dans un cycle de vie et rendant nécessaire d'anticiper des évolution techniques et des adaptations à long terme. L'application de cette norme ISO 14721 se fait de façon privilégiée grâce à PREMIS, qui reprend le cadre conceptuel de l'OAIS et qui est à la fois standard d'organisation et standard d'expression. L'édition électronique ouverte est concernée par ces standards visant à une conservation pérenne des documents, car ce sont les revues qui font l'objet des développements les plus avancés en matière d'archivage pérenne, en particulier dans les grandes bibliothèques nationales. Comme les principaux formats de métadonnées en usage sont pris en charge, les éditeurs qui s'y conforment n'ont néanmoins pas de démarche particulière à effectuer, leurs éventuelles obligations de dépôt légal mises à part.

OpenURL, imaginé par Herbert Van de Sompel, inclut pour sa part une réflexion sur le processus de transmission de métadonnées, qui sont en quelque sorte « augmentées » parallèlement à la transmission. On peut s'appuyer sur l'excellent article de Ann Apps et Ross MacIntyre pour en détailler les grandes lignes[57]. OpenURL repose sur l'identification de sources et de cibles. Les sources sont les « liens de » (lieux d'où proviennent les documents), et les cibles sont les « liens vers » (lieux vers lesquels l'utilisateur peut se diriger pour les consulter). OpenURL peut être mis en application selon deux principales organisations, l'organisation d'origine inventée par Herbert Van de Sompel, ou la norme NISO Z3988. Plus largement, le principe ici employé est celui de la résolution de lien, qu'on pourrait aussi nommer, en référence à l'étymologie, « solution » de lien.

Les liens sont « cassés » pour permettre à un service tiers de fonctionner, au bénéfice (en théorie, et souvent en pratique) de l'utilisateur désireux de consulter une source. La plupart des identifiants pérennes sont des standards d'expression qu'il ne convient pas d'évoquer ici, mais le cas du DOI est particulier et doit être abordé dans le cadre des

[55] (Morel-Pair 2007)
[56] (Roszkiewcz 2008)
[57] (Apps et MacIntyre 2006)

standards de conception. En effet, c'est par résolution de lien que le code DOI est exploité, ce qui permet d'adapter l'usage à la situation de la requête.

Dans l'ensemble des standards de métadonnées particulièrement « fluides », outre OpenURL et ceux exploitant les résolveurs de liens, on peut évoquer les « fils » d'échange de métadonnées, dont les plus connus sont les fils RSS. Les principes de la syndication, ou la conception des rétroliens, sont finalement assez simples : à chaque appel à une source identifiée par une URL, quelques données, encapsulées dans un nombre réduit de métadonnées, sont transmises et insérées dans la page d'accueil. Ce principe se décline en divers formats impliquant organisation de la transmission, quantités et types de métadonnées et de données transférés : RSS 1.0, RSS 2.0, ATOM....

Mais ces liens entre une page « cliente » et une page « serveur » font partie de l'ensemble des normes ou protocoles de transmission de données. Cet ensemble peut être considéré comme représentatif d'une approche particulière de la question des métadonnées. Ces formats autorisent les échanges organisés des métadonnées, en réponse à des besoins particuliers sur ce point. Le protocole Z3950 par exemple part du principe qu'il y a un intérêt à échanger des notices bibliographiques sans forcément échanger les documents correspondants. Deux nuances dans ces standards de transmission sont repérables : les métadonnées sont transmises seules, comme nous venons de le dire pour le protocole Z3950; ou bien les métadonnées sont accompagnées de tout ou partie du document d'origine, comme dans le cas de PAM et du RSS.

On pourrait dire qu'à l'autre bout de l'échelle de la complexité descriptive, les formats MAchine Readable Catalog (MARC), intimement liés à la norme ISO 2709, forment une toute autre famille de métadonnées, avec de nombreux membres conçus de façon similaire (MARC21, UNIMARC, INTERMARC). A l'origine en effet MARC proposait dans des fichiers (forcément séparés du document d'origine) des renseignements bibliographiques. Chaque étiquette de donnée était repérée par un code alphanumérique. En fonction des besoins des bibliothèques et réseaux de bibliothèques, cette matrice a généré des variantes assez semblables : les champs obligatoires et optionnels, les règles d'expression du contenu, des champs spécifiques à des types de documents constituent les principales nuances entre UKMARC, INTERMARC, UNIMARC.... C'est afin de transmettre ces données que la norme Z3950 a été forgée; l'existence même de cette norme prouve que la conception d'origine de MARC requérait une adaptation aux nouvelles possibilités et nouvelles nécessités en matière de transmissions de métadonnées.

Par la suite, c'est à partir de certains standards d'organisation de MARC qu'ont été créés des langages à balises. XML-MARC par exemple a adapté MARC21 à l'organisation du XML. Le standard d'organisation fondé sur la séparation physique entre les métadonnées et les données correspondantes n'a pas disparu (on rencontre souvent cette pratique dans certains réservoirs OAI-PMH, détournés en ce sens de leur conception d'origine), mais il n'est plus en usage dans le domaine de l'édition scientifique en sciences humaines et sociales. C'est pourquoi dans notre cadre d'étude ce sont plutôt les langages MARC qu'il serait opportun de prendre en compte, pour des conversion et des échanges de métadonnées.

Les principes des Microformats forment une dernière conception de génération et d'utilisation des métadonnées. La difficulté d'expression particulière les concernant est qu'il s'agit du nom d'une initiative, mais ce nom est si évocateur qu'on est tenté de l'employer comme un nom commun. L'idée principale est de limiter la définition d'un format de métadonnées à très peu de labels (de 1 à 6 environ), d'où ce nom de microformat. La légèreté qui en résulte permet de placer ces balises sans contenu visible dans n'importe quelle page web; un module de navigateur (Operator pour Firefox, par exemple) permet ensuite à l'utilisateur de récolter ou d'exploiter ces métadonnées. Les microformats les plus utilisés *volontairement* sont hCard, hCalendar et geo (le microformat tag récupérant tout le contenu des balises pourvues d'un attribut « rel », très fréquentes par exemple sur les blogs mais non destinées à cet usage à l'origine). On pourrait dire qu'il n'y a pas d'intermédiaire entre le principe des microformats et la traduction correspondante en expression de métadonnées : pas d'organisation régulée dans les pages, pas de limite ou d'ordre dans les microformats pris en compte.

Potentiellement très utiles et assez simples à mettre en oeuvre, les microformats devraient connaître un fort développement, sous réserve d'une meilleure régulation des usages. Dans le double contexte de l'édition électronique ouverte et des sciences humaines et sociales, il sera intéressant de voir quelle place ils prennent par rapport aux autres méthodes de circulation des métadonnées.

Standards d'organisation

On désignera par standard d'organisation les schémas de métadonnées qui s'attachent à décrire les outils à disposition pour l'expression contrôlée des caractéristiques du document.

RDFa est un mode d'application de plus en plus utilisé des concepts à l'oeuvre dans le RDF. Comme l'indiquent Emmanuelle Morlock et Raphaël Tournoy, les moteurs de recherche l'utilisent pour enrichir les résultats avec des données liées[58].

Toujours sur le modèle conceptuel de RDF, Simple Knowledge Object Schema (SKOS) définit des règles d'organisation pour des thésaurus et des bases sémantiques. Il est déjà expérimenté avec un certain succès dans les sciences humaines et sociales : en philosophie, le travail de distinction conceptuelle gagne en richesse avec le dialogue autour de bases SKOS peu à peu constituées. La distinction entre SKOS, qui s'attache à décrire les rapports conceptuels sur la base des triplets Sujet-Prédicat-Objet, et OWL, qui définit des ontologies selon l'alternance entre classes et instances, rend ces deux standards d'organisation complémentaires. Notons que OWL ne peut pas être considéré, dans notre typologie, comme un standard de conceptualisation des métadonnées mais comme une modélisation plus large, touchant à l'information dans son ensemble.

Les fondements conceptuels les plus employés restent néanmoins ceux du XML.

[58] (Morlock et Tournoy 2009)

L'ensemble des standards relevant de la TEI (TEI Lite, TEI P5, etc.) sont fondés sur les principes du XML. Néanmoins on pourrait considérer que la TEI comporte une part d'innovation conceptuelle par son approche particulière de la granularité des objets sur lesquels faire porter des métadonnées, autant que par sa prise en compte du processus éditorial. Reste que son principe de fonctionnement est le même que celui des autres langages à balises, même si ce standard encadre incomparablement mieux la diffusion des documents que le HTML par exemple. Ce qui nous fait ranger la TEI dans les standards d'organisation est la distinction qu'elle invite à faire entre une en-tête de métadonnées, applicables à l'ensemble, et les métadonnées encapsulées grâce aux balises. Elle se décline par ailleurs en de très nombreux standards d'expression, qui héritent de ses règles d'organisation.

Le Dublin Core Abstract Model[59], toujours sur le principe des métadonnées encapsulées grâces à des balises XML, est une recommandation proposée par le DCMI pour faciliter l'agencement de profils d'application différents issus du Dublin Core. En cela c'est bien un standard d'organisation, et ce qu'on devrait désormais appeler *les* Dublin Core (qualifié ou non, déclinés selon des profils d'application) relèvent de la catégorie des standards d'expression.

Le format METS est de tous les standards issus de l'héritage MARC celui qui est le plus employé dans l'édition électronique (ouverte ou non) : les plates-formes Erudit, Persée ou Revues.org utilisent toutes la capacité de ce standard à gérer des éléments composites, à schématiser les structures et à s'adapter aisément aux modifications du tout ou des parties. Par essence il s'agit d'un standard d'organisation, un contenant pour des modules de métadonnées (certains de ces modules pouvant eux-mêmes être des standards d'expression de métadonnées). Réunissant de nombreux critères de qualité identifiés plus haut (proportionnalité, fonctionnement par héritage, caractère structurant), il montre au passage que le monde des bibliothèques peut participer activement à l'élaboration des standards d'avenir dès que s'établit un dialogue constructif entre l'héritage bibliographique et les innovations de l'ingénierie documentaire. La seule limite évidente du standard METS est qu'il est difficile à exploiter directement pour une entité disposant de peu d'expertise ou d'expérience en matière d'ingénierie informatique. Le fonctionnement du METS est décrit avec beaucoup de pédagogie dans une présentation de Rick Beaubien disponible sur le site de la bibliothèque du Congrès[60].

Ainsi qu'on l'a vu plus haut, OpenURL se décline en 2 branches, l'organisation d'origine (détaillée, pour la dernière version en date, par Van de Sompel, Horchstenback et Beit-Arie[61]) ou l'organisation normée. Mais comme OpenURL est un format ouvert, tous les développements sur cette base sont possibles. COinS (Context Object in Span) est justement l'un des standards d'organisation qui prend pour fondement la conception d'OpenURL. En effet, sur le principe d'un point d'accès au document enrichi en métadonnées, COinS permet d'adapter à la situation de l'utilisateur l'outil qui sera utilisé pour récolter ces métadonnées. Ceci permet notamment de récupérer toute une bibliographie ou tous les articles d'un numéro de revue en une seule opération.

PAM est également un standard d'organisation, puisqu'il encadre les opérations à effectuer pour diffuser et récolter les métadonnées dans un contexte d'édition. Il indique

[59] http://dublincore.org/documents/abstract-model/
[60] (Beaubien 2007)
[61] (Van de Sompel, Hochstenback, et Beit-Arie 2001)

également les standards d'expression des métadonnées à utiliser : PRISM, Dublin Core, et des balises PAM. On pourrait considérer que PAM est la formalisation des principes d'organisation sous-jacents à PRISM, qui lui est antérieur. La valeur de ce standard pour l'édition électronique est exposée avec une grande clarté par Tony Hammond, qui y justifie l'emploi de ce format dans le cadre des serveurs OAI de *Nature*[62]. Le point qui ressort particulièrement est que justement pour les silos OAI, les véhicules de transmission de métadonnées doivent être interprétables par le biais de définitions de schémas en XML (XSD) et non en SGML (comme c'est le cas pour les DTD).

ONIX[63] s'apparente plutôt à un format de fichier même s'il reprend les principes du XML. Il est très utilisé par les éditeurs dits « commerciaux » pour les monographies. Dans le secteur de l'édition scientifique ouverte, sa principale fonction, la mise en vente du document, n'a plus de sens, ce format n'est donc pas utilisé.

PREMIS, format de type XML dont l'évolution est prise en charge par la bibliothèque du Congrès, fixe des règles d'organisation des métadonnées dans un but de conservation pérenne. Il propose d'autre part des méthodes d'expression de métadonnées qui lui sont propres, toujours dans l'idée d'améliorer la conservation conjointe des documents et des métadonnées. La question de la conservation à long terme des documents est plus prégnante dans les bibliothèques que dans le secteur de l'édition scientifique; c'est pourquoi même si les sciences humaines et sociales ont un rapport particulier au temps documentaire, nous ne détaillerons pas spécifiquement ce format dans le cadre de cette étude.

Standards d'expression

Un standard d'expression peut être défini comme un ensemble de cadres d'expression régissant directement le codage des métadonnées.

On peut subdiviser cette catégorie de standards en deux, les standards de grammaire et les standards de vocabulaire. Tous ce que le monde des bibliothèques nomme « vocabulaire contrôlé » relève de ces derniers : les RFC3066 qui donnent les chaînes de caractères à utiliser pour désigner les langues, les thésaurus, les autorités RAMEAU, le MeSH, le LCSH, le fichier VIAF, etc. On entend donc par là les règles qui s'imposent à l'expression du contenu des métadonnées.

Parmi ces standards de vocabulaire, dans le domaine des sciences humaines et sociales, les autorités de lieux du Getty sont utiles et de plus en plus employées. En ce qui concerne l'identification du contenu d'un document, dans le cadre d'une internationalisation croissante des autorités il faut certainement préférer VIAF à une référence au seul RAMEAU. En concordance avec l'usage croissant du PRISM dans l'édition électronique, il est intéressant de noter que ce format (dans sa version 2.1) permet l'expression de références en XML Topic Maps (XTM), qui est une alternative pratique aux mentions d'autorité matière plus riches mais plus complexes à manipuler.

[62] (Hammond 2009)
[63] Ce standard international de commercialisation du livre est décrit à la page http://www.editeur.org/8/ONIX/

Les standards qu'on pourrait appeler « grammaires de métadonnées » précisent les règles d'expression des champs de métadonnées. Par exemple, dans le cas du Dublin Core, il est possible de spécifier pour la langue qu'on utilise une méthode de déclaration. L'intitulé de la méthode utilisée elle-même est du ressort d'un vocabulaire contrôlé, le fait de pouvoir attribuer à un champ une propriété vient du standard d'organisation qu'est le XML, mais le fait de pouvoir faire cette déclaration pour ce champ spécifique est encadré par la grammaire du Dublin Core. Au travers de cet exemple on comprend aussi que de nombreux standards, dont le Dublin Core, relèvent du standard d'expression mais sont intimement liés à des principes d'organisation des métadonnées. C'est logique étant donné que dans notre typologie les standards de grammaire (règles sur l'expression des champs) sont entre les standards de vocabulaire (règles sur le contenu) et les standards d'organisation (règles sur l'utilisation des champs).

C'est aussi le cas de PRISM. PRISM est une grammaire de métadonnée particulièrement adaptée au contexte étudié, car comme son nom l'indique (il s'agit de « Publication Requirements ») les nécessités propres à la publication ont spécifiquement été prises en compte. La déclaration se fait par le biais des espaces de nom, ce qui permet de combiner plusieurs grammaires. Il faut noter que l'utilisation, à un niveau plus abstrait, du PAM, restreint les balises à disposition pour la description en PRISM des documents.

Dans le domaine de l'édition scientifique, le standard NLM[64] est assez répandu. Il a même été envisagé un temps par Google pour les métadonnées de Google Scholar. Ce format est plutôt riche, bien adapté à la publication scientifique, et plus précis de ce point de vue que PRISM qui s'intéresse à la publication en général. Toutefois le format ne joue pas, c'est le moins qu'on puisse dire, la carte de l'interopérabilité car tout ce que propose la NLM en partenariat avec la NCBI prend pour base certes XML, mais uniquement avec un vocabulaire NLM. Pas de crosswalk évoqué, pas d'outils de transformation d'un set de métadonnées vers NLM. En revanche, le processus de ré-ingénierie des métadonnées de citation lors de l'introduction d'une nouvelle référence est particulièrement intéressant et toute opération similaire dans un autre univers de métadonnées pourrait avec intérêt s'en inspirer.

Il est difficile désormais de favoriser l'accès aux publications de la recherche sans faire une place particulière au référencement vis-à-vis des moteurs de recherche, de Google en particulier. Ce dernier est en effet le premier outil de la recherche en nombre d'utilisations, et progresse même dans un secteur traditionnellement aussi porté sur la consultation directe des dépôts documentaires que la physique des hautes énergies[65]. C'est pourquoi la grammaire d'expression des métadonnées de Google Scholar, finalement assez récente en comparaison des formats tels que Dublin Core ou surtout la TEI, connaît une croissante très rapide.

Google jouant plus sur les capacités de traitement liées à l'ingénierie informatique que sur une politique de métadonnées structurée pour valoriser les informations, non sans succès à court terme, cela nous invite à rester vigilant sur les préconisations qu'il sera possible de formuler en matière de métadonnées. La qualité intrinsèque d'un standard ne faisant pas tout, il convient donc de s'attacher à tirer les conclusions que ce travail de

[64] www.**nlm**.nih.gov/libserv.html
[65] cf. Anne Gentil-Beccot, http://arxiv.org/abs/0804.2701

synthèse appelle pour les éditeurs en sciences humaines et sociales, avec toujours à l'esprit la situation particulière de l'édition électronique ouverte.

Préconisations : des conceptions aux usages

Dans un premier temps, en fonction des constatations concrètes et des possibilités offertes par l'environnement technologique de l'édition électronique ouverte, nous formulerons des propositions d'évolution des formats utilisés par un éditeur en sciences humaines tel que Revues.org, tout d'abord selon les types de documents puis de façon générale. Pour certains standards en cours d'implémentation ou envisagés, nous pourrons ensuite donner quelques principes visant à conserver la pertinence et la cohérence des métadonnées. Des règles de conversion d'un format à l'autre sont également utiles, c'est pourquoi nous évoquerons les plus fondamentales en ayant à l'esprit les critères de qualité identifiés précédemment. Des « mappings » détaillés, présentés en annexe, complèteront ces propositions.

PROPOSITIONS D'EVOLUTIONS SUR LE PLAN DES STANDARDS

Selon les documents concernés

Certains standards sont plus adaptés à la situation de publication des articles de revues, d'autres au contexte de l'édition d'un carnet de recherche. C'est pourquoi nous traiterons des standards de métadonnées préconisés en fonction des documents sur lesquels ils portent.

Articles de revues

Le Dublin Core est utilisé de façon irrégulière par les revues et les livres de Revues.org, de façon semblable à ce qu'on peut constater dans l'ensemble des publications scientifiques. Ce caractère de disparité a sans doute participé au déclin du Dublin Core dans les usages : il est appliqué presque systématiquement par principe, mais rarement pour son utilité descriptive ou pour la visibilité qu'il pourrait conférer. Pour augmenter la qualité du traitement en Dublin Core nous pouvons rejoindre Frédéric Martin sur la nécessité d'un ensemble peut-être moins large, mais plus constamment utilisé (Frédéric Martin parle de « complétude »). Cela se traduit par exemple par une obligation (ou tout au moins une très forte recommandation), faite aux auteurs ou aux comités de rédaction des revues, de remplir les champs essentiels à l'identification et au traitement de l'article ou du chapitre. Sur un plan plus global, nous pourrons aussi revenir sur l'utilité de conserver le Dublin Core dans le cadre des projets à long terme du TGE ADONIS et des infrastructures de la recherche à l'échelle européenne.

Le XHTML est certainement une meilleure option de présentation des pages que le HTML, en raison des possibilités de conversion afférentes au caractère strict du XHTML. L'arrivée prochaine du HTML 5 pourrait alimenter la réflexion, dans la mesure où sa capacité de valorisation des documents semble particulièrement efficace. Mais peut-être sera-t-il alors envisagé par le W3C une évolution parallèle du XHTML, ou une version XML du HTML 5; comme souvent la stabilité de la base présente n'incite pas à conseiller un changement sans un temps d'expérimentation préalable.

L'emploi de la TEI est justifié tout d'abord du point de vue des partenariats de continuité avec Cairn[66]: cet éditeur de revues en SHS tire profit du balisage en TEI pour valoriser ses documents, c'est donc dès leur création que les articles sont balisés en TEI, l'extension de la TEI à toutes les revues permettant finalement une amélioration homogène de la précision des métadonnées. D'autre part, l'application aux documents de balises issues de la TEI ouvre à plusieurs disciplines des sciences humaines et sociales des champs de recherche supplémentaires : la linguistique, l'histoire, les sciences de l'information, la sociologie, toutes représentées parmi les revues de Revues.org, disposent ainsi de corpus supplémentaires.

Les lacunes des formats actuellement employés dans l'expression des droits moraux, des droits patrimoniaux et des conditions d'utilisation trouveraient une solution adéquate dans l'emploi de PRISM, aussi bien au niveau des articles que pour les monographies.

Dans la perspective d'une meilleure visibilité, l'utilisation de COinS dans les revues en ligne serait légitime, d'autant plus que les notices des articles sont versées à terme dans le SUDOC qui emploie également ce format. Or pour l'instant ce n'est pas le cas dans les observations faites sur Revues.org. C'est d'autant plus dommageable que par exemple de façon quasi-systématique les autres articles rédigés par un auteur sont cités en fin de document; ce serait l'occasion de favoriser les rebonds en valorisant les métadonnées jusque là « silencieuses ».

L'adéquation des balises d'en-tête avec les usages des moteurs de recherche doit être régulièrement vérifiée afin de ne pas se couper d'un public élargi : en particulier, les changements éventuels des labels reconnus par Google Scholar devront être suivis. Enfin l'insertion des microformats hCard (pour les auteurs) et geo (pour les lieux) pourrait être intéressante, même si le développement correspondant dans le moteur de Lodel n'irait pas sans poser des difficultés.

Chapitres de livres

De nombreuses préconisations concernant les articles de revues restent valables pour les chapitres de livres. Toutefois si nous nous concentrons sur les aspects plus spécifiques à cette forme éditoriale, il est possible de proposer, sans un surplus de traitement démesuré, les deux granularités pour les livres : métadonnées pour chaque chapitre et métadonnées pour le livre dans son ensemble. Les métadonnées pour les chapitres sont encadrées par des standards adaptés à la situation actuelle, à la nuance près que l'emploi de la TEI peut apporter un véritable gain qualitatif.

[66] http://www.cairn.info/accueil.php?PG=START

Dans le contexte technologique qui est le notre, le choix du XHTML est crucial dans le cas des monographies numériques. C'est en effet sur la base du XHTML que la génération de livres au format EPUB est la plus aisée (il nous a été donné de le constater dans un atelier d'Adrien Gardeur organisé dans le cadre de l'Université d'Eté du CLEO), et même un poids lourd de la documentation numérique tel qu'Adobe soutient le développement de ce format ouvert. La seule limite de ce potentiel réside dans la nécessité de définitions de publication pour les différentes plates-formes, ce que ne nécessite pas par exemple le PDF; mais l'existence même de cette possibilité fait toute la valeur du standard qu'est le XHTML.

L'importance de l'iconographie dans le fonctionnement de certaines monographies électroniques implique de penser à la fois aux références mutuelles que doivent porter le texte et l'image, et aux dangers techniques que représentent les incarnations ultérieures des documents. Se lancer sans précaution dans la diffusion sur téléphones mobiles est une entreprise vouée à l'échec pour la majeure partie des documents géographiques, historiques, ethnologiques ou sociologiques, dans la mesure où l'image fixe ou animée peut apporter un avantage considérable dans la compréhension des phénomènes analysés. Si le lien organique entre les matières différentes constitutives des documents est rompu, la synergie mise en place avec une première structuration sera vaine. C'est pourquoi, a minima, il faudrait s'assurer de la complétude des métadonnées pour les ressources iconographiques et de l'existence systématique d'équivalents textuels en cas de problème à l'affichage. Si les images sont par ailleurs disponibles sur Internet et disposent d'une URI, il peut être tentant de placer cette URI comme alternative à l'image « encapsulée » dans le texte.

Revues.org s'étant lancé dans l'utilisation du DOI pour ses documents, on peut souligner l'intérêt, peut-être paradoxal de prime abord, d'attribuer un DOI non aux chapitres d'un livre mais au livre dans son ensemble. Deux éléments nous amènent à cette proposition : puisque le DOI doit pointer vers un élément toujours disponible, si l'un des auteurs fait jouer une forme de droit de retrait ou limite les droits d'accès à « sa » partie du livre, aucune modification ou suppression de DOI ne sera nécessaire, seul le document lui-même devra être amendé. L'autre raison est qu'il est toujours plus facile, dans le cadre imposé par le DOI, d'augmenter la granularité si le besoin s'en fait sentir que de la diminuer.

Articles de blog

Dans le cadre de la valorisation des articles de blog, la catégorisation qu'effectue Postgenomic sur les documents pourrait être utilisée par Revues.org. De façon générale, les chercheurs peuvent et doivent être associés à la politique des métadonnées de leur éditeur. La plate-forme Hypothèses est peut-être la meilleure porte d'entrée par laquelle engager clairement ce progrès vers une co-gestion des standards de métadonnées: pour les rédacteurs de carnets de recherche qui le souhaitent, un autre choix de standards peut être fait. Au lieu des seuls XHTML, COinS et RSS, des chercheurs peuvent s'engager sur une voie plus officielle en expérimentant pour leurs billets des métadonnées en Google Scholar, ou en déposant leurs documents dans un entrepôt OAI. Un blog comme *Homosexus* pourrait proposer un fichier EPUB à ses lecteurs désireux de consulter les

différents chapitres du « livre sans titre » sur une liseuse. De façon générale, pour mettre en relation l'activité de recherche et les produits de la recherche, l'agencement des standards de publication (revues et monographies) et des standards de blogging par une coordination accrue pourraient rencontrer un écho très favorable auprès des chercheurs.

Annonces d'événements et appels à contribution

Dans un courriel récent aux personnels du CLEO Marin Dacos signalait la nécessité de mieux organiser l'écheveau des fils de suivi des flux que proposaient les différentes plates-formes Revues.org, Hypothèses et Calenda. Le format Open Content Syndication (OCS), hérité du RDF et du XML mais encore peu connu, permettrait apparemment de répondre à ce besoin de structuration améliorée[67]. Il sera surtout intéressant de s'inspirer du guide de bonnes pratiques sur les RSS de tables des matières, rédigé en commun par ticTOCs et CrossRef, et paru fin 2009[68].

Le point le plus important, dans une logique de valorisation réciproque des différents « rouages » du CLEO, serait de faire une continuité, un lien, entre les événements annoncés dans Calenda et les publications scientifiques de Revues.org. La continuité avec la plate-forme Manuscrits serait évidemment à mettre en place de façon similaire. Le lien avec Hypothèses serait plus discutable et surtout plus difficile à assurer. Pour assurer ce lien, le METS paraît s'imposer dans un premier temps, avec sans doute l'emploi de PRISM pour rendre compte des différentes étapes de génération du document. Mais on voit rapidement l'intérêt que pourrait avoir une conversion vers le web sémantique, l'événement et le document qui en est issu se trouvant liés par le graphe des prédications.

Esquisse d'une politique globale des métadonnées tirant les conséquences des analyses précédentes

De toutes les conceptions en matière de métadonnées, l'insertion au plus près des documents est vraiment la méthode la plus intéressante pour l'édition électronique (sur le mode englobant du XMP pour les images et éléments audio-visuels, ou par le biais des balises pour ce qui relève plutôt du texte). Un point mérite notre attention, il s'agit de l'expression des droits liés au document et à son exploitation. Dans le contexte d'un accès que Marin Dacos définit comme libéré « en amont », il est crucial de signaler les mentions d'autorité et les conditions d'utilisation de façon explicite. Car trop souvent les lecteurs / utilisateurs de la ressource pensent que le texte est utilisable sans contrainte. Trop souvent surtout ils considèrent que l'image ou le son tirés d'un document scientifique en libre accès sont comme « libre de droits ».

Il est donc nécessaire de procéder à une mention systématique des titulaires de droits, explicitement et dans les métadonnées, en s'appuyant sur les standards qui le permettent. L'utilisation de PRISM est dans l'exacte ligne de ces besoins, grâce au module dédié à

[67] Voir à ce sujet le rapport de Ian Davis signalé par (Cover 2001)
[68] (Bilder et al. 2009)

l'expression des droits proposé par ce standard. Mais puisque l'expression des droits doit être reçue par les utilisateurs, il faut aussi penser à d'autres modes de cette expression. Puisque les bibliothèques ne peuvent généralement pas utiliser le PRISM, des indications pourraient par exemple leur être proposées dans des formes explicites ou dans des formats plus traditionnels.

Parmi les tendances de fond dans le domaine de l'édition électronique ouverte, le « risque » est grand que se développe dans les années à venir la lecture des documents sur des appareils portables (liseuse, téléphone portable, agenda électronique, console de poche...). En conséquence, parmi les métadonnées des images et autres documents multimédia insérés dans le corps du texte, il faudrait toujours inclure l'identification d'un ancrage de façon à ne pas perdre la pertinence de cet ancrage lors des transferts et conversions vers des périphériques physiques différents. On l'a vu cette question est particulièrement aiguë dans les transformations matérielles impliquées par le passage du web à d'autres supports de lecture plus récents. La conversion de fichiers EPUB en MOBI puis en AZW, déjà évoquée[69], pose des problèmes particuliers dans ce domaine selon la note technique d'Adobe.

La question des images dans les textes nous amène aussi à considérer celle des mélanges de standards dans un document donné, ou de la multiplication des fichiers pour un document unique. L'emploi de plusieurs standards conjoints dans un fichier tend à se généraliser, grâce notamment à l'utilisation massive des espaces de nom; mais comme cela alourdit les documents, le choix est parfois fait d'éclater le document en plusieurs parties, dont le « corps » du document et des documents annexes. Toutefois pour conserver le lien entre ces fichiers, les mentions d'URL ne donnent pas assez de garantie.

Sans que la situation de l'édition électronique ouverte soit particulière de ce point de vue, on peut envisager deux issues pour les problèmes de gestion des briques éparses dans un document publié, avec les implications souvent néfastes sur le plan de la qualité des métadonnées:

- l'utilisation du METS comme véhicule structurant, déjà très répandue pour les publications traditionnelles de la recherche en sciences humaines (articles, monographies en ligne) mais qui peut être prolongée pour absorber les autres formes de diffusion de l'information scientifique
- l'emploi de charnières sémantiques en RDF, comme le proposent Kai Eckert, Magnus Pfeffer et Heiner Stuckenschmidt, qui permettrait d'après eux de réifier les métadonnées et donc de les traiter plus systématiquement lors des opérations de « crosswalk » impliquées par les agrégats[70].

De façon générale il est utile de suivre l'apparition de profils d'application des formats dominants adaptés à la situation de l'édition électronique ouverte et / ou des sciences humaines et sociales. On peut indiquer par exemple l'existence d'un profil d'application dédié à l'édition électronique scientifique, SWAP[71]. Il peut être aussi légitime de construire, si possible dans le cadre de partenariats, des profils d'applications répondant aux besoins, pour des formats de métadonnées « légers ». Par exemple, les recommandations de CrossRef et ticTOCs concernant les fils RSS de tables des matières

[69] (Adobe 2009)
[70] (Kai Eckert, Magnus Pfeffer, et Heiner Stuckenschmidt 2009)
[71] http://www.ukoln.ac.uk/repositories/digirep/index/Eprints_Application_Profile

pourraient être mises en adéquation avec les spécificités stylistiques et techniques des sciences humaines et sociales.

Une autre veille à effectuer avec soin concerne les initiatives ou services de conversion. Le fait de ne pas utiliser le format ONIX n'empêche pas de savoir, à titre d'illustration, que le service « Crosswalk » d'OCLC peut effectuer la conversion depuis et vers ce type de métadonnées en divers standards plus intéressants du point de vue de l'édition électronique ouverte : Dublin Core (qualifié ou non), MARC, MODS, etc.

Un autre principe général d'encadrement de la politique de métadonnées dans l'édition électronique ouverte est la nécessité de réemployer autant que possible les données. Dans l'idéal ce réassort devrait se faire automatiquement par reconstruction de prédications indirectes (dont un modèle possible est celui des syllogismes); mais pour l'instant il peut s'agir d'un signalement fait par les auteurs eux-mêmes, informés préalablement des outils mis à disposition. Concrètement, il est en effet dommageable que de très riches informations soient disponibles à la lecture pour les individus, mais que ces mêmes informations ne soient pas utilisables par des automatismes pour faciliter le repérage, le traitement et la construction d'informations. La mise en relation, encadrée par des outils lors de la soumission des articles, de lieux évoqués par les géographes ou par d'autres chercheurs avec la base de données du Getty (thésaurus TGN) permettrait de générer automatiquement un index géographique pour l'ensemble de la plate-forme de publication. Chaque « facette » de recherche gagnée, même si elle implique des contraintes réparties sur l'ensemble des acteurs de la chaîne éditoriale, valorise conjointement les métadonnées de chaque document et de l'ensemble des documents.

Dans le contexte des sciences humaines et sociales, les principaux points forts d'une stratégie de réutilisation des données pourraient être les lieux géographiques, les périodes historiques (en tenant bien compte des difficultés liées aux modèles de représentation des dates et des durées), les personnes et/ou les collectivités. La prise en compte des sujets ne pourrait faire l'objet d'un traitement systématique semblable que plus tard, lorsqu'auraient été rôdées les méthodes techniques et humaines concernant les points forts identifiés ci-dessus.

CADRES GENERAUX DE LA CARTOGRAPHIE DES METADONNEES APPLICABLE A L'EDITION SCIENTIFIQUE « OUVERTE » EN SCIENCES HUMAINES

On désigne par cartographie ce que la langue anglaise nomme « mapping » : il s'agit d'une description structurée indiquant les champs de métadonnées utilisables par un standard ou dans un contexte donné.

Cartographie générale

Nous pouvons mettre à profit l'expérience d'Europeana en matière de métadonnées et d'interopérabilité. Europeana propose un ensemble de balises qui sont classées selon leur degré de nécessité. Mais là où la plus grande part des dispositifs d'interprétation des

métadonnées propose 3 catégories, Europeana différencie les éléments absolument essentiels des éléments « simplement » nécessaires. Les champs peuvent également être simplement recommandés. La dernière catégorie regroupe les champs accessoires[72].

Si on adopte un principe de hiérarchie similaire pour les métadonnées de l'édition électronique ouverte, on doit distinguer les articles de revue des chapitres de livre. D'autres documents spécifiques ont pu être identifiés, mais il serait nécessaire de raffiner l'analyse pour chacun d'eux.

Dans l'ensemble, on peut observer une insuffisante catégorisation des différentes mentions de responsabilité dans les formats à balises courants et surtout dans les modalités d'utilisation de ces formats (certains formats autorisant des précisions satisfaisantes sont sous-utilisées). La problématique de l'autorité personnelle n'est certainement pas la même en sciences humaines et sociales qu'en mathématiques ou en biochimie : d'une part la responsabilité est souvent concentrée dans un nombre réduit de personnes (on peut encore une fois citer le travail plus volontiers solitaire des chercheurs en sciences humaines et sociales identifié par Jenny Fry[73]). D'autre part les responsabilités annexes se font selon des modalités très particulières, plus spécifiques que la simple « participation », courante en sciences expérimentales : il est rare que le produit du travail d'un traducteur soit une des sources principales utilisée par un chercheur en mathématiques.

Si nous listons les principales balises rencontrées dans l'usage, en précisant leur caractère général de nécessité (sans tenir compte des types de document cette fois ci), voici la cartographie générale des métadonnées qui peut s'appliquer à la plupart des documents de sciences humaines et sociales :

Champ	Valeur du champ	Nécessité d'usage
clé	Identifiant interne dans la base de données	nécessaire
création	Date de création des métadonnées	recommandé
indexeur	Auteur des métadonnées	recommandé
type	Type de document	nécessaire
genre	Type d'article (commentaire, thèse, édito…) Type de chapitre (dédicace, corps, conclusion…)	important important
titre	Titre de l'article Sous-titre de l'article Titre du chapitre Sous-titre du chapitre	nécessaire nécessaire
auteur	Nom, Prénom de l'auteur	nécessaire
traducteur	Nom, Prénom du traducteur	important
directeur	Nom, Prénom du directeur de publication	important
éditeur de la série	Nom, Prénom du directeur de collection	recommandé
auteur critique	Nom, Prénom du commentateur (différent d'auteur de la critique = auteur)	important
contributeur	Nom, Prénom du collaborateur	accessoire
collaborateur	Nom, Prénom du collaborateur	accessoire
résumé	résumés en différentes langues	important

[72] (Verleyen 2009)
[73] (Fry 2009)

Champ	Valeur du champ	Nécessité d'usage
publication	Nom de la revue	nécessaire
titre du livre	Titre du livre	nécessaire
série	Nom de la collection	recommandé
volume	Nombre d'années ou année de publication (chaque année étant subdivisée en numéros, numérotation recommencée chaque année) Numéro de volume du livre dans la collection	accessoire accessoire
numéro	Numéro de la revue (à préférer) ou Titre du numéro de la revue Numéro du chapitre	accessoire accessoire
nbr de volumes	Nombre de volumes dans la collection	accessoire
édition	Numéro d'édition	important
première édition	Année de première édition	important
pages	Pagination	recommandé
jour	Jour de publication	important
mois	Mois de publication	important
date	année de publication	nécessaire
éditeur	Editeur (titulaire des droits d'exploitation)	nécessaire
institution	Institution à l'origine de la création	important
lieu	Lieu de publication	accessoire
abréviation de journal	Titre abrégé de la revue	recommandé (même si identique au titre long)
titre abrégé du livre	Titre abrégé du livre	recommandé (même si identique au titre long)
langue	nom de la langue	nécessaire
DOI	DOI	important
ISBN	Numéro ISBN du livre	nécessaire
ISSN	Numéro ISSN de la revue Numéro ISSN de la collection	nécessaire recommandé
section	Partie contenant le document	accessoire
titre abrégé	Titre abrégé de l'article	accessoire
élément commenté	Identifiant unique du document commenté	important
URL	URL de l'article URL du chapitre	nécessaire
accédé le	Date de consultation : jour mois année heure:minute:seconde	nécessaire
dépôt	Droits d'auteur	nécessaire
nom de la base de données	Nom de la base de données	accessoire
autorisation	Licence d'utilisation	nécessaire
disponibilité	Date d'entrée en libre accès	important
prix	Prix	important
couverture spatiale	zone géographique traitée par le document	important
couverture temporelle	période historique traitée par le document	important

Champ	Valeur du champ	Nécessité d'usage
marqueurs	mot-clef 1 mot-clef 2 keyword 1 keyword 2	accessoire (standards actuels)
notes attachées	contenu de la note 1 contenu de la note 2 …	accessoire
pièces jointes	Titre du lien 1 Titre du lien 2	important
autres versions	URL des autres versions et des autres formats disponibles	important

L'ordre utilisé est le suivant, du moins nécessaire au plus nécessaire :

- accessoire
- recommandé
- important
- nécessaire

Un principe applicable serait de dire que les champs considérés comme « nécessaires » le sont indépendamment des circonstances, bloquant même éventuellement le processus de publication en cas d'absence, tandis que les champs classés comme « importants » doivent être renseignés dès qu'ils sont disponibles.

Transformations d'un format à l'autre

Les traductions d'une grammaire à l'autre engendrent fréquemment un nivellement des qualités des métadonnées créées, du fait des transpositions souvent imparfaites d'un « set » de métadonnées à l'autre. C'est pourquoi la plus grande attention doit être portée aux sens et emplois des différents champs de métadonnées dans chaque format.

Dans l'ensemble des équivalences possibles identifiées pour les articles de revue et pour les chapitres de livres, on peut pointer quelques éléments particuliers :

- Ce qu'on pourrait appeler le format Zotero s'avère souvent trop pauvre ou trop peu structurant sur certains points, et presque trop riche sur certains on est donc amené à le compléter en puisant dans d'autres formats existants. Les types de mention de responsabilité sont plus nombreux que dans d'autres formats, mais ces types d'auteur / de créateur ne sont pas toujours adaptés au genre de document concerné. A l'inverse les champs destinés à exprimer l'époque ou la zone géographique sont insuffisants dans Zotero.
- Les standards qui mettent en place des fonctionnements conditionnels (si ce champ a telle valeur, alors ce champ est nécessaire, etc.) sont souvent bien adaptés aux documents pour lesquels ils sont conçus. Mais ils sont aussi incomparablement plus complexes à cartographier, comme on peut le constater quand on tente de convertir un fichier Endnote.
- Lorsque deux sens différents sont concurrents pour un champ de métadonnées, on a toujours intérêt à en privilégier un par rapport à l'autre et à expliquer ce choix, en fonction des pratiques dominantes par exemple.

- L’une des précautions à prendre est de s’assurer que les informations qu’on consigne dans les métadonnées pourront avoir un rôle à jouer jusqu’à l’utilisateur final, surtout pour l’utilisateur final. Mais en contrepartie on se doit de toujours évaluer le coût qu’implique la prise en compte d’une catégorie d’information supplémentaire, voire d’un standard d’organisation ou d’expression supplémentaire.

SOURCES ET OUTILS DISPONIBLES POUR SOUTENIR LES EFFORTS D'AMELIORATION DE LA QUALITE DES METADONNEES

Il est de plus en plus difficile, dans notre contexte technologique et social, d'envisager une action isolée pour un éditeur scientifique désireux d'augmenter la qualité de ses métadonnées. C'est au contraire un avantage que de pouvoir s'appuyer sur les expériences et les résultats de différentes communautés, de différents groupements d'intérêt commun (en anglais, langue la plus couramment utilisée sur Internet, on parle de « special interest group », SIG). Confrontés à une multitude de documents et d'outils ayant un rapport avec les métadonnées, nous ne pouvons pas encore actuellement disposer du recul nécessaire pour faire le tour de toutes ces aides potentielles.

Recommandations sur l’environnement logiciel

Les tendances sur les réseaux changent très vite, c'est une évidence lorsqu'on pense aux « modes » qui affectent les Content Manager Setup (ou CMS). L'impact des outils sur les usages en matière de métadonnées ne doit pas être oublié dans l'appréciation de leur intérêt. Ainsi lorsqu'une plate-forme de blogs aussi dominante que WordPress fait des choix sur les possibilités de description des billets ou sur des services nouveaux, c'est tout le secteur de l'édition de blogs scientifiques qui doit en tenir compte, pour l'imiter ou s'en différencier. La généralisation du RDF et des ontologies dans la publication scientifique, de la même façon, est un peu suspendue à l'adoption de ces standards par des outils de rédaction et de publication.

Il serait intéressant et utile d'étudier l'évolution de l'usage de la plate-forme de citation collaborative CiteULike, dans la mesure où les sciences humaines et sociales se prêtent plus aisément à l'apport incident d'un public élargi au-delà des limites de la recherche institutionnelle. De plus CiteULike s'interface avec plus de facilité, si on en croit les propos tenu à l'adresse des éditeurs, avec la littérature scientifique pourvue d'identifiants pérennes (voire de DOI) et des fils de déclarations en PAM : autant d'évolutions envisagées pour les mois à venir par Revues.org.

Concernant l'opportunité de proposer des fonctionnements propres au web sémantique, la situation de l'édition scientifique en sciences humaines et sociales est semblable au secteur de l'information généraliste ou spécialisée. Il est presque inévitable que les besoins des usagers rendent nécessaire la mise en place de ce que Tim Berners-Lee

nomme « le web de données ». Il est toutefois pour l'instant à la fois risqué et coûteux de prendre en charge ce développement.

Les expériences de dimensions réduites (on peut penser à ce que proposait Gautier Poupeau dès 2007 avec sa « knowledge box », sur son blog *Les Petites Cases*[74]) et les grands projets largement collaboratifs (comme DBpedia[75]) semblent pour l'instant les meilleures échelles à partir desquelles se développement le « web 3.0 ». L'édition électronique ouverte n'a pas actuellement les marges financières et les ressources humaines suffisantes pour s'atteler à une tâche aussi immense qu'inéluctable. La réutilisabilité se limite encore, mais ce sont des pas importants, à la création des identifiants pérennes (tel le DOI de la fondation CrossRef), à l'exploitation croissante de vocabulaires contrôlés et à l'utilisation de standards de métadonnées « ouverts » sur les fonctionnements par triplets.

Recommandations sur les échanges avec d'autres acteurs

L'adhésion au consortium TEI, en projet à très court terme, serait en effet un excellent choix dans la mesure où il donnerait à la plate-forme adhérente les moyens de peser sur les développements autour de ce standard d'expression très adapté aux sciences humaines et sociales. Le MODS, envisagé un temps mais laissé de côté, pourrait par exemple être réintégré dans les formats disponibles grâce au travail de conversion entre TEI et MODS de l'université de l'Indiana[76]. L'adéquation entre METS et MODS ayant été particulièrement travaillée, on voit les garanties fonctionnelles qu'une plate-forme de publication peut tirer de ces échanges entre standards « pivots ».

Dans un même ordre d'idée, l'IDPF[77] qui développe le format EPUB se retrouve régulièrement animé par des arbitrages qui auront une implication dans des évolutions futures. Ainsi est régulièrement débattue la question des processus de mise en page adaptative, qui fonctionnent soit avec des feuilles de style en cascade (CSS) soit avec des feuilles de transformation de forme (XSL-FO). Mais ces discussions, qui peuvent avoir un impact fort sur les métiers de l'édition électronique ouverte, peuvent être infléchies si justement les éditeurs prenaient une part active en apportant leur point de vue et leur expérience.

Devant l'inquiétude de certains, dont Marin Dacos, sur la position dominante dont pourrait abuser le DOI, la seule alternative crédible serait un réseau de partenaires faisant des références croisées *via* les standards légers de type RSS 1.0. La spécificité de l'édition électronique ouverte est qu'elle se place souvent en partenariat avec des entités similaires pour les développements techniques, mais entre en « concurrence » avec celles-ci du point de vue de l'économie de l'attention. Sans doute, afin de contourner cet obstacle, serait-il préférable de nouer des partenariats avec des plates-formes du même domaine scientifique mais appartenant à d'autres aires culturelles (Scielo, au Brésil,

[74] Voir son article à l'adresse : http://www.lespetitescases.net/amusons-nous-avec-rdfa
[75] Dont le projet est exposé à l'adresse : http://wiki.dbpedia.org/ .
[76] Cf. https://wiki.dlib.indiana.edu/confluence/display/ETDC/TEI+to+MODS+Mapping+Issues+and+Challenges
[77] http://www.idpf.org/

pourrait en faire partie[78]), ou avec des plates-formes de la même aire culturelle mais traitant d'un domaine scientifique différent (on pense par exemple au CEDRAM[79]).

Les acteurs de l'édition électronique ouverte, conscients du rôle des métadonnées dans la diffusion de la recherche en sciences humaines et sociales, sont pour l'instant contraints de s'adapter aux mouvements impulsés par d'autres (grands éditeurs commerciaux, consortia et fondations tels que le W3C, CrossRef ou la fondation Mellon, acteurs économiques majeurs des télécommunications, juridictions nationales et internationales). Leur capacité d'innovation est réduite malgré l'implication et la compétence des personnels qui y oeuvrent. Peut-être la donne serait-elle changée, pour la recherche française et européenne, si était mieux comprise l'importance de l'avantage comparatif que procure une analyse prospective continue sur les métadonnées et sur les flux d'information. Que l'univers des bibliothèques joue pleinement son rôle dans l'économie générale des standards de métadonnées serait donc un pas important dans ce sens.

[78] Scientific Electronic Library Online, http://www.scielo.org/php/index.php
[79] http://www.cedram.org/

Rôle des bibliothèques dans l'amélioration de la qualité des métadonnées

Comme l'a fort bien montré Emma Bester[80], dans une économie de l'attention, les choix techniques et les héritages de pratiques des bibliothèques universitaires ont un impact sur l'audience des publications en sciences humaines et sociales. Vis-à-vis de l'édition électronique ouverte, cet impact n'est pas forcément positif, pour deux raisons principales distinctes :

- les bibliothèques, par manque d'expertise, ont dû se reposer sur un ensemble d'intermédiaires de gestion des ressources électroniques qui favorisaient les gros éditeurs traditionnels et les entreprises puissantes (parmi d'autres, Thomson, Springer, Elsevier, Nature Publishing Group, etc.);
- les bibliothèques de recherche ont directement bénéficié (et se sont contenté) de la voie verte du libre accès, celle des archives ouvertes. Pierre Mounier parle de façon très explicite d'un « Yalta du libre accès » issu des conclusions de la conférence de Budapest sur le libre accès, où les bibliothèques ont trouvé leur place dans le fonctionnement des archives ouvertes tandis que les éditeurs avaient toute la leur dans les revues en libre accès. Les deux voies ont dès lors pris le risque de ne pas laisser de place à « l'autre », l'éditeur ouvert dans la voie verte et la bibliothèque dans la voie dorée[81].

Il semblerait donc bénéfique qu'un dialogue renouvelé s'instaure entre bibliothèques et éditeurs de revues et monographies en libre accès, d'autant que les objectifs à long terme des uns et des autres se rejoignent. Le rôle des bibliothèques dans la qualité générale des métadonnées, leur rôle plus spécifique vis-à-vis des standards de métadonnées méritent donc d'être examinés, à la suite de l'analyse que nous avons menée dans le domaine de l'édition électronique en sciences humaines.

Dans cette analyse du rôle des bibliothèques dans la qualité des métadonnées, nous nous concentrerons sur le cas des bibliothèques universitaires, des bibliothèques de grands établissements scientifiques, des bibliothèques patrimoniales et des bibliothèques nationales. On gardera toutefois à l'esprit que les grandes campagnes de numérisation des documents concernent aussi des bibliothèques municipales, et que celles-ci ont donc également à ce titre un rôle à jouer dans les métadonnées des réservoirs de documents numérisés que sont par exemple Google Book Search[82] ou HathiTrust[83].

Quatre rôles différents peuvent être joués par les bibliothèques dans le cycle de vie des métadonnées :

- les bibliothèques utilisent les métadonnées créées par d'autres
- les bibliothèques créent des métadonnées
- les bibliothèques valident le contenu de certaines métadonnées

[80] (Bester 2009)
[81] (Bester et Mounier 2009)
[82] http://books.google.com/books
[83] http://www.hathitrust.org/

- les bibliothèques, en accompagnant l'utilisation ou la création des métadonnées par des tiers, influencent les usages et la qualité de ces métadonnées

ROLE D'UTILISATRICES DES METADONNEES

Les principaux schémas de métadonnées utilisés en bibliothèque sont des standards de description bibliographique ou à vocation bibliographique. Mais ce ne sont pas les seuls, ainsi qu'on le verra.

Les bibliothèques voient peu à peu se réduire la part de travail consacrée au catalogage (ce qui n'est pas sans poser des questions sur le rôle et la formation du personnel, questions que nous n'aborderons toutefois pas ici). La principale raison de ce phénomène est la possibilité d'accéder à des sources de plus en plus complètes de notices bibliographiques, qu'il est plus aisé de récupérer à distance que de créer soi-même – une opération reste néanmoins présente, et cette opération est essentielle, c'est l'exemplarisation, qui est de fait l'attribution d'un identifiant unique –.

On le sait pourtant, une notice de document est toujours plus précise lorsqu'elle est faite avec soin en fonction de l'univers dans lequel ce document sera utilisé, en particulier sur le plan de l'indexation du contenu. Importer le code Dewey d'un ouvrage est ainsi souvent source d'erreurs de classement. Pourquoi ne pas opter systématiquement pour la solution offrant la meilleure qualité ? Dans un volume de milliers de documents entrant dans les fonds de la bibliothèque, il est impossible de tenir ce niveau d'exigence à des coûts temporels et financiers raisonnables. Il est d'autant plus logique de profiter du travail bibliographique des autres professionnels de la documentation que la qualité des métadonnées créées est directement influencée par la maîtrise que l'agent peut avoir du document; or on peut aisément, pour un novice, confondre quark p et quark q dans la création d'une notice traitant de physique des hautes énergies, alors qu'un spécialiste distinguera nettement les deux particules.

Après une forte diversification dans les années 70 – 80 liée aux multiples adaptations locales des formats MARC, on assiste de ce fait à une uniformisation progressive des métadonnées utilisées par les bibliothèques. Les SIGB, en nombre de plus en plus restreints mais de plus en plus diffusés, exploitent les formats dominants fondés sur le XML. Ces logiciels de gestion de bibliothèques automatisent les transferts de notices en proposant plusieurs protocoles qui sont autant de standards d'organisation des métadonnées : si le Z3950 garde un rôle fondamental pour les publications discrètes, il sera sans doute à court terme supplanté par les requêtes OAI et éventuellement le PAM. De toutes façons pour les publications continues, la livraison de lots de notices en fonction des bouquets de revues reçus est devenue courante et dominante.

Par ailleurs ce qu'on peut appeler rétrospectivement les premiers identifiants pérennes (ISBN, ISSN) ont été valorisés et promus par l'utilisation quotidienne qu'en faisaient les bibliothèques. Si les bibliothèques, comme elles semblent devoir le faire, valorisent le DOI en préférant les documents numériques qui en comportent un, ce standard d'identification pérenne s'imposera d'autant plus à l'édition électronique (ouverte ou non).

ROLE DE CREATRICES DE METADONNEES

Les bibliothèques, en particulier lorsqu'elles sont héritières d'une longue histoire, renferment des trésors, pour les sciences humaines et sociales comme pour les autres champs scientifiques. Or les documents non indexés, ou répertoriés sur catalogue papier ou fiches cartonnées (hors de tout processus de diffusion élargie), ne sont pas visibles pour la plus grande part de la communauté scientifique. C'est pourquoi il est important d'encourager les projets visant à améliorer la diffusion de ces données « dormantes ». Il est tout aussi essentiel d'intégrer dans ces projets la prise en compte des usages de l'édition scientifique et de la circulation des données sur les différents réseaux. Lorsqu'une bibliothèque diffuse le signalement de son patrimoine, elle crée consciemment ou non des métadonnées sur ces documents patrimoniaux. Le choix des standards employés s'est souvent fait à partir des usages en cours, sans dimension prospective suffisante. Mais quoi qu'il en soit, il est indéniable que les bibliothèques font un travail quotidien de création de métadonnées : celles-ci ont souvent les défauts de leurs qualités, c'est à dire que, admirablement adaptées au contexte de leur création elles sont difficilement utilisables dans d'autres contextes, d'autres usages, d'autres formats. Du point de vue de l'interopérabilité, la volonté de bien faire des bibliothécaires est finalement parfois néfaste : on l'a vu plus haut, la qualité qui ne peut pas être maintenue au fil des échanges est doublement coûteuse, surtout si elle n'est pas conjuguée avec une interprétation du standard accessible aussi bien aux machines qu'aux individus.

La dimension archivistique des fonds patrimoniaux, et surtout la difficulté de traiter d'emblée tout un fonds, ont poussé certaines bibliothèques à utiliser le format EAD (qui s'accommode très bien de granularités de description différentes) pour les décrire. L'avantage de l'EAD est d'être une organisation et une grammaire issues du XML, ce qui facilite grandement les conversions. Les bibliothèques l'utilisent donc, même si son utilisation est désormais plutôt le fait du monde des archives. L'édition ne l'a pas intégré dans son fonctionnement mais à terme l'édition électronique ouverte aurait tort de délaisser ce format, par lequel des documents pour la recherche et des publications scientifiques sont mis à disposition. Il serait important, dans cette optique, que le schéma EAD puisse être interprété par les serveurs des éditeurs de revues pour d'éventuelles citations ou inclusions à titre d'exemple (par le biais des espaces de noms, dans une enveloppe en METS par exemple, cela devrait se faire sans difficulté insurmontable).

De façon moins spécifique que les documents patrimoniaux, pour certaines publications échappant au circuit traditionnel de l'édition, les bibliothèques ont créé et créent aussi des notices descriptives, donc utilisent des standards de métadonnées. La spécificité de certains formats génère là encore son lot de problèmes : convertir un fichier Filemaker ne se fait pas aussi facilement, dans le contexte actuel, qu'une notice en UNIMARC.

Enfin la création de standards de métadonnées n'aurait pas connu la même histoire sans le travail de quelques bibliothèques, au premier rang desquelles la bibliothèque du Congrès. Elles ont été parmi les premières institutions à avoir eu besoin de ce qu'on ne nommait pas alors standards de métadonnées, et qui ont alimenté les premières bases, à

but interne tout d'abord. Ainsi même si la masse gigantesque de documents que les bibliothèques avaient à gérer a été dépassée par la masse incommensurable des documents disponibles sur les réseaux informatisés, les métadonnées créées par les bibliothèques forment encore une part importante de toutes celles disponibles sur Internet.

ROLE DE REFERENCE STABLE DES DONNEES

Les bibliothèques, dans le domaine de l'identification des documents, ont depuis des siècles acquis une certaine légitimité. Cette identification, cette description se sont fait avec une perspective à très long terme. C'est le cas des bibliothèques universitaires, acquérant et conservant les produits de la recherche pour les transmettre aux utilisateurs présents et à venir. C'est le cas des bibliothèques nationales, souvent titulaires du privilège de dépôt légal et nanties du rôle de recensement des publications parues sur le territoire de leur juridiction. On peut noter à ce propos que les besoins des grandes bibliothèques en matière de conservation ont été décisifs dans l'émergence de standards de métadonnées spécifiquement pensés dans ce but.

La publication sur Internet a nécessairement modifié les conditions matérielles de l'exercice de ce rôle d'identification. Néanmoins l'expérience tirée de quelques mois d'existence de Google Books montre que la dématérialisation des documents, paradoxalement, rend plus nécessaire le recours au témoignage des fonds et des connaissances accumulés dans les bibliothèques. Là où une recherche sur Internet peut faire croire que Voltaire a écrit un ouvrage en 1905, ou qu'un Dictionnaire des peuples d'Amérique du Nord a pour sujet les prénoms d'enfants, les fonds des bibliothèques permettent de s'appuyer sur un travail continu de plusieurs siècles (et une utilisation correspondante) pour valider les informations.

C'est donc dans un autre domaine que les bibliothèques peuvent prétendre à une légitimité certaine, celui des autorités. L'habitude d'avoir recours à des thésaurus, à des bibliographies, à ce qu'on appelait il y a peu des documents secondaires, pour guider la recherche des usagers de bibliothèque a certainement participé à l'acquisition pour ce secteur de la galaxie Gutenberg de cette compétence particulière. Il n'est donc pas surprenant de voir les grandes bibliothèques conserver presque intact ce rôle de désignation des autorités (VIAF est par exemple l'agrégat des autorités des bibliothèques nationales de plusieurs pays).

ROLE DANS LA PRESCRIPTION D'USAGES

Auprès de certains utilisateurs de métadonnées

Les bibliothèques universitaires sont particulièrement concernées par l'encadrement des personnes utilisant les métadonnées. Bien avant qu'une exigence soit formulée dans la formation des étudiants aux outils de la recherche documentaire, les enseignants-chercheurs et les bibliothécaires ont souvent échangé sur les outils technologiques qui étaient à disposition de la recherche. Parmi ces outils, on comptait les catalogues des bibliothèques, devenus OPAC sans qu'ils progressent nettement en clarté de fonctionnement pour des non-spécialistes. La conception des « champs » de recherche induisait une analyse opératoire des métadonnées. De ce fait les exigences des chercheurs en matière de métadonnées se limitaient la plupart du temps à quelques éléments, ceux rencontrés couramment dans la page d'accueil du catalogue informatisé.

Ces utilisateurs des métadonnées ont tiré profit, à partir des années 80, des logiciels de gestion bibliographique, tel Endnote. Certains enseignants chercheurs ont alors construit leurs méthodes bibliographiques au contact des bibliothécaires, qui ont donc là encore joué un rôle de prescripteurs d'usage en matière de métadonnées.

Actuellement, ce sont désormais aussi les étudiants (certes souvent de niveau licence – master) qui sont formés à l'appréhension de métadonnées dans le cadre de l'apprentissage de la recherche documentaire. Il ne s'agit pas de cours sur la DTD EAD, mais les outils tels que Zotero sont parfois évoqués. Cette transition de l'outil bureautique de gestion de métadonnées à l'outil tourné vers les usages actuels d'Internet est significative mais n'efface pas le rôle que les bibliothèques peuvent jouer auprès des chercheurs et apprentis chercheurs dans l'utilisation des gisements de métadonnées.

Auprès de certains créateurs de métadonnées

Les bibliothèques ont une place à part dans l'univers mental de nombreux intervenants en sciences de l'information, dans celui des éditeurs comme dans celui des ingénieurs qui développent des outils de création de métadonnées. Et ce n'est pas sans effet sur leur travail dans le domaine des métadonnées. Trois exemples, parmi d'autres, tendent en tout cas à le prouver.

Pour reprendre le cas évoqué plus haut des enseignants chercheurs, on peut souligner que lorsqu'ils écrivent un article et entrent dans un processus de publication, ils ont parfois leur mot à dire dans la qualification de leur document : les mots-clés par exemple. A leur mesure, pour leurs propres recherches, ils construisent également des bibliothèques virtuelles qui sont autant d'exercices de création de métadonnées.

Le rôle de MARC dans l'univers des données sur les données n'est plus le même qu'il y a 20 ans; toutefois, pour des raisons historiques et parce que les bibliothèques abritent elles aussi des pionniers dans la recherche sur les sciences de l'information, ces institutions gardent une certaine audience pour leurs conseils sur la gestion des données

bibliographiques. L'exigence de précision dans la rédaction des mentions de responsabilité, l'idée de prendre en compte les différentes versions d'un document, ou la prospective sur la description du contenu sémantique rencontrent directement des problématiques au long cours du monde des bibliothèques.

Comme l'ont noté par ailleurs Ann Apps et Ross MacIntyre, les bibliothèques ont pesé pour l'adoption par les éditeurs des standards OpenURL, dans la mesure où elles étaient systématiquement à la recherche de moyens de signaler les ressources qu'elles proposaient localement[84]. C'est un témoignage supplémentaire attestant du rôle que jouent et pourraient jouer, consciemment ou non, les réseaux de bibliothèques dans l'évolution des métadonnées. La proximité entre le monde de la recherche et celui des bibliothèques, sans doute plus grande en sciences humaines et sociales, nous invite à penser avec optimisme à l'opportunité de nouveaux échanges dans ce champ disciplinaire.

[84] (Apps et MacIntyre 2006)

Conclusion

Le secteur des Sciences humaines et sociales pousse à la multiplication des formats de métadonnées, car les contenus de ce pan de la recherche sont comme des poupées russes, ou comme les motifs des figures fractales, toujours susceptibles de contenir des unités plus petites ou plus grandes. A l'inverse, le secteur de l'édition pousse à une rationalisation dans un but de meilleure visibilité.

La combinaison des deux n'en est que plus fructueuse du point de vue de la recherche sur les métadonnées. Preuve en est, d'un monde des métadonnées d'identification, nous passons peu à peu à un monde des métadonnées support d'agrégation : or l'agrégat n'existe que dans un sens, toujours changeant, toujours à re-contextualiser, comme l'est la connaissance en sciences humaines et sociales.

L'édition électronique ouverte nous montre donc que la qualité des métadonnées ne saurait être purement technique. On comprend également que l'attention à la rigueur des formats ne peut éviter ni la réflexion à haut niveau (conceptions nouvelles) ni les précautions d'application (gestion des erreurs de description ou de tagging). Et enfin que la qualité dépend autant de l'adaptation à la situation qu'à l'adaptation à une évolution dans le temps.

François Moreau, dans « Méta-information et économie numérique » (Moreau 2008), souligne « un déplacement de la valeur vers la méta-information » (p. 231). Les bibliothèques auraient donc intérêt à jouer pleinement leur rôle dans la création de cette valeur; et puisque par essence la méta-information est destinée à faire des liens autour des informations, ce serait l'occasion de tisser des liens nouveaux avec les éditeurs, en particulier avec ceux dont l'objectif est d'offrir au plus grand nombre un accès libre aux connaissances dans toute leur diversité.

Bibliographie

Adobe. 17 Septembre 2009. *InDesign to Kindle white paper* | Adobe Developer Connection. http://www.adobe.com/devnet/digitalpublishing/articles/indesigntokindle.html.

Apps, Ann et MacIntyre, Ross. 2006. "Why OpenURL?" in *D-Lib Magazine* 12, n° 5 (Mai). doi:10.1045/may2006-apps. http://www.dlib.org/dlib/may06/apps/05apps.html.

Beaubien, Rick. 2007. *METS : an introduction*. Présenté à l'*University of California Computing Services Conference*, Berkeley. http://www.loc.gov/standards/mets/presentations/METSUCCSC.ppt.

Beaudry, Guylaine. "Le numérique et les mutations dans la structuration du champ éditorial de l'ouvrage en sciences humaines et sociales" in *Mémoires du livre* 1, no. 1. http://id.erudit.org/iderudit/038633ar.

Bester, Emma. 9 novembre 2009. *L'Economie de l'attention pour le Libre Accès. Le cas de Revues.org dans les bibliothèques universitaires.* Diplôme de chef de projet en ingénierie documentaire, INTD.

Bester, Emma et Mounier, Pierre. Novembre 2009. "Usages des ressources électroniques en libre accès dans les BU et SCD". Conférence présentée à *Ressources électroniques académiques*, Lille.

Bilder, Geoffrey, Dodds, Leigh, Hammond, Tony, O'Beirne, Richard et J Rogers, Lisa. 19 octobre 2009. *CrossTech: Recommendations on RSS Feeds for Scholarly Publishers.* Éd. Crossref et TicTOC. Oxford university press. http://oxford.crossref.org/best_practice/rss/.

Chumbe, Santiago et Macleod, Roderick. 4 juin 2009. "TicTOCron: an Automatic Solution for Propagating Quality Metadata to Scholarly TOC RSS Feed Metadata". Conférence présentée à la *30ème conférence annuelle de l'IATUL*, Leuven (Belgique). http://eprints.rclis.org/15946/.

Cover, Robin. 20 septembre 2001. *Open Content Syndication (OCS).* http://xml.coverpages.org/ocs.html.

Crossref. 31 novembre 2009. *CrossRef Fast Facts.* http://www.crossref.org/01company/16fastfacts.html.

Dacos, Marin. 21 Septembre 2009. "La citabilité des textes en ligne" in *L'édition électronique ouverte*. http://leo.hypotheses.org/2597.

Dalbin, Sylvie. 2008a. 2 Octobre 2008. "Représentation et accès à l'information : transformation à l'œuvre" in *Métadonnées : mutations et perspectives. Séminaire INRIA, 29 septembre-3 octobre 2008, Dijon*, éd. Calderan, Lisette, Hidoine, Bernard et Millet, Jacques. Pages 9-54. Paris: ADBS.

———. 2008b. 2 Octobre 2008. "Métadonnées et normalisation" in *Métadonnées : mutations et perspectives. Séminaire INRIA, 29 septembre-3 octobre 2008, Dijon*, éd. Calderan, Lisette, Hidoine, Bernard et Millet, Jacques. Pages 113-157. Paris: ADBS.

DCMI. 20 décembre 2004. *Dublin Core Metadata Element Set, Version 1.1: Reference Description*. http://dublincore.org/documents/dces/.

Doueihi, Milad. 17 janvier 2008. *La Grande Conversion numérique*. Seuil.

Equipe des rédacteurs de Calenda. 2009. "Les Digital humanities aujourd'hui : centres, réseaux, pratiques et enjeux". http://calenda.revues.org/nouvelle15009.html.

Fry, Jenny. 2 juin 2009. "The disciplinary shaping of information landscapes" présenté à *Diversité des pratiques numériques*, ENSSIB, Villeurbanne. http://pratiquesnum.enssib.fr/PPT/Fry.pdf.

Gèze François, « Le livre dématérialisé », présenté à *Horizon 2019 - Bibliothèques en prospective*, ENSSIB, Villeurbanne.

Gonzalez-Quijano, Yves. 20 décembre 2009. "No (Coranic) Logo !" in *Culture et politique arabes*. http://cpa.hypotheses.org/1519.

Hammond, Tony. 8 mai 2009. "CrossTech: PRISM Aggregator Message" in *CrossTech*. http://www.crossref.org/CrossTech/2009/05/post_2.html.

Kai Eckert, Pfeffer, Magnus et Stuckenschmidt, Heiner. 2009. "A Unified Approach for Representing Metametadata" in *International Conference on Dublin Core and Metadata Applications; DC-2009 Seoul Proceedings*. http://dcpapers.dublincore.org/ojs/pubs/article/view/973/948.

Luther, Judy. 30 juin 2009. *Streamlining Book Metadata Workflow*. Vol. 1. Dublin, Ohio: NISO & OCLC. http://www.oclc.org/fr/fr/news/releases/200940.htm.

Martin, Frédéric. 26 avril 2007. "La mise en oeuvre de l'OAI-PMH à la BnF". Conférence présentée aux *Journées professionnelles du groupe français de l'AIBM*, Périgueux. http://www.aibm-france.org/journees_pro/perigeux_2007/jp07_compte_rendu_2.htm.

Martin, James. 1982. *Strategic Data Planning Methodologies*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Moreau, François. 2 octobre 2008. Méta-information et économie numérique. Dans *Métadonnées : mutations et perspectives. Séminaire INRIA, 29 septembre-3 octobre 2008, Dijon*, éd. Calderan, Lisette, Hidoine, Bernard et Millet, Jacques.

Pages 223-238. Paris: ADBS.

Morel-Pair, Catherine. 2007. "Métadonnées et XML : des standards efficients de l'environnement numérique" in *Ingénierie des systèmes d'information* 12, n° 2. Pages 9-39. Prépublication disponible à l'adresse http://www.enssib.fr/bibliotheque-numerique/notice-1842.

Morlock, Emmanuelle, et Raphaël Tournoy. 10 septembre 2009. "Formats et enjeux des métadonnées en édition - Edition électronique en sciences humaines et sociales". Atelier présenté à l'*Université d'été du CLEO*, Marseille. http://www.digitalhumanities.cnrs.fr/wikis/edelec-shs/index.php/Formats_et_enjeux_des_m%C3%A9tadonn%C3%A9es_en_%C3%A9dition.

Mounier, Pierre. 11 décembre 2009. "Revues.org : Fédération de revues de sciences humaines et sociales en libre accès" in Journée d'étude de l'*Institut des Amériques*, Paris. http://www.slideshare.net/revuesorg/prsentation-de-revuesorg.

Nunberg, Geoff. "Google Books: A Metadata Train Wreck" in *Language Log*. http://languagelog.ldc.upenn.edu/nll/?p=1701.

———. 28 août 2009. "Goog Book MetadataSh" présenté au *Google Book Settlement Conference*, UC Berkeley. http://people.ischool.berkeley.edu/~nunberg/GBook/GoogBookMetadataSh.pdf.

Roszkiewcz, Ron. 2008. *XMP Primer*. http://www.idealliance.org/filefolder/XMPPrimer.pdf.

Schaffner, Jennifer. 2009. *The Metadata is the Interface*. Rapport de recherche. OCLC Research. http://www.oclc.org/programs/publications/reports/2009-06.pdf.

Serres, Alexandre et CLEO. 15 décembre 2009. "Blogosphère scientifique en SHS" présenté à la Formation URFIST de Rennes, Rennes. http://www.slideshare.net/revuesorg/le-blogosphre-en-sciences-humaines-et-sociales-prsentation-de-la-plateforme-hypothses.

Van de Sompel, Herbert, Hochstenback, Patrick et Beit-Arie, Orie. 2 novembre 2001. *OpenURL syntax description*. ExLibris.

Verleyen, Julie. 19 juin 2009. "L'ouverture et l'interopérabilité - Europeana contributions" présenté à la formation *Numérisation et numérique en bibliothèque*, ENSSIB.

W3C. 11 février 2004. *RDF Primer*. http://www.w3.org/TR/REC-rdf-syntax/.

Wenz, Romain. Janvier 2009. *Avenir des catalogues (L') : Formats, données, outils, usages*. ENSSIB. http://www.enssib.fr/bibliotheque-numerique/notice-21205.

Table des annexes

Glossaires

GLOSSAIRE DES ABREVIATIONS

API = Application Programming Interface (interface de programmation d'application, ou plus simplement interface de programmation). Ensemble de briques mises à disposition pour le lancement de procédures à distance sur des données.

COinS = Context Object in Span (Objet contextuel décrit dans une balise « span »). Ainsi que la traduction en français l'indique, le COinS est un ensemble de métadonnées insérées dans une balise « span » de HTML, donc non visibles par l'utilisateur mais précisant pour les automates, à un endroit précis du texte enrichi de la sorte, des informations bibliographiques. La déclaration est traduite par un résolveur de lien selon une grammaire définie par la norme Z3988 du NISO.

DTD = Document Type Definition. Fichier de définition de grammaire utilisée dans un fichier XML. Il associe les entités possibles aux propriétés qui peuvent y être attachées. Les DTD sont exprimées en SGML (à la différence des fichiers .xsd - XML Schema Definition -, qui sont en XML même s'ils remplissent à peu près la même fonction).

EAD = Encoded Archival Description. L'EAD est une DTD. Format de description spécifiquement conçu pour l'indexation des fonds d'archives, il est élaboré avec le soutien de grandes bibliothèques (dont la bibliothèque du Congrès, qui en assure le développement).

FOAF = Friend Of A Friend (en français, « ami d'un ami »). Grammaire de description de l'identité personnelle et des relations d'un individu. Elle se fonde sur le modèle de conception du RDF.

IDPF = International Digital Publishing Forum. Association à l'origine du format ouvert EPUB, qui continue à en assurer les principaux développements.

JSON = Java Script Object Notation. Format de fichier, décrit par le RFC 4627, permettant l'interrogation et l'analyse de ressources dans un autre contenant, *via* une récolte de flux venus d'une autre localisation en ligne. Le fichier .json est mis à jour régulièrement avec par exemple des statistiques de publication des éléments concernés.

LCSH = Library of Congress Subject Headings (« autorités matière de la bibliothèque du Congrès »). Vocabulaire contrôlé de description du contenu des documents, élaboré par la bibliothèque du Congrès des Etats-Unis.

MADS = Metadata Authority Description Schema. Schéma de description des notices d'autorité, exprimé sous forme de XSD et conçu en complémentarité avec MODS et METS par la bibliothèque du Congrès.

MARC = MAchine Readable Catalog. Format de description bibliographique interprétable par des automates. La matrice d'origine, datant de la fin des années 1960, s'est déclinée en de nombreuses versions nationales (UKMARC, AUSMARC, InterMARC) ou internationales (MARC21, UNIMARC). Il a été adapté plus récemment au mode d'expression du XML, sous le nom de MARCXML.

METS = Metadata Encoding & Transmission Standard. Standard définissant les articulations entre différents documents et différents standards de métadonnées afin d'en faciliter le traitement. C'est donc un standard englobant qui peut régir l'utilisation d'autres standards tels que MODS, MADS ou XMP. La bibliothèque du Congrès est à l'origine de sa conception et assure le suivi de cet ensemble.

MeSH = Medical Subject Headings (« autorités matière en médecine »). Thésaurus international des termes médicaux.

MODS = Metadata Object Description Schema. Schéma de description de fichier XML hérité de MARC21. Le MODS est un format de description bibliographique portant sur des objets numériques (livres, revues, images, sons).

NLM = National Library of Medicine. Par extension, DTD décrivant un standard de métadonnées, dérivé du Dublin Core.

OpenURL = Open Unified Ressource Locator. Protocole d'identification d'une ressource intégrée dans son adresse URL. Créé par Herbert Van de Sompel, il a été certifié par l'agence de normalisation des Etats-Unis (NISO), sous le label Z3988. C'est la société Exlibris Group qui prend en charge le développement de ce standard.

OPML = Outline Processor Markup Language. Langage à balise permettant les échanges groupés de flux de syndication.

OWL = Web Ontology Language. Langage de description d'ontologies (ensembles structurés d'éléments sémantiques interdépendants). Plus développé que le SKOS, il est lui aussi fondé sur le RDF et développé par le consortium W3C.

PAM = PRISM Aggregator Message. Protocole de transfert des données en PRISM. Comme le PRISM il est géré par IDEALLiance. Le RSS 1.0 prend en compte ce protocole, mais pas (encore) le RSS 2.0.

PREMIS = PREservation Metadata : Implementation Strategies. Standard et vocabulaire de métadonnées destiné à une conservation pérenne des documents et des métadonnées. Il est géré par la bibliothèque du Congrès des Etats-Unis.

PRISM = Publishing Requirements for Industry Standard Metadata. Définit une organisation et des vocabulaires permettant l'agrégation, la structuration, la description et la diffusion multiple de données. C'est en même temps un schéma de métadonnées et une association de schémas existants, grâce à l'utilisation des espaces de nom. Il repose

sur les spécifications du XML et du RDF. Le PAM repose sur PRISM. Il est géré par IDEALLiance.

RDA = Resource Description and Access. Profil d'application du Dublin Core Metadata Initiative, en relation avec les Anglo - American Cataloguing Rules. L'initiative semble toutefois abandonnée.

RDF = Resource Description Framework. Principes grammaticaux d'expression des métadonnées. Mis en place par le W3C, cet ensemble de préconisations repose sur l'expression de triplets, un identifiant unique (« sujet ») étant associé à une propriété (« objet ») par un prédicat.

RDFa = Resource Description Framework (le « a » n'a pas de sens certain, il s'agit sans doute de « applied », en français « appliqué », ou de « attributes », en français « attributs »). Appliquant les préconisations du RDF, le RDFa insère des informations complémentaires dans le corps des pages au format XHTML. Il propose à la fois de reprendre une partie de la grammaire XHTML et d'utiliser des attributs propres, afin d'ajouter un verni « sémantique » à une page statique.

RFC = Request For Comments. Type de publication, liée à l'histoire d'Internet, visant à améliorer les propositions normatives en fonction des commentaires des utilisateurs. Avec le temps certains de ces documents deviennent des standards de fait (c'est le cas par exemple de la désignation des langues avec la RFC 3066).

RSS = acronyme signifiant soit Really Simple Syndication (syndication vraiment simple), soit Rich Site Summary (riche résumé de site), soit RDF Site Summary (résumé de site en RDF). Schéma de description des flux de documents, qui permet en particulier à un utilisateur de « s'abonner » à ce flux pour avoir l'information des mises à jour récentes. La syndication peut déclencher une interaction avec le flux suivi.

SKOS = Simple Knowledge Organization System. Principes de structuration de thésaurus ou de bases de connaissances reposant sur les recomandations du RDF. C'est une recommandation du W3C depuis le mois d'août 2009, ce qui laisse à penser que le SKOS sera largement adopté à l'avenir. Le vocabulaire d'autorités-matière de la bibliothèque du congrès, LCSH, est par exemple formulé en SKOS.

SVG = Scalable Vector Graphics (graphiques vectoriels à échelle adaptable). Format de fichier graphique en XML. Le dessin étant vectoriel, il peut être redimensionné sans perte de précision.

TEI = Texte Encoding Initiative. La TEI est un consortium fondé en 1987, encadrant la création de vocabulaires de description de contenu, en particulier de contenu textuel et iconographique. Par extension on parle de TEI pour désigner les vocabulaires correspondants. Ces préconisations de description fonctionnant par modules et par profils d'application, il existe de multiples versions de la TEI. Mais la TEI qu'on pourrait qualifier d' « officielle » et complète en est à sa cinquième version (on parle de TEI P5).

TGE ADONIS = Très Grand Etablissement ADONIS. Structure du ministère de l'enseignement et de la recherche français, dédiée à l'accompagnement des projets de recherche dans le cadre des technologies de la communication.

URI = Unified Resource Identifier (identifiant unifié de ressource). Identifiant propre à l'objet portant cette métadonnée, qui le distingue des autres objets. L'URL est une forme (imparfaite) d'URI.

URL = Unified Resource Locator (adresse unifiée de ressource). Adresse de page internet, incluant parfois des éléments dynamiques et des requêtes de l'utilisateur (comme dans le cas d'OpenURL).

VIAF = Virtual International Authority File (fichier international virtuel d'autorités). Base de données agrégeant et harmonisant les notices d'autorité de nombreuses bibliothèques nationales (dont la BnF).

VDX = Virtual Document eXchange (Echange de document virtuel).

W3C = World Wide Web Consortium. Consortium de conseil et de développement concernant le World Wide Web.

XML = eXtensible Markup Language. Langage à balises en alphabet Unicode. Dérivation simplifiée du SGML (Standard Generalized Markup Language), il repose sur le lien avec des grammaires définies à l'extérieur du fichier « .xml », les DTD (Document Type Definition), ou les XSD (XML Schema Description) qui peuvent être combinées grâce à l'utilisation des espaces de nom.

XMP = eXtended Metadata Platform. Enveloppe de métadonnées portant sur un fichier inclus, n'ayant pas au départ une aussi grande finesse de description (typiquement, actuellement, un fichier PDF).

XSD = XML Schema Description. Fichier en XML décrivant de manière systématique les champs et les contenus possibles d'un type de fichier XML donné.

GLOSSAIRE DES EXPRESSIONS

Atom : format de syndication de contenu, concurrent de RSS.

Crosswalk (en anglais, « passage piéton ») : processus, décrit en théorie et / ou en pratique, permettant de passer d'un standard de métadonnées à un autre, c'est-à-dire de convertir des métadonnées dans un autre format. La description la plus fréquente est une mise en correspondance des « mappings ».

Espace de nom (en anglais on parle de « namespace ») : déclaration insérée dans le balisage des fichiers XML permettant au logiciel interprétant ce fichier de se référer à la grammaire de métadonnées adaptée.

Mapping (en anglais, « cartographie ») : liste, la plupart du temps ordonnée et organisée, des labels et éventuellement des contenus des champs de métadonnées, pour un standard donné.

Primer : Document d'introduction technique à un schéma de métadonnées. Il contient généralement une description fonctionnelle étendue, des exemples et des listes de termes employés par le schéma, mais n'évoque pas l'ensemble des particularités ou possibilités offertes. Les Primers sont de plus en plus répandus pour encadrer l'usage des standards de métadonnées.

Profil d'application : déclinaison d'un standard de métadonnées (typiquement le Dublin Core) adaptée à une situation particulière.

Résolveur de lien : application en ligne permettant de rediriger l'utilisateur vers le document cherché à partir d'une adresse codée, mais en lui proposant des services associés. Le site de Crossref a par exemple un résolveur de lien dédié à l'exploitation du DOI.

Conversions de formats

La conversion de Zotero à MODS, qui a servi de base aux autres transformations, avait été initialement imaginée par Nicolas Barts, Jean-François Rivière et Jean-Baptiste Bertrand. Zotero a servi de format de départ en raison des possibilités de vérification qu'il offrait, mais les spécificités des formats d'arrivée ont été également étudiées pour le cas d'une opération inverse.

ZOTERO VERS MODS

MODS a été abandonné pour l'instant par Revues.org

Pour les articles

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
Item Type	Article de revue	<genre authority="local">journalArticle</genre>	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
Titre (item.title)	Titre de l'article Sous-titre de l'article	<titleInfo> <title>**Titre de l'article**</title> <subTitle>**Sous-titre de l'article**</subTitle> </titleInfo>	Si le titre est suivi d'un sous-titre, il faut faire suivre le titre d'un point et d'un espace sinon le titre et sous-titre seront collés dans Zotero
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="family">**Nom**</namePart> <role> <roleTerm type="code">aut</roleTerm> <roleTerm type="text">author</roleTerm> </role> </name>	pour l'instant, le traducteur MODS de Zotero est défectueux et ne prend pas en compte les rôles des contributeurs (il les considère tous comme auteurs). Hypothèse : la double mention par code et texte pose-t-elle le problème? Ou bien est-ce à cause du "else if" dans le Translator qui fait un "ou" exclusif?
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="family">**Nom**</namePart> <role> <roleTerm type="code">trl</roleTerm> <roleTerm type="text">translator</roleTerm> </role> </name>	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="Family">**Nom**</namePart> <role> <roleTerm type="code">edt</roleTerm> <roleTerm type="text">editor</roleTerm> </role> </name>	
Auteur critique (creatorTypes.commenter)	Nom, Prénom du commentateur (à ne pas confondre avec l'auteur de la critique, auteur)	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="Family">**Nom**</namePart> <role> <roleTerm type="code">cmm</roleTerm> <roleTerm type="text">commentator</roleTerm> </role> </name>	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
Contributeur (creatorTypes.contributor)	Nom, Prénom du collaborateur	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="Family">**Nom**</namePart> <role> <roleTerm type="code">ctb</roleTerm> <roleTerm type="text">contributor</roleTerm> </role> </name>	
résumé (item.abstractNote)	résumé abstract resumen …	<abstract lang="fr" xml:lang="fr">**résumé**</abstract> <abstract lang="en" xml:lang="en">**abstract**</abstract> <abstract lang="es" xml:lang="es">**resumen**</abstract> …	S'il y a des résumés en différentes langues, il est nécessaire d'insérer un séparateur entre eux, sinon ils apparaitront collés dans Zotero
publication (item.publicationTitle)	Nom de la revue	<relatedItem type="host"> ... <titleInfo> <title>**Nom de la revue**</title> </titleInfo> ... </relatedItem>	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
volume (item.volume)	volume du numéro dans lequel est contenu l'article ? Ou nombre d'années de publication (chaque année étant subdivisée en numéros, numérotation recommencée chaque année)?	<relatedItem type="host"> ... <part> <detail type="volume"> <number>**volume de la revue**</number> </detail> </part> ... <relatedItem type="host">	
numéro (item.issue)	numéro de la revue ou titre du numéro de la revue	<relatedItem type="host"> ... <part> <detail type="issue"> <number>**numéro de la revue**</number> </detail> </part> ... <relatedItem type="host">	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
pages (item.pages)	pagination	<relatedItem type=“host”> <part> <extent unit="page"> <start> **page de début** </start> <end> **page de fin**</end> </extent> </part> </relatedItem>	
date (item.date)	date de publication	<relatedItem type="host"> <originInfo> ... <dateIssued>**date de publication**</dateIssued> ... </originInfo> </relatedItem>	
abréviation de journal (item.journalAbbreviation)	Titre abrégé de la revue	<relatedItem type="host"> <titleInfo type = "abbreviated"> <title>**Titre abrégé de la revue**</title> </titleInfo> </relatedItem>	
langue (item.language)	nom de la langue	<language> <languageTerm type="text">**langue**</languageTerm> </language>	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
DOI (item.DOI)	DOI	<relatedItem type="host"> ... <identifier type="doi">**DOI**</identifier> ... </relatedItem>	
ISSN (item.ISSN)	numéro ISSN	<relatedItem type="host"> ... <identifier type="issn">**ISSN**</identifier> ... </relatedItem>	
Titre abrégé (item.shortTitle)	Titre abrégé de l'article	Dans le XSD MODS 3, dans le "complexType" "titleInfoType", voilà ce qui pourrait aller : <xsd:attribute name="type" use="optional"> <xsd:simpleType> <xsd:restriction base="xsd:string"> <xsd:enumeration value="abbreviated">**Titre abrégé de l'article**</xsd:enumeration> </xsd:restriction> </xsd:simpleType> </xsd:attribute>	champ peu utilisé, mais quand il l'est ce serait dommage de ne pas en tenir compte, certains titres sont parfois très longs…
url (item.url)	URL de l'article	<location> <url>**url de l'article**</url> </location>	

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	<relatedItem type="host"> <originInfo> … <dateCaptured>**AAAA-MM-JJ heure:minute:seconde**</dateCaptured> ... </originInfo> ... </relatedItem>	
Dépôt (item.repository)	Est-ce que cela permettrait de distinguer droit d'auteur (dépôt) et conditions d'utilisation (autorisation)?	<accessCondition type="restriction on access"> **Licence portant sur le document** </accessCondition>	**voir aussi http://www.loc.gov/standards/mods/v3/mods-userguide-elements.html#accesscondition pour plus de précisions** **Une autre balise exploitable, avec du développement mais mieux adapté aux besoins, serait <holdingExternal>**
autorisation (item.rights)	droits d'auteur	<accessCondition type="use and reproduction"> **Droits d'auteur** </accessCondition>	
marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	<subject> <topic xml:lang="fr">mot clef 1</topic> <topic xml:lang="fr">mot clef 2</topic> <topic xml:lang="en">keyword 1</topic> <topic xml:lang="en">keyword 2</topic> </subject>	prend seulement en compte les balises de type <topic>. Ne prend pas en compte les balises de type <geographical>

Champ Zotero	Valeur du champ	balisage MODS	Commentaires
notes attachées	contenu de la note 1 contenu de la note 2 …	<note type="content">contenu de la note 1</note> <note type="content">contenu de la note 2</note>	
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	<location> ... <url access="raw object" displayLabel="titre du lien 1">url du lien 1</url> <url access="raw object" displayLabel="titre du lien 2">url du lien 2</url> ... </location>	Permet de donner des liens vers différents types de documents **MODS: est-ce que la syntaxe de relatedItem est ici utilisée ? Il me semble qu'elle est adaptée à la situation (illustrations et compléments de l'article, même si sont plutôt visés les articles citants ou contextes) ?**

Pour les chapitres de livres

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
Item Type	Chapitre de livre	<genre authority="local">bookSection</genre> Avec également dans le XSD cette portion qui peut être utile pour cette information et d'autres (mais s'utilise-t-elle pour le niveau article ou le niveau livre?): <xsd:complexType name="detailType">	S'il y a une balise book ou bookSection Zotero identifie le marcGenre comme livre et le type texte (si reconnaissance conforme il réfère au standard)

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
		<xsd:choice maxOccurs="unbounded"> <xsd:element name="number" type="xsd:string"/> <xsd:element name="caption" type="xsd:string"/> <xsd:element name="title" type="xsd:string"> <xsd:annotation> <xsd:documentation>Use only if different than main title of resource being described</xsd:documentation> </xsd:annotation> </xsd:element> </xsd:choice> <xsd:attribute name="type" type="xsd:string"> <xsd:annotation> <xsd:documentation>Suggested values: part, volume, issue, chapter, section, paragraph, track....</xsd:documentation> </xsd:annotation> </xsd:attribute> <xsd:attribute name="level" type="xsd:positiveInteger"> ... </xsd:attribute> </xsd:complexType>	
Titre (item.title)	Titre du chapitre Sous-titre du chapitre	<titleInfo> <title>**Titre du chapitre**</title> <subtitle>**Sous-titre du chapitre**</subtitle> </titleInfo>	Si le titre est suivi d'un sous-titre, il faudrait faire suivre le titre d'un point et d'un espace, sinon le titre et sous-titre apparaitront collés

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="family">**Nom de famille**</namePart> <role> <roleTerm type="code" authority="marcrelator">aut</roleTerm> <roleTerm type="text" authority="marcrelator">author</roleTerm> </role> </name>	pour l'instant, le traducteur MODS de Zotero est défectueux et ne prend pas en compte les rôles des contributeurs (il les considère tous comme auteurs). Hypothèse : la double mention par code et texte pose-t-elle le problème? Ou bien est-ce à cause du "else if" dans le Translator qui fait un "ou" exclusif? **A préférer, quand indiqué, l'auteur / collaborateur / traducteur concerné par le chapitre et non le livre dans son ensemble. La mention d'éditeur de série / directeur de collection n'est que rarement renseignée.**
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="family">**Nom**</namePart> <role> <roleTerm type="code" authority="marcrelator">trl</roleTerm> <roleTerm type="text" authority="marcrelator">translator</roleTerm> </role> </name>	
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="Family">**Nom**</namePart> <role> <roleTerm type="code">edt</roleTerm> <roleTerm type="text">editor</roleTerm> </role> </name>	

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
Editeur de la série (creatorTypes.seriesEditor)	Nom, Prénom de l'éditeur de la série	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="Family">**Nom**</namePart> <role> <roleTerm type="code">pbd</roleTerm> <roleTerm type="text">publishing director</roleTerm> </role> </name>	
Contributeur (creatorTypes.contributor)	Nom, Prénom du collaborateur	<name type="personal"> <namePart type="given">**Prénom**</namePart> <namePart type="Family">**Nom**</namePart> <role> <roleTerm type="code">ctb</roleTerm> <roleTerm type="text">contributor</roleTerm> </role> </name>	
Résumé (item.abstractNote)	résumé abstract resumen	<abstract lang="fr" xml:lang="fr">**résumé**</abstract> <abstract lang="en" xml:lang="en">**abstract**</abstract> <abstract lang="es" xml:lang="es">**resumen**</abstract> …	S'il y a des résumés en différentes langues, il est nécessaire d'insérer un séparateur entre eux, sinon ils apparaitront collés dans Zotero
titre du livre (item.bookTitle)	Titre du livre	<relatedItem type="host"> ... <titleInfo> <title>**titre du livre**</title> </titleInfo> ... </relatedItem>	

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
Série (item.series)	Nom de la collection	<relatedItem type="series"> <titleInfo> <title>**Nom de la collection**</title> </titleInfo> </relatedItem>	
Numéro de la série (itemFields.seriesNumber)	Numéro de la collection	<relatedItem type="host"> ... <identifier type="issn">**ISSN**</identifier> ... </relatedItem>	La numérotation de collection la plus courante et la plus utile est l'ISSN, donc même si Zotero distingue les 2 on peut utiliser ISSN en MODS
Volume (item.Volume)	numéro du volume ou titre du volume (à éviter)	<relatedItem type="host"> <part> <detail type="volume"> <number>**numéro du volume**</number> </detail> </part> </relatedItem>	
Nbr de volumes (itemFields.numberOfVolumes)	nombre de volumes dans la collection	N'existe pas en MODS En revanche, il serait possible d'indiquer l'adresse du site de la collection avec <location> <url displayLabel="URL de la collection">http://...</url> </location>	L'information possible avec MOD (identifiant web de collection) est plus utile au quotidien que l'item Zotero, très secondaire (et pas disponible dans le cas, fréquent, d'une collection "ouverte").
Edition (itemFields.edition)	Edition (version du texte)	<originInfo> <edition>**Date ou rang de l'édition**<edition> </originInfo>	Pour des questions de présentation il est préférable d'afficher le contenu de la balise sous la forme : Edition : xxx car les usages ne sont pas normalisés entre par exemple "édition 2009", "édition 2.3" et "3ème édition"

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
éditeur (item.publisher)	nom de l'éditeur	<relatedItem type="host"> <originInfo> <publisher>**nom de l'éditeur**</publisher> </originInfo> </relatedItem>	
lieu (item.place)	lieu d'édition	<relatedItem type="host"> <originInfo> …. <place> <placeTerm type=“text”>**lieu d'édition**</placeTerm> </place> …. </originInfo> </relatedItem>	
date (item.date)	date de publication	<relatedItem type="host"> <originInfo> ... <dateIssued>**date de publication**</dateIssued> ... </originInfo> </relatedItem>	Préférer la date de l'article (= du chapitre), en particulier dans le cas des compilations d'articles de dates différentes. Si le livre et le chapitre sont publiés par Revues.org, il faudrait indiquer à la fois la date pour chaque chapitre et la date du livre (option d'application automatique à toutes les parties?).
pages (item.pages)	pagination	<relatedItem type=“host”> <part> <extent unit="page"> <start>**page de début**</start> <end>**page de fin**</end> </extent> </part> </relatedItem>	

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
	section contenant le document	<detail type="section"> <number>**Numéro de section**</number> </detail> *ou* <detail type="section"> <text>**Nom de section**</text> </detail>	**Ce champ peut être utile**
langue (item.language)	nom de la langue	<language> <languageTerm type="text">**langue**</languageTerm> </language>	Ne prend en compte la balise <languageTerm> que lorsqu'elle a un attribut type="text". Ne prend par exemple pas en compte les balises du type <languageTerm authority="rfc3066" type="code">
DOI (item**Fields**.DOI)	DOI	<relatedItem type="host"> ... <identifier type="doi">**DOI**</identifier> ... </relatedItem>	Non nécessaire pour Zotero pour ce type de document, il peut être utile de prévoir son existence et d'en assurer la transmission en MODS
ISBN (item.ISBN)	numéro ISBN	<relatedItem type="host"> …. <identifier type="isbn">**numéro ISBN**</identifier> …. </relatedItem>	
Titre abrégé du livre (itemFields.shortTitle)	Titre abrégé du livre	<relatedItem type="host"> <titleInfo type = "abbreviated"> <title>**Titre abrégé du livre**</title> </titleInfo> </relatedItem>	
url (item.url)	URL du chapitre	<location> <url>**url du chapitre**</url> </location>	

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	<relatedItem type="host"> <originInfo> … <dateCaptured>**AAAA-MM-JJ heure:minute:seconde**</dateCaptured> ... </originInfo> ... </relatedItem>	
Dépôt (item**Fields**.repository)	Est-ce que cela permettrait de distinguer droit d'auteur (dépôt) et conditions d'utilisation (autorisation)?	<accessCondition type="use and reproduction"> Droits d'auteur </accessCondition>	**voir aussi http://www.loc.gov/standards/mods/v3/mods-userguide-elements.html#accesscondition pour plus de précisions** **Une autre balise exploitable, avec du développement mais mieux adapté aux besoins, serait <holdingExternal>**
autorisation (item.rights)	droits d'auteur	<accessCondition type="restriction on access"> Licence portant sur le document </accessCondition>	
onglet marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	<subject> <topic>mot-clef 1</topic> <topic>mot-clef 2</topic> <topic>keyword 1</topic> <topic>keyword 2</topic> </subject>	prend seulement en compte les balises de type <topic>. Ne prend par exemple pas en compte les balises de type <geographical>.
notes attachées	contenu de la note 1 contenu de la note 2 …	<note type="content">contenu de la note 1</note> <note type="content">contenu de la note 2</note>	

Champ Zotero	Valeur du champ	Balisage MODS	Commentaires
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	<location> ... <url access="raw object" displayLabel="titre du lien 1">url du lien 1</url> <url access="raw object" displayLabel="titre du lien 2">url du lien 2</url> ... </location>	Permet de donner des liens vers différents types de documents **MODS: est-ce que la syntaxe de relatedItem est ici utilisée ? Il me semble qu'elle est adaptée à la situation (illustrations et compléments de l'article, même si sont plutôt visés les articles citants ou contextes) ?**

ZOTERO VERS BIBTEX

Pour les articles

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
Item Type	Article de revue	l'entête d'un item de type article est @article (	

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
variable interne, numéro dans la base Zotero	numéro dans la base	clé d'identification établie par le créateur du fichier BibTeX	La mention du type de fichier est suivie dans l'exemple de "nom-titre-2009," mais il est difficile de savoir d'où Zotero tire cette identification de source, sachant que "nom" a une majuscule dans les champs Zotero, et "titre" n'est pas isolé dans un champ. Quoi qu'il en soit le risque est fort de créer des conflits d'identifiants sans réflexion préalable sur le nommage
Titre (item.title)	Titre de l'article Sous-titre de l'article	title = {**Titre de l'article Sous-Titre de l'article**},	
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	author = {**Nom, Prénom** and **Nom, Prénom** and **Nom, Prénom}** Préférable à l'ordre "Prénom Nom", qui est aussi possible, pour une meilleure distinction des parties du nom	A l'export Zotero inverse les mentions de responsabilités pour faire en BibTeX "Prénom Nom" Le champ "author" est obligatoire
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	tranlator = {**Nom, Prénom**}	
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication	editor= {**Nom, Prénom**}	Mais pas de balise spécifique pour collaborateur
collaborateur (creatorTypes.collaborator) ou auteur critique (creatorTypes.commenter)	Nom, prénom du collaborateur ou du commentateur		

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
résumé (item.abstractNote)	résumé abstract resumen …	abstract = {**résumé abstract resumen**},	S'il y a des résumés en différentes langues, il est nécessaire d'insérer un séparateur entre eux, sinon ils apparaitront collés dans Zotero et BibTeX mettra les mots avec majuscule incluse entre parenthèses. De façon générale d'ailleurs le traitement de la casse est très sensible sous BibTeX, c'est pourquoi Zotero rajoute des {} partout où il peut y avoir litige.
publication (item.publicationTitle)	Nom de la revue	journal = {**Nom de la revue**}	"journal" est obligatoire
Volume (itemFields.volume)	volume de la revue	volume = {**numéro d'année de la revue**},	
numéro (item.issue)	numéro de la revue *ou* *titre du numéro de la revue*	number = {**numéro de la revue**},	
pages (item.pages)	pagination	pages = (**pageDébut**--**pageFin**)	NB l'insertion d'un tiret à l'export Zotero pour coller à la grammaire BibTeX
date (item.date)	date de publication	month = **moisdepublication,** ET year = {**année de publication**},	"month" devrait être un abrégé en trois lettre du nom anglais du mois, s'il est présent "year" est obligatoire
Pas pour les articles	Lieu de publication	address = {**adresse physique de l'éditeur**}	

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
abréviation de journal (item.journalAbbreviation)	Titre abrégé de la revue		Le nom d'un journal peut être abrégé en utilisant une "chaîne", même si le rapport entre travail et bénéfice d'information est ici élevé
langue (item.language)	nom de la langue	langue = {**langue**},	Zotero n'exporte pas les informations contenues dans le champ Langue, car ce champ n'est pas toujours reconnu dans BibTeX. Il n'est cependant pas difficile de le proposer, de façon normalisée ou non
DOI (itemFields.DOI)	DOI	doi = {**DOI**},	Les références consultées en ligne indiquent aussi le champ "crossref" pour les références croisées type DOI, mais puisqu'il existe un champ "doi" supporté, par exemple, par Jabref…
ISSN (item.ISSN)	numéro ISSN	issn = {**ISSN**},	Attention, l'ISSN n'est pas toujours pris en compte par les logiciels BiBTeX. Mais qui peut le plus peut le moins…
Titre abrégé (itemFields.shortTitle)	Titre abrégé de l'article		Pas forcément nécessaire de proposer la transformation de Zotero en BibTeX
url (item.url)	URL de l'article	url = {**URL**},	
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	urldate = {**Date de dernière visite de la page**}, ? Par l'auteur / éditeur ou par l'utilisateur ? - champ initié par Jurabib	

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
Dépôt (itemFields.repository) ou autorisation (itemFields.rights)	Droits d'auteur et Licence	publisher = {**éditeur commercial**}, et copyright = {**expression du droit d'auteur**},	Les différents duos ne se recoupent pas forcément, mais en BibTeX il serait intéressant de référencer à la fois les droits "patrimoniaux" et les droits "moraux". "copyright" n'est toutefois pas toujours pris en compte. Zotero produit pour ce genre de document "copyright" mais pas "publisher".
marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	keywords = {**keyword 1**, **keyword 2**, **mot-clef 1**, **mot-clef 2}**,	
notes attachées	contenu de la note 1 contenu de la note 2 …	contents = {**table des matières**}, annote = {…}, annote = {…}, annote = {…}	Contents n'est pas toujours reconnu en BibTeX; l'uage n'est pas figé sur la hiérarchisation et les ponctuations… Beaucoup d'incertitudes donc, "annote" est sans doute préférable.
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	crossref = {{**URL1**}}, crossref = {{**URL2**}}, crossref = {{**URL3**}},	L'utilisation de "crossref" pourrait bien s'adapter, d'après moi, à des liens attachés dans un contexte "tout en ligne". Attention, il faut certainement placer aussi des pointeurs (et des retours ?) sur les documents cités.
		institution = {**Nom de l'institution**}	Une mention de responsabilité intéressante, qui correspond aussi aux besoins des partenaires de Revues.org (revues éditées par un laboratoire, sous la direction d'une personne physique par ex.)
		price = {**Prix du document**}	Balise assez rare pour être signalée. Elle n'est toutefois pas reconnue par tous les interprètes BibTeX

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
		)	ferme le fichier (parenthèse ouverte après le type de document)

Pour les chapitres de livres

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
Item Type	Chapitre de livre	l'entête d'un item de type chapitre de livre est @incollection (	Dans le cas d'une référence à un livre complet, le type "livre" s'identifie en BibTeX avec @book Est préférable à @inbook pour la possibilité de nommer le chapitre. Au pire, ce nom peut être un numéro... Optional fields: editor, pages, organization, publisher, address, month, note, key
variable interne, numéro dans la base Zotero	numéro dans la base	clé d'identification établie par le créateur du fichier BibTeX	La mention du type de fichier est suivie dans l'exemple de "nom-titre-2009," mais il est difficile de savoir d'où Zotero tire cette identification de source, sachant que "nom" a une majuscule dans les champs Zotero, et "titre" n'est pas isolé dans un champ. Quoi qu'il en soit le risque est fort de créer des conflits d'identifiants sans réflexion préalable sur le nommage
Titre (item.title)	Titre du chapitre Sous-titre du chapitre	title = {**Titre de l'article Sous-Titre de l'article**},	nécessaire

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
Titre du livre (item.bookTitle)	Titre du livre	booktitle = {**Titre du livre Sous-titre du livre**}	nécessaire
Série (item.series)	Nom de la collection	series = {**Nom de la collection**}	
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	author = {**Nom, Prénom** and **Nom, Prénom** and **Nom, Prénom}** Préférable à l'ordre "Prénom Nom", qui est aussi possible, pour une meilleure distinction des parties du nom	A l'export Zotero inverse les mentions de responsabilités pour faire en BibTeX "Prénom Nom" Le champ "author" est obligatoire Il est préférable d'identifier les mentions de responsabilité au niveau du chapitre et non dans les métadonnées du livre.
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	tranlator = {**Nom, Prénom**}	Il est préférable d'identifier les mentions de responsabilité au niveau du chapitre et non dans les métadonnées du livre.
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication	editor = {**Nom, Prénom**}	
éditeur de la série (creatorTypes.seriesEditor) ou collaborateur (creatorTypes.contributor)	Nom, prénom de l'éditeur de collection ou du collaborateur		La mention d'éditeur de série / directeur de collection n'est que rarement renseignée.

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
résumé (item.abstractNote)	résumé abstract resumen …	abstract = {**résumé abstract resumen**},	S'il y a des résumés en différentes langues, il est nécessaire d'insérer un séparateur entre eux, sinon ils apparaitront collés dans Zotero et BibTeX mettra les mots avec majuscule incluse entre parenthèses. De façon générale d'ailleurs le traitement de la casse est très sensible sous BibTeX, c'est pourquoi Zotero rajoute des {} partout où il peut y avoir litige.
Série (item.series)	Nom de la collection	series = {**Nom de la collection**}	
Numéro de la série (itemFields.seriesNumber)	Numéro de la collection	issn = {**Numéro ISSN de la collection**},	La numérotation de collection la plus courante et la plus utile est l'ISSN, donc même si Zotero distingue les 2 on peut utiliser ISSN en BibTeX
Volume (item.Volume)	numéro du volume ou titre du volume (à éviter)	volume = {**Numéro du volume du livre**},	volume est plus adapté que number
Nbr de volumes (itemFields.numberOfVolumes)	nombre de volumes dans la collection	n'existe pas en BibTeX	il n'est pas nécessaire de renseigner et de tenir compte de ce champ
Edition (itemFields.edition)	Edition (version du texte)	edition = {**Nombre ordinal indiquant le numéro d'édition**}	Pour des questions de présentation il est préférable d'afficher le contenu de la balise sous la forme : Edition : xxx car les usages ne sont pas normalisés entre par exemple "édition 2009", "édition 2.3" et "3ème édition". BibTeX préconise plutôt l'écriture en lettres de nombres ordinaux

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
lieu (item.place)	lieu d'édition	address = {**adresse physique de l'éditeur**}	
pages (item.pages)	pagination	pages = (**pageDébut**--**pageFin**)	NB l'insertion d'un tiret à l'export Zotero pour coller à la grammaire BibTeX
date (item.date)	date de publication	month = **moisdepublication,** ET year = {**année de publication**} OU date = {**aaaa-mm-jj**}	"month" devrait être un abrégé en trois lettre du nom anglais du mois, s'il est présent "year" est obligatoire Préférer la date de l'article (= du chapitre), en particulier dans le cas des compilations d'articles de dates différentes. Si le livre et le chapitre sont publiés par Revues.org, il faudrait indiquer à la fois la date pour chaque chapitre et la date du livre (option d'application automatique à toutes les parties?). La norme préconisée par les utilisateurs de BibTeX est ISO 8601
langue (item.language)	nom de la langue	langue = {**langue**},	Zotero n'exporte pas les informations contenues dans le champ Langue, car ce champ n'est pas toujours reconnu dans BibTeX. Il n'est cependant pas difficile de le proposer, de façon normalisée ou non
DOI (itemFields.DOI)	DOI	doi = {**DOI**},	Les références consultées en ligne indiquent aussi le champ "crossref" pour les références croisées type DOI, mais puisqu'il existe un champ "doi" supporté, par exemple, par Jabref…
ISBN (item.ISBN)	numéro ISBN	isbn = {**ISBN**},	Attention, l'ISBN n'est pas toujours pris en compte par les logiciels BiBTeX. Mais qui peut le plus peut le moins…

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
Titre abrégé (itemFields.shortTitle)	Titre abrégé du livre		Pas forcément nécessaire de proposer la transformation de Zotero en BibTeX
url (item.url)	URL du chapitre	url = {**URL**},	
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	urldate = {**Date de dernière visite de la page**}, ? Par l'auteur / éditeur ou par l'utilisateur ? - champ initié par Jurabib	
Dépôt (itemFields.repository) ou autorisation (itemFields.rights)	Droits d'auteur et Licence	publisher = {**éditeur commercial**}, et copyright = {**expression du droit d'auteur**},	Les différents duos ne se recoupent pas forcément, mais en BibTeX il serait intéressant de référencer à la fois les droits "patrimoniaux" et les droits "moraux". "copyright" n'est toutefois pas toujours pris en compte. Zotero produit pour ce genre de document "copyright" mais pas "publisher".
marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	keywords = {**keyword 1**, **keyword 2**, **mot-clef 1**, **mot-clef 2}**,	
notes attachées	contenu de la note 1 contenu de la note 2 …	contents = {**table des matières**}, annote = {…}, annote = {…}, annote = {…}	Contents n'est pas toujours reconnu en BibTeX; l'uage n'est pas figé sur la hiérarchisation et les ponctuations… Beaucoup d'incertitudes donc, "annote" est sans doute préférable.

Champ Zotero	Valeur du champ	balisage BibTeX	Commentaires
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	crossref = {{**URL1**}}, crossref = {{**URL2**}}, crossref = {{**URL3**}},	L'utilisation de "crossref" pourrait bien s'adapter, d'après moi, à des liens attachés dans un contexte "tout en ligne". Attention, il faut certainement placer aussi des pointeurs (et des retours ?) sur les documents cités.
		chapter = {**Numéro de chapitre**}	Optionnel si tous les chapitres ont un nom; utile dans le cas contraire…
		institution = {**Nom de l'institution**}	Une mention de responsabilité intéressante, qui correspond aussi aux besoins des partenaires de Revues.org (revues éditées par un laboratoire, sous la direction d'une personne physique par ex.)
		price = {**Prix du document**}	Balise assez rare pour être signalée. Elle n'est toutefois pas reconnue par tous les interprètes BibTeX
		)	ferme le fichier (parenthèse ouverte après le type de document)

ZOTERO VERS RIS

Pour les articles

Champ Zotero	Valeur du champ	balisage RIS	Commentaires

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
Item Type	Article de revue	TY avec la valeur JOUR ou éventuellement MGZN (pour un magazine) / GEN (pour un document général)…	Doit être la première balise du fichier
Titre (item.title)	Titre de l'article Sous-titre de l'article	T1 -- Titre de l'article T2 -- Sous-titre de l'article	Pour les champs de plus de 70 caractères (dont parfois les titres) il faut insérer des retours chariot. Plusieurs titres de même niveau peuvent être déclarés, par exemple avec des langues différentes.
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	A1	L'asterisque (caractère 42) n'est pas autorisé pour les champs auteur. Zotero ne reconnaît que les balises A1 ou AU (->Auteur) et A2 ou ED (->Collaborateur), pas A3.
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	A2	
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication	ED	
collaborateur (creatorTypes.collaborator) ou auteur critique (creatorTypes.commenter)	Nom, prénom du collaborateur ou du commentateur	A3	

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
résumé (item.abstractNote)	résumé abstract resumen …	N2	S'il y a des résumés en différentes langues, il est nécessaire d'insérer un séparateur entre eux, sinon ils apparaîtront collés dans Zotero
publication (item.publicationTitle)	Nom de la revue	JF	L'asterisque (caractère 42) n'est pas autorisé dans le nom de revue
Volume (itemFields.volume)	volume de la revue	VL	
numéro (item.issue)	numéro de la revue *ou* *titre du numéro de la revue*	IS	
pages (item.pages)	pagination	SP -- **page de début** EP -- **page de fin**	Attention, il faut 2 champs RIS pour un champ Zotero ou Lodel; en se mettant s'accord pour un signe séparateur (par exemple le tiret haut, très utilisé "-") ce devrait être gérable
date (item.date)	date de publication	PY	
Pas pour les articles	Lieu de publication	CY	
abréviation de journal (item.journalAbbreviation)	Titre abrégé de la revue	JA ou JO, J1, J2 (non produits par Zotero mais reconnus par le format)	

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
langue (item.language)	nom de la langue		Zotero n'exporte pas les informations contenues dans le champ Langue
DOI (itemFields.DOI)	DOI	M3	Les champs M1à M3 sont "divers" et définis par convention d'application. Zotero place le DOI en M3
ISSN (item.ISSN)	numéro ISSN	SN	
Titre abrégé (itemFields.shortTitle)	Titre abrégé de l'article	T1	Plusieurs titres de même niveau peuvent être indiqués (par exemple dans différentes langues). Il est préférable de placer le titre abrégé en T1 et le sous-titre en T2
url (item.url)	URL de l'article	UR	
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	Y2 -- AAAA/MM/JJ/ Consultation à h;min;sec	Proposition d'adaptation, sachant que les "/" doivent rester présents dans ce champ. Le texte à la fin est libre, mais il faut vérifier que les ";" ne posent pas de problème d'interprétation par les logiciels
Dépôt (itemFields.repository) ou autorisation (itemFields.rights)	droits d'auteur et Licence		RIS ne gère directement ni droits d'auteur ni droits d'accès. La Champ AV (availability) est en fait un pointeur vers le texte complet. Restent les champs définis par les usages (M..., U...), mais ils sont déjà assez "occupés" par les autres informations utiles sur le document.

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
variable interne, numéro dans la base Zotero	numéro dans la base	ID	Pourrait servir, avec une table des correspondances, à pallier l'absence d'indication sur les droits (le champ compte jusqu'à 255 caractères, de préférence en capitales, ce qui est assez ample)
marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	KW KW KW KW	L'asterisque (caractère 42) n'est pas autorisé pour les champs mots-clés
notes attachées	contenu de la note 1 contenu de la note 2 …	N1 N1 N1	
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	L1 -- Lien vers un PDF L2 -- Lien vers du plein-texte (en HTML par exemple) L3 -- Lien vers des documents en relation L4 -- Lien vers une image (par exemple couverture, image d'accroche, illustration)	AV permet de pointer vers un PDF ou Postscript, mais la gestion avec Pdfroot est un peu complexe. Il est préférable de décliner, en en respectant les différences, entre les lignes L1 à L4. Chaque balise peut être répétée.
		ER	Doit être la dernière balise du fichier

Pour les chapitres de livres

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
Item Type	Chapitre de livre	TY avec la valeur CHAP (chapitre de livre) ou éventuellement BOOK (pour un livre complet) / GEN (pour un document général) / MAP (pour une carte) / RPRT (pour un rapport)…	Doit être la première balise du fichier
Titre (item.title)	Titre du chapitre Sous-titre du chapitre	T1 -- Titre du chapitre T1 -- Sous-titre du chapitre T1 -- titre abrégé du chapitre	Pour les champs de plus de 70 caractères (dont parfois les titres) il faut insérer des retours chariot. Plusieurs titres de même niveau peuvent être déclarés, par exemple avec des langues différentes.
Titre du livre (item.bookTitle)	Titre du livre	T2	
Série (item.series)	Nom de la collection	T3	
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	A1	L'asterisque (caractère 42) n'est pas autorisé pour les champs auteur. Zotero ne reconnaît que les balises A1 ou AU (->Auteur) et A2 ou ED (->Collaborateur), pas A3.
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	A2	

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication	ED	
collaborateur (creatorTypes.collaborator) ou éditeur de la série (creatorTypes.seriesEditor)	Nom, prénom du collaborateur ou de l'éditeur de collection	A3	
résumé (item.abstractNote)	résumé abstract resumen …	N2 -- résumé N2 -- abstract N2 -- resumen ou N2 -- résumé + abstract + resumen	S'il y a des résumés en différentes langues, il sera utile de les placer dans des champs N2 séparés pour le format RIS. Mais tous ne seront pas forcément récupérés par les logiciels bilbliographiques, donc il est sans doute plus prudent de garder 1 champ N2 et de séparer les résumés langue par langue.
Volume (itemFields.volume)	volume dans la collection	VL	Peu utile, mais possible
numéro de la série (itemFields.seriesNumber)	numéro ISSN de la collection	M2 -- ISSN xxxxx xxxxx	La balise SN est valable aussi bien pour l'ISSN que l'ISBN. Il est donc préférable, pour un livre, de garder SN pour l'ISBN (presque toujours présent) et de proposer en M2 par exemple "ISSN" et le numéro ISSN de la collection

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
Nbr de volumes (itemFields.numberOfVolumes)	nombre de volumes dans la collection	Pas intéressant de façon générale, de toutes façons	
Edition (itemFields.edition)	Edition (version du texte)	Dans PY (voir plus bas)	
pages (item.pages)	pagination	SP -- **page de début** EP -- **page de fin**	Attention, il faut 2 champs RIS pour un champ Zotero ou Lodel; en se mettant s'accord pour un signe séparateur (par exemple le tiret haut, très utilisé "-") ce devrait être gérable
date (item.date)	date de publication	PY	La balise peut comporter, outre la date d'édition, une mention du genre "2 édition"; c'est donc là qu'il vaut mieux appliquer les données du champ Zotero Edition.
éditeur (item.publisher)	nom de l'éditeur	PB	
lieu (item.place)	Lieu de publication	CY	
abréviation de journal (item.journalAbbreviation)	Titre abrégé de la revue	JA ou JO, J1, J2 (non produits par Zotero mais reconnus par le format)	

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
langue (item.language)	nom de la langue		Zotero n'exporte pas les informations contenues dans le champ Langue car RIS ne les gère pas d'origine. Il serait utile de placer la (les) langues dans des champs RIS U1 à U5 (ou U1 à U3, + d'autres informations contenues en U4 et 5) - ces champs contenant des "user-defined informations". L'usage semble en passe d'être établi de mettre en M3 le DOI, donc mettre autre chose que des éléments d'identification universelle ou globale dans M1 et M2 ne serait pas cohérent.
DOI (itemFields.DOI)	DOI	M3	
ISBN (item.ISBN)	numéro ISBN du livre	SN	La balise SN est valable aussi bien pour l'ISSN que l'ISBN. Il est donc préférable, pour un livre, de garder SN pour l'ISBN (presque toujours présent) et de proposer en M2 par exemple "ISSN" et le numéro ISSN de la collection
Titre abrégé du livre (itemFields.shortTitle)	Titre abrégé du livre	T2 -- Titre abrégé du livre	Voir plus haut
url (item.url)	URL de l'article	UR	Il est possible de proposer plusieurs fois le champ UR (plutôt que de mettre toutes les adresses dans la même balise, ce qui est pourtant possible). Ce peut être utile pour les versions du texte en plusieurs langues. Pour les autres liens, voir plus bas

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	Y2 -- AAAA/MM/JJ/ Consultation à h;min;sec	Proposition d'adaptation, sachant que les "/" doivent rester présents dans ce champ. Le texte à la fin est libre, mais il faut vérifier que les ";" ne posent pas de problème d'interprétation par les logiciels
Dépôt (itemFields.repository) ou autorisation (itemFields.rights)	droits d'auteur et Licence		RIS ne gère directement ni droits d'auteur ni droits d'accès. La Champ AV (availability) est en fait un pointeur vers le texte complet. Restent les champs définis par les usages (M..., U...), mais ils sont déjà assez "occupés" par les autres informations utiles sur le document.
variable interne, numéro dans la base Zotero	numéro dans la base	ID	Pourrait servir, avec une table des correspondances, à pallier l'absence d'indication sur les droits (le champ compte jusqu'à 255 caractères, de préférence en capitales, ce qui est assez ample)
marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	KW KW KW KW	L'asterisque (caractère 42) n'est pas autorisé pour les champs mots-clés
notes attachées	contenu de la note 1 contenu de la note 2 …	N1 N1 N1	

Champ Zotero	Valeur du champ	balisage RIS	Commentaires
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	L1 -- Lien vers un PDF L2 -- Lien vers du plein-texte (en HTML par exemple) L3 -- Lien vers des documents en relation L4 -- Lien vers une image (par exemple couverture, image d'accroche, illustration)	AV permet de pointer vers un PDF ou Postscript, mais la gestion avec Pdfroot est un peu complexe. Il est préférable de décliner, en en respectant les différences, entre les lignes L1 à L4. Chaque balise peut être répétée.
		ER	Doit être la dernière balise du fichier

ZOTERO VERS ENDNOTE

Pour les articles

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
Item Type	Article de revue	%0 **Journal Article**	%0 **Journal Article**	Le type de référence implique des balises obligatoires. Les styles Endnotes font que si les deux version d'un article sont disponible il faut préférer la description "classique" avec mention d'adresse internet.

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
Titre (item.title)	Titre de l'article Sous-titre de l'article	%T **Titre de l'article Sous-titre de l'article** %Q **Titre de l'article traduit Sous-titre de l'article traduit**	%T **Titre de l'article Sous-titre de l'article** %Q **Titre de l'article traduit Sous-titre de l'article traduit**	Si le titre est suivi d'un sous-titre, il faut faire suivre le titre d'un point et d'un espace sinon le titre et sous-titre seront collés dans Zotero De façon générale il faut éviter dans les champs Endnote les signes diacritiques, qui sont remplacés à l'import par des "*", sauf ceux utilisés par la grammaire d'Endnote (par exemple "," dans les noms auteur). La séparation se fait plutôt par sauts de ligne. Endnote propose un champ spécifique pour les versions traduites des mentions (titre, auteur)
auteur (creatorTypes.author)	Nom, Prénom de l'auteur	%A **Nom, Prénom Nom, Prénom Nom, Prénom** %H **Nom traduit, Prénom traduit Nom traduit, Prénom traduit Nom traduit, Prénom traduit**	%A **Nom, Prénom Nom, Prénom Nom, Prénom** %H **Nom traduit, Prénom traduit Nom traduit, Prénom traduit Nom traduit, Prénom traduit**	La "," sépare le nom de famille du prénom; forme préférable à l'ordre prénom nom. Si l'auteur est une institution, il faut la faire suivre d'une virgule. Les traductions de noms d'auteurs contemporains seront sans doute très rares.
traducteur (creatorTypes.translator)	Nom, Prénom du traducteur	%A **Nom, Prénom**	%A **Nom, Prénom**	Endnote ne différencie pas traducteur et auteur (attention, %H = "Translated author" est le nom traduit de l'auteur, en tout cas dans l'usage tel que je l'ai constaté...). Il faut un seul champ %A !
éditeur (creatorTypes.editor)	Nom, Prénom du directeur de publication			Endnote ne prend pas en compte la mention de directeur de publication (= éditeur) pour ces 2 types de documents

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
collaborateur (creatorTypes.collaborator) ou auteur critique (creatorTypes.commenter)	Nom, prénom du collaborateur ou du commentateur			En article papier, Endnote ne reconnaît pas les auteurs secondaires ou tertiaires ! C'est regrettable d'autant plus que ce serait utile… Il peut être intéressant de les intégrer malgré cela, car les champs Endnote sont conservés si présents.
résumé (item.abstractNote)	résumé abstract resumen …	%X **Résumé abstract resumen**	%X **Résumé abstract resumen**	S'il y a des résumés en différentes langues, il suffit d'insérer un saut de ligne entre eux dans Endnote
publication (item.publicationTitle)	Nom de la revue	%J **Titre de la revue**	%B **Titre de la revue**	La mention de titre de publication peut se faire selon plusieurs champs dans Endnote, mais l'usage distingue les deux cas.
Volume (itemFields.volume)	volume de la revue	%V **Volume**	%V **Volume**	
numéro (item.issue)	numéro de la revue ou titre du numéro de la revue	%N **Numéro de revue** (pas de mention de titre de numéro possible)	%N **Numéro de revue** (pas de mention de titre de numéro possible)	les champs Zotero Titre de série et Texte de série peuvent être utilisés, mais l'usage ne semble pas bien défini à leur propos

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
pages (item.pages)	pagination	%P **Nombre de pages** *ET* %& **Page début**	%P **Nombre de pages**	les pages sont optionnelles en version électronique; à convertir en nombre de paragraphes (c'est une indication volumétrique non moins indicative que le nombre de pages) ?
date (item.date)	date de publication	(%D **Année** OU %8 **Date**) ET %7 **Date de publication électronique**	%D **Année** et %= **Date de dernière modification**	
Pas pour les articles	Lieu de publication	%C **Lieu de publication**	%C **Lieu de publication**	Pas géré par Endnote pour les articles papier, mais bien pour les articles électroniques : il est logique de générer systématiquement ce champ.

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
abréviation de journal (item.journalAbbreviation)	Titre abrégé de la revue	%! **Titre court**	%! **Titre court**	Le titre court n'est pris en charge par Endnote que pour les publications papier; le champ n'étant pas utilisé pour les articles en ligne, on peut néanmoins aussi le renseigner avec %! **Titre court**
langue (item.language)	nom de la langue	%G **Langue**	%G **Langue**	
DOI (itemFields.DOI)	DOI	%R **DOI**	%R **DOI**	
ISSN (item.ISSN)	numéro ISSN	%@ ISSN	%@ ISSN	
Titre abrégé (itemFields.shortTitle)	Titre abrégé de l'article			Réserver le titre court à la revue (très utilisé) et non à l'article (plus marginal) en format Endnote. Il est déconseillé d'utiliser le titre traduit comme titre court.
url (item.url)	URL de l'article	%U **URL**	%U **URL**	
accédé le (item.accessDate)	date de consultation : jour mois année heure:minute:seconde	%[**Date d'accès**	%8 **Date de dernier accès** (= date accessed) OU %[**Date d'accès (= access date)**	Date Accessed pour la publication en ligne, ne précise pas s'il s'agit du dernier accès auteur ou lecteur, ou de la dernière modification de notice. Access Date, optionnel pour tous les "sets" Endnote, a-t-il la même signification ?

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
Autorisations (item.rights)	droits d'accès	%1 **Notification des droits**	? Inclus dans les %Z **Notes** ?	Attention à la différence nette entre %1 pour article papier et %1 pour article électronique (année de citation)
Dépôt (itemFields.repository)	droits d'auteur	%W **Fournisseur de la base de données ?**	%W **Fournisseur de la base de données ?**	Endnote ne propose pas directement d'expression des droits d'auteur
marqueurs (item.tags)	mot-clef 1 mot-clef 2 keyword 1 keyword 2	%K **mot-clef 1** **mot-clef 2** **keyword 1** **keyword 2**	%K **mot-clef 1** **mot-clef 2** **keyword 1** **keyword 2**	les mots-clefs sont séparés par des sauts de ligne
notes attachées	contenu de la note 1 contenu de la note 2 …	%Z **Note 1** **Note 2**	%Z **Note 1** **Note 2**	Les notes sont séparées par des sauts de ligne (une seule prévue à l'origine) - limite de 64K par champ. Les "research notes" d'Endnote, introduites par "%<", sont à placer dans les notes normales "%Z"
Liens (pièces jointes)	Titre du lien 1 Titre du lien 2	%> fichiers joints	%> fichiers joints	Maximum de 45 fichiers joints. Les fichiers images sont traités différemment, avec l'insertion d'une légende à la suite, apparemment introduite par "%^", et le placement dans un répertoire DATA spécifique.

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
	Date de citation		%1 **Année de citation** OU %2 **Date de citation**	Pour l'instant, je ne sais pas s'il s'agit de la date de dernière citation, ou de la date de fabrication de la notice Endnote citant le document (plus logique et plus vraisemblable). En cas de doublon, toujours préférer la date, plus précise; mais parfois mieux vaut une année correcte qu'une date fausse. S'il est généré automatiquement ce contenu doit l'être en date, si c'est manuel il faut préconiser plutôt le champ année.
variable interne, numéro dans la base Zotero	numéro dans la base	%L **Numéro d'appel** *OU* %M **Numéro d'accès ?**	%M **Numéro d'accès**	Dans Endnote, le champ %F **Label** est parfois utilisé comme étiquette "manuelle" du document
	Genre de l'article	%9 **Type d'article**	%9 **Type d'article**	Utile pour distinguer les commentaires des travaux originaux
	Nom de la base de données	%~ **Nom de la base de données**	%~ **Nom de la base de données**	Peut servir de complément d'identification de la ressource. Ce champ rique d'être très utile dans le cadre d'une multiplication des portes d'accès à une ressource (par exemple, Revues.org via Cairn)
Elément connexe (non-signifiant dans Zotero)?	Référence commentée dans l'article	%* **Elément commenté**	%* **Elément commenté**	Champ très utile dans le cadre d'une revue. Il peut être rendu nécessaire par l'identification d'un champ "Type d'article" avec la valeur "review / commentaires". On peut préconiser de renseigner ce champ avec un identifiant pérenne plutôt que le titre. Fonctionnement également intéressant en guise de test pour un passage au web sémantique.
	Première édition	%(**Première édition**	%(**Première édition**	Même si la question de la version en ligne est moins posée pour des revues scientifiques que pour les

Champ Zotero	Valeur du champ	balisage Endnote papier + électronique	balisage Endnote électronique	Commentaires
	Numéro d'édition	%) **Edition**	%) **Edition**	archives ouvertes, étant donné qu'il y a de fait des modifications a posteriori il serait rigoureux de signaler ces modifications. Or au niveau des métadonnées il n'y a souvent pas de mention de ce type, celle-ci étant portée la plupart du temps dans le texte du document. Il serait intéressant tout au moins d'indiquer la date de dernière modification, sans forcément la numéroter (mention de première version pas nécessaire).

Index des standards