Métadonnées et XML
Des standards efficients de l'environnement numérique

MOREL-PAIR, Catherine
Institut de l'information scientifique et technique (INIST)

MOREL-PAIR, Catherine. Prépublication de « Métadonnées et XML : des standards efficients de l'environnement numérique ». *Ingénierie des systèmes d'information*, 2007, vol.12, n°2, p. 9-39. ISSN 1633-1311 ISBN 978-2-7462-1913-7
Disponible sur : <http://www.enssib.fr/bibliotheque-numerique/notice-1842>

Métadonnées et XML

Des standards efficients de l'environnement numérique

Catherine Morel-Pair

Service Edition Electronique – INIST-CNRS
2, Allée du Parc de Brabois
F-54541Vandoeuvre-lès-Nancy Cedex
catherine.morel@inist.fr

RÉSUMÉ. *Les métadonnées, ensembles de données structurées décrivant des ressources physiques ou numériques, sont un élément fondamental des pratiques de « curation digitale » du document électronique. Pour qu'elles remplissent leurs fonctions, l'utilisation de standards est incontournable, sur les éléments de métadonnées comme sur leur format d'implémentation. Le langage XML est à juste titre le format d'implémentation privilégié de nombreux standards de métadonnées. Après un panorama sur les fonctions des métadonnées puis sur XML et son intérêt, cet article décrit les principaux standards, quelques grandes applications, puis introduit RDF, outil XML d'une interopérabilité accrue entre applications*

ABSTRACT. Metadata, structured data sets about physical or digital resources, are a basic tool in digital curation. To assume their main functions, standard must be used, for metadata sets as for implementation formats. XML language is rightly the most important format for a lot of metadata standards. This document gives first a review of metadata functions and XML in context, describes main metadata sets and applications examples, then introduces a particular XML format, RDF, which will reach a real resources interoperability

MOTS-CLÉS : métadonnées, ressources électroniques, systèmes d'information, curation digitale, standards interopérabilité, Dublin Core, XML, RDF, applications, synthèse.

KEYWORDS : metadata, digital resources, information systems, digital curation, standards, Dublin Core, XML, RDF, interoperability, applications, review.

1. Introduction

Dans une société de l'information de plus en plus numérique, a émergé le concept de « curation digitale », traduction littérale de *digital curation*, ensemble de pratiques visant à « maintenir et ajouter de la valeur à un corpus fiable d'information électronique pour les utilisations actuelles et futures »[1], de la création des données à leur archivage en passant par leur diffusion. Ce processus repose sur un élément central, une « colonne vertébrale » (Higgins 2006); les métadonnées, qui sont « des ensembles de données structurées pour décrire, expliquer, localiser les ressource et en faciliter la recherche, l'usage et la gestion » (NISO, 2004-1). Pour aller vers la pérennité et l'interopérabilité des données, à l'intérieur d'un système ou entre systèmes, l'utilisation de standards s'impose, sur les éléments de métadonnées utilisés, sur leurs valeurs et sur leurs formats d'implémentation.

Il existe aujourd'hui quelques dizaines de standards de métadonnées, basés sur un modèle de données spécifique, et complémentaires dans leurs objectifs. Par rapport à d'autres formats d'implémentation, XML, eXtensible Mark-up Language, standard soutenu par le W3C, donne aux métadonnées leur puissance opérationnelle en matière de recherche, de gestion et d'interopérabilité . XML est d'abord un langage de balisage qui permet de créer des ressources logiquement structurées, pérennes, conformes à des modèles et donc facilement partageables ; et avec XML, les fichiers de métadonnées deviennent des ressources à part entière. C'est également un langage de programmation puissant pour l'ingénierie du document. XML est aujourd'hui le support incontournable des principaux jeux d'éléments, des grandes applications partagées, mais également de beaucoup d'applications plus locales.

Après un tour d'horizon de ce contexte, nous passerons en revue les grands standards, dont Dublin Core, et des projets les utilisant ; nous expliciterons l'intérêt de RDF pour l'interopérabilité, nous terminerons par quelques questions sur l'implémentation de métadonnées.

Cet état de l'art, qui a pour point de départ une optique documentation-bibliothèques, s'est construit en lien avec des professionnels d'autres métiers. Aujourd'hui, autour des ressources numériques, tous les acteurs du système d'information sont en interaction technique et décisionnelle, dans des métiers qui évoluent, avec des besoins de références communes, de la recherche d'information aux questions d'archivage, ou du processus de publication traditionnel à la gestion des documents de travail du chercheur ; il est certain que ces standards ont une place à tenir dans les diverses communautés.

---

[1] "maintaining and adding value to a trusted body of digital information for current and future use", Chris Rusbridge (DCC), in "The challenge of managing and preserving e-Research", http://www.apsr.edu.au/documents/rusbridge_NLA_talk.pdf

2. Des métadonnées, pourquoi, comment ?

2.1. *Des besoins nouveaux de description, gestion et interopérabilité des ressources*

Par rapport à l'objet d'information traditionnel, géographiquement situé, et conçu selon une chaîne de production et de diffusion spécifique, le développement rapide de l'information numérique et des techniques du Web modifient profondément l'économie et le cycle de vie du document. C'est d'abord une opportunité extraordinaire d'accès à une masse exponentielle de ressources multiformes et permettant des parcours de lecture divers, sur Internet ou à l'intérieur d'un système d'information ; les grands laboratoires scientifiques entre autres, et notamment en astronomie, physique des particules, bioinformatique ou médecine, disposent de masses de données énormes issues d'appareils de mesure et d'exploration variés, ou de simulation par ordinateur (Day, 2005). C'est aussi une opportunité d'évolution des pratiques de production et de mise à disposition de l'information : nouvelles technologies, intégration de ressources préexistantes, liens multiples entre ressources, activité collaborative d'écriture, plasticité du document dans le temps…

Cette évolution crée en parallèle de nouveaux besoins de traitement : comment aider des utilisateurs divers à retrouver l'information pertinente, voire évaluée, dans cet ensemble ? Comment faire évoluer les droits d'usage dans le respect des droits d'auteur et les gérer ; qui est responsable des ressources ? Comment garder la notion de contexte de production, d'intégration de l'objet numérique de base dans un ensemble cohérent situé dans le temps, tout en développant l'interopérabilité des ressources entre elles pour en créer d'autres dans de nouveaux contextes ? Comment garantir l'authenticité et assurer la traçabilité de versions successives ? Et comment rendre les documents pérennes en termes de conservation et d'accessibilité à long terme, alors que machines et logiciels sont eux-mêmes peu pérennes ?

Tout système d'information doit répondre à ces questions par de bonnes pratiques de curation digitale, dans lesquelles les métadonnées jouent un rôle important ; leurs fonctions peuvent être déclinées en six groupes principaux (Day 2005, Higgins 2006) :

– améliorer la recherche d'information et la « découverte » (*discovery*) des ressources avec des *métadonnées descriptives du contenu* : titre, résumé, mots-clés, classement … Cet objectif, premier dans l'histoire des métadonnées, peut inclure des éléments sur la qualité du document et les pratiques émergentes de folksonomie ;

– gérer les ressources, avec deux grands sous-ensembles de métadonnées :

- d'une part, des *métadonnées administratives*, portant sur la propriété intellectuelle, la responsabilité, les droits d'usage et les sources utilisées ;

- d'autre part, des *métadonnées instancielles*, ou *techniques et de structure*, regroupant les caractéristiques physiques ou informatiques : format, propriétés techniques détaillées de documents particuliers comme les images, dates significatives dans le cycle de vie, structure ou place dans une hiérarchie, logiciels de consultation…

– gérer les « archives », au sens du *Record management* (norme ISO 15489) ; les archives sont ici l'ensemble des documents utiles à court et moyen terme. Dans ce processus, il s'agit « d'identifier, authentifier, localiser, et contextualiser les données, ainsi que les personnes, les processus et les systèmes qui les créent, les gèrent ou les utilisent et les politiques qui les régissent » pour garantir la qualité, la fiabilité, l'accessibilité et la pérennité des ressources (Day 2005) ; la commission ISO/TC46/SC11 responsable de la norme travaille actuellement à un standard de métadonnées répondant aux besoins identifiés ;

– faciliter le partage de données et leur réutilisation ; cette fonction, importante pour limiter les coûts, passe par l'amélioration de l'interopérabilité au travers de standards, ainsi que par la présence d'informations contextuelles pour guider l'interprétation des données. Dans le cadre du Web sémantique, l'interopérabilité entre ensembles de données concerne d'abord les machines ;

– participer à la pérennité des ressources numériques, qui garantit que « l'essence du contenu est accessible pour toujours »[2] . Il y a là un enjeu fort. Par exemple, et ce n'est que la partie émergée de l'iceberg, 30 % des publications en ligne citées dans des articles de « D-Lib magazine » entre juillet 1995 et août 2004 sont inaccessibles en février 2005, pour 28 % en septembre 2004 (McCown 2005) ; 41 % des citations 1995-1999 de « Computer and Communications of the ACM » sont inaccessibles en 2003, dont 23 % dès 2000°(Jantz 2005). Les *métadonnées de préservation* décrivent « entre autres les actions réalisées en vue d'assurer la pérennité et l'accès pérenne, telles migrations ou contrôles d'intégrité des fichiers » (Day 2005);

– décrire les utilisateurs pour gérer les accès, leur permettre des personnalisations de consultation, analyser les comportements d'usage…

Cette dernière fonction ne sera pas traitée ici, où nous parlerons essentiellement des métadonnées pour les ressources : *métadonnées descriptives du contenu, métadonnées de gestion administratives et instancielles, et métadonnées de préservation.*

Le terme de métadonnées est utilisé depuis longtemps dans certains domaines d'activité comme la description des documents géographiques, la gestion des

[2] Chris Rusbridge (DCC), "The challenge of managing and preserving e-Research", http://www.apsr.edu.au/documents/rusbridge_NLA_talk.pdf

ressources images et multimédias ou les bases de données ; sous un autre vocabulaire, le concept est au cœur de certains métiers comme ceux des bibliothèques et de l'archivistique. Il concerne aujourd'hui tous les acteurs de l'environnement numérique.

Selon l'usage attendu, les éléments sont présents dans la ressource elle-même (documents HTML, métadonnées natives des images…) ou dans un fichier associé (bases de données, métadonnées des entrepôts des archives ouvertes, métadonnées images exportées …). Dans le premier cas, les métadonnées sont plus fortement liées à la ressource, mais leur utilisation pour la recherche d'information et la gestion est limitée ; la seconde possibilité est indispensable pour assurer ces fonctions.

2.2. Autour de l'interopérabilité

L'interopérabilité peut être définie fonctionnellement comme la capacité d'une ressource ou d'un service à s'intégrer dans un ensemble plus vaste ; plus techniquement c'est la « capacité d'échanger des données entre systèmes multiples disposant de différentes caractéristiques en terme de matériels, logiciels, structures de données et interfaces, et avec le minimum de perte d'information et de fonctionnalités » (NISO 2004-1)

L'interopérabilité des ressources met en jeu trois niveaux techniques complémentaires :

- une description des ressources avec des sémantiques communes issues de différents jeux de métadonnées standardisés,
- un contexte générique d'implémentation de ces descriptions dans des langages structurés standardisés, interprétables par les machines,
- des protocoles informatiques d'échange de ces données normalisées.

	Standards traditionnels	Standards récents
Jeux de métadonnées	MARC	Dublin Core MARC-XML, MODS EAD LOM…
Cadre générique d'implémentation	ISO 2709 ISAD(G)	XML RDF espaces de nom URL
Protocoles	WAIS FTP Z39.50	HTTP OAI-PMH SRU/SRW

Tableau 1. Niveaux d'interopérabilité et positionnement de divers standards.

Un quatrième niveau d'interopérabilité, plus organisationnel que technique, implique des pratiques communes dans l'utilisation des éléments et dans les valeurs de ces éléments (codes, vocabulaires, données d'autorité ou ontologies reconnus). L'importance de ce niveau a été mise en évidence plus récemment, après l'établissement d'un consensus relatif sur les standards. Il fait partie des spécifications de tout bon jeu de métadonnées. Ces spécifications sont formalisées dans des registres de métadonnées qui décrivent chaque élément (définition, hiérarchie, usage, valeurs attendues, parfois correspondances avec d'autres jeux…), et dans des « guidelines », qui apportent des recommandations de bonnes pratiques.

2.3 Des acteurs et des initiatives

Le Consortium W3C, World Wide Web Consortium, www.w3.org, instance internationale permanente chargée de l'avenir du Web composée surtout d'industriels de l'informatique et des TIC, travaille activement à l'interopérabilité du Web au travers de l'élaboration et de la diffusion des protocoles et des formats syntaxiques et structurels, de HTTP et HTML à RDF et au Web sémantique.

Les premières versions du langage HTML prévoient déjà des métadonnées dans l'en-tête du fichier, couvrant sommairement les trois premières grandes fonctions décrites ci-dessus (description du contenu, gestion administrative et technique). Cependant, ces éléments souffrent d'une standardisation insuffisante des usages et valeurs, et sont sous-utilisées, tant par les producteurs de sites que par les outils de recherche. De plus, ils ne peuvent décrire que des documents HTML et sont peu utilisables en ingénierie de l'information.

Le jeu d'élément Dublin Core a donc été créé en 1995 par le DCMI, Dublin Core Metadata Initiative[3], groupe international de professionnels du milieu documentation-bibliothèques et d'autres origines (musées, IETF[4]…) réuni à Dublin dans l'Ohio. Il correspond à une description « généraliste » s'appliquant à tout type de ressource, pour améliorer la recherche d'information dans le cadre d'une large utilisation. Le DCMI, devenu instance permanente, continue à travailler à l'adaptation de ce jeu aux différents contextes d'usage, au travers de groupes thématiques ouverts et d'un congrès/atelier de travail (*workshop*) annuel.

D'autres grands organismes ou regroupements d'organismes, souvent américains et leaders dans le traitement de l'information de leur domaine d'activité, ont élaboré en parallèle des jeux sémantiques plus spécifiques, répondant à des besoins particuliers de description et de gestion. Citons la FGDC, Federal Geographic Data Committee pour les métadonnées géospatiales, le RLG, Research Library Group, et

[3] Dublin Core Metadata Initiative, http://www.dublincore.org
[4] Internet Engineering Task Force, http://www.ietf.org/

l'OCLC, On Line Computer Library Center, pour les métadonnées liées au patrimoine culturel et à la préservation, l'Université de Bekerley pour l'archivistique et la fédération des grandes universités américaines dans les programmes « Digital Library Initiative »[5] ; l'IPTC, International Press and Telecommunication Council pour la standardisation des métadonnées images, la Library of Congress pour le développement d'ensembles à orientation bibliographique et la maintenance de nombreux autres jeux, les éditeurs commerciaux de livres et revues et les éditeurs d'outils pédagogiques.

Les organismes normalisateurs et standardisateurs jouent un rôle important dans le processus de développement et de diffusion. Au sein de l'ISO, l'Organisation internationale de normalisation, plusieurs comités techniques sont impliqués dans la normalisation des standards actuels. Le NISO, National Information Standards Organization « identifie, développe, maintient et publie des standards techniques pour gérer l'information dans un monde évolutif et de plus en plus numérique » en lien étroit avec le monde industriel, et s'intéresse fortement aux métadonnées (NISO 2004-1, NISO 2004-2).

D'autres organismes travaillent à un niveau plus global, autour de l'usage, de l'implémentation et de l'interopérabilité des métadonnées. Ainsi, l'OCLC est porteur d'une réflexion conceptuelle et développe des outils de type passerelles entre jeux[6]. Le DDC, Digital Curation Center[7], association d'universités du Royaume Uni soutenue par le JISC, Joint Information Systems Committee, a une activité de recherche, d'expertise et de conseil sur les bonnes pratiques de traitement du document numérique, d'abord destinée aux universités du pays, mais dont l'intérêt dépasse largement ce cadre.

Aujourd'hui, plusieurs dizaines de jeux sont reconnus comme des standards, normalisés ou en cours de normalisation, et largement utilisés dans le monde. Dans ce qui pourrait apparaître comme un foisonnement d'initiatives, la majorité de ces ensembles sont plutôt complémentaires dans leurs d'objectifs, pour l'objet décrit ou en terme de métier. Ces grands standards seront décrits en partie 4. Presque tous sont implémentés en XML, et les autres évoluent rapidement dans ce sens.

En France, malgré l'absence de projet global sur la communication scientifique, ces standards sont utilisés par les grands opérateurs comme l'ABES, la BNF, l'INIST (voir partie 5) et certaines unités. Ils sont enseignés dans les cursus pédagogiques. L'AFNOR, représentant de l'ISO pour la normalisation, participe à leur adaptation aux besoins et spécificités françaises. Elle est notamment impliquée dans le projet TEF, Thèses Electroniques Françaises avec d'autres partenaires

---

[5] Digital Library Initiative, http://www.dli2.nsf.gov/

[6] Voir Metadata Schema Transformation Services,
http://www.oclc.org/research/projects/mswitch/1_schematrans.htm
Et ERROLs : Extensible Repository Resource Locators
http://errol.oclc.org/schemaTrans.oclc.org.html

[7] http://www.dcc.ac.uk/

institutionnels. Divers travaux relaient, complètent ou adaptent les réalisations internationales : recherche et publications, liste de diffusion francophone Dublin Core[8], forum métadonnées sur la plateforme ARTIST[9], billets de blogs professionnels[10]...

3. XML, un langage puissant et reconnu

3.1 XML pour des documents structurés et pérennes

Issu de SGML, premier langage de balisage structuré, XML permet de structurer les documents de manière logique et arborescente ; le document XML est un ensemble de nœuds imbriqués correspondant à des éléments signifiants sur le plan structurel et sémantique. C'est uniquement une structuration du contenu, sans propriétés de mise en page ; diverses présentations du même document pourront être générées ensuite. Le rôle de chaque élément, le type de la valeur, les liens entre éléments sont précisés par des attributs.

XML est un standard ouvert qui s'est imposé dans l'échange de documents. Intégralement basé texte, il est indépendant des formats des fichiers binaires ou des systèmes d'exploitation[11]. Il s'associe à n'importe quel jeu de caractères, notamment Unicode. Il est donc pérenne pour le stockage de documents à long terme et interopérable. Il est de plus modulaire et extensible : c'est un métalangage dont les bases peuvent être utilisées pour créer d'autres langages reposant sur les règles du XML.

Par exemple, le mécanisme XLink relie un élément à un autre fichier, avec une standardisation fine de la sémantique du lien, et XPointer permet de plus de désigner des nœuds spécifiques dans le fichier cible ; les objets XML complexes constitués ainsi sont donc précisément structurés. Citons également la technologie XForm, dédiée aux formulaires, SVG, outil de description vectoriel de ressources graphiques ou SMIL pour les ressources multimedia.

3.2 Modèles de documents : DTD et schémas

Pour favoriser l'interopérabilité, XML propose l'implémentation de modèles de documents décrivant la structure, les éléments et attributs, voire les valeurs autorisées, et le parseur XML effectue la validation de chaque instance de document

---

[8] Liste DCMI francophone, http://listserv.inist.fr/wwsympa.fcgi/info/dcmi-fr

[9] Appropriation par la Recherche des Techniques de l'IST, http://artist.inist.fr , voir « Espace groupes de travail »

[10] Par exemple www.figoblog.org/ et www.lespetitescases.net/

[11] L'édition du fichier est indépendante d'un logiciel particulier, mais les éditeurs XML permettent seuls d'utiliser les fonctions évoluées du processeur/parseur XML

par rapport au modèle déclaré. Ces modèles sont de deux types : la DTD, Document Type Description et le schéma XML.

La DTD, issue du monde SGML, a été le modèle le plus utilisé au départ. Sa syntaxe est relativement simple et ramassée, mais sa puissance reste limitée.

Le schéma XML, de syntaxe plus complexe, présente de nombreux avantages :

- c'est un document XML ;

- il permet d'imposer des contraintes plus précises sur l'occurrence des éléments et sur le modèle des valeurs attendues, des expressions régulières par exemple ;

- il permet de déclarer des espaces de nom, ensembles d'éléments fermés décrits de manière formelle, par exemple jeux d'éléments standards ou ensembles de valeurs. La déclaration de chaque espace de nom inclut son adresse et un préfixe qui sera précisé devant les éléments correspondants ; ceux-ci restent donc toujours reliés à leur origine ;

- il autorise l'import de schémas XML externes et leur adaptation à l'application locale par divers mécanismes ; selon la priorité donnée au schéma distant, on utilise la syntaxe « xs:import » ou « xs:include ».

Le schéma XML permet ainsi de créer des « profils d'application » spécifiques, adaptés aux applications locales tout en respectant une interopérabilité de base ; la majorité des projets métadonnées sont basés sur le schéma standard le plus adapté à un objectif, mais en modifient les contraintes et y ajoutent des éléments issus d'autres espaces de nom ou locaux, comme des éléments de gestion sans intérêt pour l'échange ; cette pratique a fait l'objet d'exposés pédagogiques pendant les derniers *workshops* Dublin Core.

XML Schéma est maintenu par le Consortium W3C ; d'autres syntaxes de schémas plus puissantes existent, notamment Schematron et Relax NG.

3.3 XML et usages du document

XPath, XSL, DOM, XQuery sont des briques de fonctions XML créées par le Consortium W3C pour le traitement et l'utilisation intégrés du document XML. Standards proposés par le W3C, ils sont normalisés ou en cours de normalisation et parfois doublés par des mécanismes créés par d'autres acteurs.

3.3.1 XPath

XPath est un premier langage de requête qui permet de désigner des noeuds de l'arbre en fonction de leur position, du nom et contenu des éléments, des attributs et valeurs d'attribut. C'est un module de base utilisé par tous les autres.

3.3.2 XSL, XSLT

Le processus de transformation XSLT fait appel à des feuilles de style XSL qui s'appliquent aux documents XML pour créer un résultat, soit à la volée et lisible par un navigateur, soit en dur, cas le plus fréquent pour les métadonnées. Les résultats

sont variés : fichiers de diffusion, notamment pages HTML et documents PDF ; nouvelles ressources issues de l'extraction de parties du document ou de la concaténation de plusieurs fichiers (ou parties de fichiers) ; changement des noms des éléments et attributs pour correspondre à un autre modèle de document… Cette dernière transformation est extrêmement utilisée dans les applications métadonnées, où le principe de décrire une fois pour réutiliser plusieurs fois au cours du temps, dans des projets d'échanges ou d'évolution interne, est particulièrement pertinent.

XSL/XSLT constitue un véritable langage de programmation dans la syntaxe XML ; il comprend des tests et des boucles imbriqués, des compteurs et variables, et une manipulation évoluée des chaînes de caractères. Des processeurs XSLT sont intégrés aux les éditeurs XML et aux divers environnements informatiques ; les programmes peuvent leur passer des paramètres et des variables, notamment pour traiter des lots de fichiers ou concaténer des données.

3.3.4 Traitement DOM

DOM, *Document Object Model*, correspond à une conception orientée objet de l'arbre-document et de ses nœuds. Le parseur DOM, composé de bibliothèques de fonctions, permet de créer et transformer les documents XML directement au travers d'un langage de programmation.

Les fonctions DOM et XSLT sont proches, et le choix de l'un ou l'autre dépend surtout de l'environnement informatique et de la formation de celui qui implémente

3.3.5 XQuery

Comme les langages de requête des SGBD, XQuery permet des requêtes sur un ensemble de documents, portant sur les éléments, les attributs et les valeurs. Suite à la longue durée de la standardisation, des alternatives orientées bases de données ont été développées : XML-QL, XQL ou SQL/XML, standard ISO 2003.

3.4 XML aujourd'hui

3.4.1 Quelques standards

Le site coverpages .org d'OASIS recense plusieurs centaines d'applications XML très diverses ; gros consortium international pour le développement, la convergence et l'adoption de standards pour le e-business envisagé de manière large, OASIS est fortement engagé sur deux axes, la standardisation des *web services* et de XML[12].

[12] OASIS, Organization for the Advancement of Structured Information Standards, http://www.oasis.org et ses deux sites techniques très complets : http://www.xml.org , « applying XML and Web services standards in industry » http://xml.coverpages.org/, « online resources for markup technologies ».

Cette liste cite entre autres des DTD comme TEI et Erudit[13], qui proposent un balisage très fin de tous les éléments signifiants d'un document, pour des utilisations ultérieures : recherche scientifique pour TEI, édition de revues en sciences humaines et sociales pour Erudit. MATH-ML[14] et OpenMath formalisent en XML les objets mathématiques des documents. DocBook[15] structure finement des documents techniques industriels, pour permettre la publication de différentes documentations dérivées adaptées à chaque contexte d'utilisation. Dans le domaine commercial, ebXML[16], standard développé par OASIS, est la base de transactions standardisées et sécurisées, et les documents ebXML ont valeur de preuve.

Les métadonnées correspondent à des documents régulièrement structurés, et se prêtent facilement à une modélisation XML, sous forme de schéma pour la plupart des standards, de Dublin Core à RDF.

3.4.2 XML et environnement informatique

L'environnement bases de données propose deux types de produits pour XML. D'une part, des SGBD compatibles XML permettent d'importer des données XML dans les différentes *tables* et de les manipuler ; elles conviennent pour des données régulièrement structurées dont l'ordre dans le document est sans importance après l'import. D'autre part, pour des documents plus complexes et variables en structure et contenus, dans lesquels les éléments ont un lien sémantique et structurel, il existe des bases de données en XML natives, qui reposent sur le traitement des documents natifs au travers des fonctionnalités XML[17].

En bureautique, Open Office a été le premier à s'intéresser à XML, et chaque document est un ensemble de fichiers XML, dont l'un pour le contenu textuel et un pour les métadonnées. Le projet Office Open XML de Microsoft, lancé en 2005, est en cours de normalisation auprès d'Ecma International puis d'approbation par l'ISO. Il est soutenu par de très grosses entreprises (Apple, Barclays Capital, BP, the British Library, Essilor, Intel Corporation, Microsoft, NextPage Inc, Statoil ASA et Toshiba) qui y voient une garantie d'interopérabilité et d'archivage pérenne. Outre un convertisseur de documents déjà disponible chez SourceForge en 2006, la technique est maintenant intégrée dans la suite Office 2007[18].

---

[13] TEI : Text Encoding Initiative, http://www.tei-c.org, Erudit, http://www.erudit.org/

[14] Math-ML, http://www.w3.org/Math/

[15] DocBook, http://www.docbook.org/

[16] www.ebxml.org

[17] XML and Databases, Bourret R, mise à jour de septembre 2005, http://www.rpbourret.com/xml/XMLAndDatabases.htm

[18] Voir par exemple http://www.microsoft.com/france/cp/2006/1/info.asp?mar=/france/cp/2006/1/2006010501_a101.html&xmlpath=/france/cp/2006/xml/1.xml&rang=1

Microsoft intègre également dans sa machine virtuelle .NET Framework, parmi tous les langages de programmation disponibles, de nombreuses bibliothèques de classes orientées XML.

L'acronyme AJaX souvent associé aux techniques du Web2, un Web plus interactif, signifie Asynchronous Javascript XML.

Enfin, XML peut aussi être utilisé pour *paramétrer* des applications. Par exemple, l'outil d'exploration, de recherche et de visualisation SERV'IST développé par le service Veille de l'INIST, programmé en C et PHP, repose sur des fichiers de configuration XML/XSL pour définir les caractéristiques de chaque serveur : choix et création des différent index de recherche, onglets de navigation, structures des pages de résultats…

Dans ce contexte, l'implémentation de métadonnées en XML est un choix raisonné pour l'interopérabilité et la valeur ajoutée, à court et long terme.

4 Grands standards de métadonnées en XML

4.1 Jeux orientés « description du contenu »

4.1.1 Dublin Core

Les premiers travaux du DCMI en 1995 ont abouti rapidement à la création d'un jeu de base de 15 éléments, normalisé ultérieurement par l'ISO sous le numéro 15836-2003, et comprenant :

- *des éléments descriptifs du contenu* : title, description, subject, coverage, language ;

- *des éléments administratifs* : creator, contributor, publisher, source, rights ;

- *des éléments instanciels* : format, date, relation, identifier.

Ce module de base, nommé Dublin Core simple, a été enrichi par :

- une cinquantaine de qualificatifs (*refinements)* précisant les éléments,

- de nouveaux éléments, créés par le DCMI ou issus d'autres jeux, l'ensemble formant le Dublin Core étendu,

- des « schémas d'encodage », soit syntaxiques (modèles d'implémentation des éléments en HTML, XML et RDF), soit sémantiques (standards recommandés pour les valeurs).

Le site officiel dublincore.org a édité un guide d'usage détaillé (Hillmann 2005), renvoyant à d'autres documents plus techniques, registre et spécifications d'implémentation.[19]

Dublin Core est un cœur d'interopérabilité sémantique qui fait aujourd'hui l'objet d'un large consensus ; il présente aussi des limites, notamment sur la gestion de collections de ressources, les métadonnées techniques, administratives et de préservation. Il est ainsi utilisé par de multiples applications, en Dublin Core simple ou qualifié, très souvent dans des profils d'application spécifiques.

Les autres jeux orientés vers la description du contenu lui sont complémentaires, avec des objectifs plus spécifiques.

4.1.2 Description de domaine de connaissance ou d'action spécifiques

CSDGM, Content Standard for Digital Geospatial Metadata, standard XML publié par la FGDC,[20] est très pratiqué pour la description détaillée des ressources géographiques, géospatiales et socio-démographiques, en recherche scientifique comme par diverses administrations. Il possède des éléments décrivant la qualité des ressources ainsi que des extensions adaptées aux domaines biologie et environnement.

IEEE-LOM [21], Learning Object Metadata décrit les outils éducatifs, notamment logiciels de e-learning. Ce jeu d'éléments très utilisé est volontiers donné en exemple pour la génération automatisée de métadonnées : d'une part, elles sont souvent issues directement du cahier des charges XML du produit, et d'autre part, l'utilisation de la ressource génère de nouvelles métadonnées sur le processus d'apprentissage (comportement de l'utilisateur, enregistrement des résultats de tests…). Par ailleurs, LOM prépare un schéma d'implémentation en RDF et l'AFNOR a normalisé en 2005 une version française de LOM[22].

IMS-Metadata[23] est un autre jeu de description des ressources éducatives. SCORM, Sharable Content Object Reference Model[24], une spécification permettant de créer des objets pédagogiques structurés, interopérables, durables et réutilisables, intègre ces deux jeux.

---

[19]- Voir aussi, en Français mais orienté XHTML, Jacquet C., Dublin Core et les métadonnées, http://www.openweb.eu.org/articles/dublin_core/

[20] Federal Geographic Data Committee, http://www.fgdc.gov/metadata/metadata.html,

[21] IEEE-LOM, Learning Objects Metadata, Institute of Electrical and Electronics Engineers, http://ltsc.ieee.org/wg12/

[22] LOMFR, http://www.educnet.education.fr/articles/lom-fr.htm et http://www.boutique.afnor.fr/NEL5DetailNormeEnLigne.aspx?CLE_ART=FA139935&nivCtx=NELZNELZ1A10A101A107&ts=9958564

[23] IMS Learning Metadata, http://www.imsglobal.org/metadata/

[24] SCORM, http://www.adlnet.gov/scorm/index.cfm

D'autres standards émergents de métadonnées décrivant des corpus de données scientifiques font partie de cette catégorie (Day 2005), ainsi que le projet FOAF, Friend of a Friend[25], décrivant en RDF les personnes, leurs activités et leurs réseaux sociaux.

4.1.3. Jeux de métadonnées orientés métiers

4.1.3.1 MARC-XML et MODS[26]

MARC et ses dérivés, dont UNIMARC, correspondent au format de référence de la description bibliographique. Nés dans les années 1960 pour normaliser la description des documents et en faciliter l'exploitation et l'échange par les ordinateurs, maintenus par la Library of Congress, ils constituent des jeux de métadonnées essentiellement descriptives, permettant un signalement extrêmement précis et contenant des valeurs standardisées ; cependant, ces formats sont surtout pratiqués par les professionnels des bibliothèques et centres de documentation car ils sont complexes, peu lisibles par l'homme et déclinés en de multiples versions nationales ou locales.

Leur implémentation en XML améliore la compatibilité des différentes versions, avec la possibilité de passer plus facilement de l'une à l'autre grâce aux feuilles de style XSL. Le jeu d'éléments MADS, également en XML, formalise les données d'autorité sur les personnes et collectivités.

MODS, né en 2003 à l'initiative de la Library of Congress, est directement issu de MARC, et adapté à l'environnement actuel : schéma XML, codage Unicode, noms des éléments explicites et proches de Dublin Core, éléments adaptés à la description des ressources numériques …

BiblioML est une représentation XML du format bibliographique UNIMARC, développée en France par le ministère de la Culture et la société de services AJLSM[27].

4.1.3.2 EAD, Encoding Archive Description et l'archivistique

La DTD EAD[28] fait l'objet d'un large consensus dans les milieux archivistiques du monde entier. Publiée par l'Université de Bekerley (Californie) en 1993, elle est maintenue par la Library of Congress. EAD formalise en XML la norme ISAD(G) sur les « instruments de recherche » (catalogues) archivistiques. Les principes bien codifiés de la description archivistique traditionnelle ont été appliqués pour EAD :

---

[25] The Friend of a Friend (FOAF) Project, http://www.foaf-project.org/

[26] MARC : Machine Readable Cataloging, MODS, Metadata Object Description Schema, voir Libray of Congress, standards, http://www.loc.gov/standards/

[27] BiblioML, http://www.biblioml.org/fr/index.html ; AJLSM est une société d'études, de développement et de formation autour de l'information numérique

[28] Voir http://www.loc.gov/standards/
et Sibille C., la DTD EAD (Encoded Archival Description), avril 2004, http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html

respect des fonds et du contexte de production, grand nombre de niveaux hiérarchiques, description allant du général au particulier avec héritage des niveaux supérieurs sans redondance entre eux (Dhérent 2003).

Le format EAC, Encoding Archives Context, également XML, a été créé un peu plus tard pour les données d'autorité.

4.1.3.3 ONIX et l'édition commerciale

Ce standard XML de métadonnées a été proposé en 1999 par le groupe EDItEUR[29] pour favoriser le commerce électronique du livre et des séries ; il permet aux acteurs des métiers du livre de partager l'information sur le produit tout le long de la chaîne de fabrication et de commercialisation. Il comprend entre autres une description bibliographique précise et de l'information sur la « vie sociale » du livre. ONIX est effectivement utilisé par toutes les librairies en ligne, et, en France par le catalogue des libraires, Electre ; EDItEUR projette une adoption plus large, par exemple par les bibliothèques. Il publie pour l'instant des feuilles de style permettant le passage d'ONIX à MARC ou Dublin Core.

4.1.4 Métadonnées liées à des formats de documents

4.1.4.1 En-têtes de documents XML

DocBook, Erudit ou TEI, par exemple, comprennent chacun un ensemble de métadonnées essentiellement descriptives du contenu dans un en-tête. Idéalement, ce *header* est généré au moment de la création et de la publication du document, à partir des éléments balisés ; il peut ensuite être facilement converti vers des formats de description externes.

4.1.4.2 Métadonnées des formats images

L'IPTC, International Press and Telecommunications Council, est une organisation internationale créée en 1965 pour développer et promouvoir des standards d'échange de données à destination de la presse. En association avec la NAA (Newspaper Association of America), l'IPTC a défini un modèle global de données, IPTC-NAA Information Interchange Model, dont un sous-ensemble a servi de base à la société Adobe pour définir dans *Photoshop* les informations associées à une image, essentiellement des éléments descriptifs du contenu et des métadonnées administratives. Ce sous-ensemble est communément appelé *métadonnées IPTC*. Il comprend des informations stockées à l'intérieur des fichiers images JPEG, TIFF ou PSD (Photoshop), et ce standard dépasse donc largement le monde de la presse.

D'autres propriétés images sont très techniques (standard EXIF, voir 4.2.1), et également stockées dans les fichiers images. Les métadonnées IPTC et EXIF y sont encodées de façon propriétaire et avec des dérives par rapport aux standards de base, par exemple sur les noms d'éléments ; certaines métadonnées sont ainsi volontiers perdues entre les différents logiciels de gestion d'images.

---

[29] EDItEUR, http://www.editeur.org/

Adobe-Photoshop implémente aujourd'hui un format plus générique, XMP, eXtensible Metadata Packet, où des *packets* RDF encapsulent les éléments Dublin Core (étendus par Adobe pour la description des droits), les éléments Photoshop et Acrobat (PDF), ainsi que tout autre jeu, en particulier EXIF. La description est toujours incluse dans le fichier image, mais *XMP Packet* permet d'accéder aux métadonnées en lecture et écriture même en l'absence de logiciel spécifique ; elle est pérenne, accessible et directement utilisable en export.

Sous l'influence de XMP, le standard IPTC a évolué en 2005 vers IPTC-Core, dont la structure est basée sur le principe des *packets* XMP encapsulant des éléments IPTC qui ont légèrement évolué. Ce standard XML remplacera probablement les précédents dans les différents formats et outils image.

4.1.4.3 MPEG-7 et le multimédia

Les métadonnées liées à ces applications sont aussi diverses aujourd'hui que les applications elles-mêmes avec un inconvénient certain : faute d'une description opérationnelle standardisée, il est difficile de faire une recherche fine et multi-bases sur les ressources multimédias. Les formats MPEG récents doivent résoudre ce problème.

Les premiers formats MPEG développés par le *Moving Pictures Expert Group* sont essentiellement des formats de compression[30]. Avec MPEG-4, apparaît la possibilité de décrire le contenu. MPEG-7, norme ISO/IEC 2001, permet de *décrire* de façon standardisée et en XML une ressource audiovisuelle, avec un haut niveau sémantique sur le contenu visuel et sonore, ce qui doit améliorer l'efficacité des recherches. Il comprend également des métadonnées techniques et administratives[31].

4.1.6 Windows et les métadonnées descriptives du contenu

Microsoft a introduit avec Windows 2000 la possibilité d'associer et de renseigner des propriétés *Titre*, *Auteur*, *Mots-clés*, *Catégorie*, *Source*, *etc.*, à un fichier quelconque, même non Microsoft Office, sur les machines disposant de systèmes de fichiers NTFS, au travers des *alternate streams*. Ce système reste malheureusement éminemment propriétaire et non interopérable, même entre plateformes Windows différentes, et ces métadonnées sont volontiers perdues lors de la manipulation des fichiers hors de leur contexte de création (Peccatte 2006).

4.2 Métadonnées techniques et de structure : exemple des formats images

Développé en 1995 par le JEIDA, Japan Electronic Industry Development Association, EXIF, EXchangeable Image File, décrit de façon extrêmement précise les paramètres de prise de vue et les réglages de l'appareil au moment de la capture

---

[30] Les normes MPEG, Collectif, http://forum.hardware.fr/hfr/VideoSon/Traitement-Video/unique-mpeg4-mpeg21-sujet_54957_1.htm

[31] Gwinner C. L., Lalaurette S., Le standard MPEG-7 http://membres.lycos.fr/psebcopathe/mpeg/Le%20standard%20MPEG7.pdf

numérique. Ces métadonnées interne à la ressource et intégrées de façons diverses dans les différents formats images, sont encapsulables aujourd'hui dans les *packets* XMP.

MIX, Metadata Image XML, a été développé par le NISO en lien avec des experts du monde industriel, pour permettre de gérer durablement et de façon non propriétaire des collections d'images, dans un but d'échange et d'archivage pérenne ; il est maintenu conjointement avec la Library of Congress[32]. Il contient une description fine des caractéristiques techniques de l'image proche des éléments EXIF, mais dans un fichier externe ; les documents MIX contiennent logiquement les propriétés EXIF extraites des différentes images du corpus.

4.3 Métadonnées administratives

Tous les ensembles de métadonnées comprennent un minimum de métadonnées administratives. Celles-ci peuvent être simplement déclaratives, ou gérer les droits utilisateurs, dans le cadre du DRM, Digital Rights Management. Quelques jeux spécifiques peuvent être cités.

L'initiative Creative Commons[33] représente une alternative au *copyright* strict en proposant un ensemble de licences qui déclinent des droits d'usage diversifiés. Lancée par un éminent professeur de droit de Harvard (Etas-Unis), elle permet de combiner partage et protection des œuvres de l'esprit, et concerne plutôt les œuvres artistiques au départ. Elle a été reconnue et adaptée au droit national par des pays divers, dont la France en 2005. Le site Creative Commons permet de choisir une licence au travers d'un formulaire, de faire pointer la ressource sur le texte correspondant et d'insérer cette licence en RDF dans des objets de formats variés. Des outils comme Google et Yahoo proposent ce critère de recherche. Aujourd'hui, ces licences sont utilisées par certaines archives ouvertes et bases de données scientifiquest[34].. La confidentialité et le partage d'œuvres scientifiques répondent cependant souvent à des contraintes particulières, et le projet Science Commons, issu de Creative Commons, a pour mission d'en étudier les caractéristiques pour définir des licences adaptées.

VCard[35], standard de signature électronique soutenu par le consortium W3C pour authentifier les documents et transactions commerciaux, est d'ores et déjà un format employé par les outils de messagerie de manière transparente.

Différents projets de DRM, en majorité propriétaires, coexistent. XrML, eXtensible Rights Markup Language, a été élaboré entre 2000 et 2003 par ContentGard, société spécialisée dans le domaine des droits pour le numérique. Il a été adopté rapidement par le groupement MPEG comme infrastructure de référence

---

[32] MIX, http://www.loc.gov/standards/MIX/
[33] http://creativecommons.org
[34] Par exemple, la base de données factuelle UniProt, http://www.pir.uniprot.org/terms
[35] http://www.w3.org/TR/2001/NOTE-vcard-rdf-20010222/

pour son propre système DRM et par Microsoft dans ses outils multimédia, et des organismes comme OASIS l'intègrent dans leurs recommandations ou spécifications techniques[36]. ODLR, Open Digital Rights Language[37], diffusé sous licence Open Source, apparaît aujourd'hui comme la principale alternative ; la technologie ODLR a été adoptée par plusieurs fournisseurs de solutions serveur et elle est recommandée par le Consortium W3C.

Le jeu de métadonnées ONIX repose sur un cadre qui facilite le transfert des métadonnées commerciales et juridiques dans le contexte du commerce électronique, et appartient donc également à cette catégorie.

4.4 Métadonnées de conservation

L'OAIS, Open Archival Information System[38], norme internationale ISO 14721, est un modèle de gestion de données destiné à garantir leur pérennité. Il définit entre autres trois *packages* d'information contenant les objets numériques et leurs métadonnées : les *packages* de soumission, archivage et diffusion.

Le jeu d'éléments PREMIS, également XML, permet d'implémenter ces *packages* d'information ; il est fondamental pour le package d'archivage. Publié en 2005 par le RLG et l'OCLC, il est maintenu par la Library of Congress[39]. Bien des applications de tailles diverses suivent le modèle OAIS, parmi lesquelles de grands projets souvent cités comme l'Australian Phototheque, le projet CEDARS en Grande Bretagne (Day 2005) et la Bibliothèque Nationale de France (Dhérent 2005).

4.5 Méta-formats

Certains jeux de métadonnées décrivent des collections de ressources comme des objets numériques complexes, Ils remplissent l'ensemble des fonctions décrites ci-dessus et intègrent d'autres jeux de métadonnées pour décrire l'objet numérique de base. Il est donc tentant de les qualifier de « méta-formats ».

Le plus connu est METS, Metadata Encoding and Transmission Standard[40], issu du projet MOA, Making of America, qui rassemble des universités américaines autour de l'Université de Cornell autour de l'American Memory, grande bibliothèque numérique de sciences sociales sur une période historique spécifique. Egalement maintenu par la Library of Congress, METS permet de créer des objets numériques complexes, très structurés et dynamiques, « rassemblant » des objets numériques plus simples, quelconques et éventuellement dispersés. Constitué de

---

[36] http://www.xrml.org/; voir aussi Crochet-Damais A., Gestion des droits numériques, les standards se précisent, http://solutions.journaldunet.com/0401/040108_drm.shtml

[37] ODRL, http://www.odrl.net/ et http://www.w3.org/TR/odrl/

[38] OAIS, http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[39] PREservation Metadata Implementation Strategies, http://www.loc.gov/standards/premis/

[40] METS, Metadata for Encoding and Transmission Standards, http://www.loc.gov/standards/

sept sections (carte de structure, section des métadonnées descriptives, administratives, section des fichiers, de leurs comportements …) liées entre elles et au travers de leurs éléments-fils, il permet de gérer et diffuser la collection de multiples façons. Pour décrire les objets de base, METS intègre des éléments de tous les jeux de métadonnées XML ou pointe sur des descriptions existantes ; il comprend un seul ensemble sémantique spécifique portant sur les droits d'usage. METS permet enfin de créer des *packages* conformes au modèle OAIS. Ce format efficace mais assez complexe est de plus en plus utilisé[41].

l'IMS Global learning Consortium propose un autre méta-format, IMS-CP, pour IMS-Content Package[42], probablement plus orienté vers les collections virtuelles d'outils pédagogiques.

.MPEG-21 est un cadre de représentation des objets multimédias basé sur un ensemble de modules XML indépendants et externes à la ressource (avec une partie spécifique en RDF). Normalisé par l'ISO (ISO/IEC 21000), il a pour objectif de donner un bon niveau d'interopérabilité contrôlée pour la diffusion de documents multimédias, notamment dans le cadre de transactions commerciales. La notion de *Digital Item*, niveau de granularité préférentiel, est au centre de la norme ; cet élément regroupe les ressources, les métadonnées et une structure. Les métadonnées sont descriptives du contenu, techniques et administratives, et la structure permet également d'implémenter les packages OAIS (NISO 2004-2).

4.6 De l'interopérabilité …

Bien que tous ces standards de métadonnées soient à l'évidence utiles dans les différents contextes de création et de gestion de documents numériques, leur multiplicité pose un problème théorique et pratique en termes d'interopérabilité et de partage de données.

Certains jeux « transversaux » ont explicitement cet objectif, comme Dublin Core et METS, ou XMP et MPEG-21 dans l'avenir. D'autres sont issus de domaines où le partage de données est un élément prépondérant, comme les géosciences, la communauté éducative, et visent également ce but, mais surtout au sein de la communauté concernée (Day 2005).

Pour des utilisations dans d'autres contextes, il faudra souvent passer d'un jeu de métadonnées à un autre, en utilisant les mécanismes de transformation XML. La question est tout à fait maîtrisable sur le plan technique, et des outils de plus en plus

---

[41] Pour des exemples pédagogiques en France, voir
- un document édité par l'Ecole des Chartes :
. le fichier METS, http://lespetitescases.net/morchesnemets
. son affichage pour navigation http://elec.enc.sorbonne.fr/morchesne
- des exemples de fichiers TEF, Thèses électroniques Françaises, http://www.abes.fr/abes/DesktopDefault.aspx?tabindex=4&tabid=403, exemples

[42] IMS Global Learning Consortium, http://www.imsglobal.org/

nombreux sont proposés sur Internet : pages de l'OCLC déjà citées, pages de la Library of Congress[43], boîtes à outils EAD[44], site d'EDItEURS...

Elle est plus complexe sur le plan conceptuel, car ce passage implique des choix, sur l'établissement d'équivalence entre des éléments spécifiques et des éléments plus génériques (le contraire étant encore plus difficile), ou sur les valeurs des éléments quand les pratiques diffèrent d'une communauté à une autre.

Dans le cadre du Web sémantique et de RDF, la question conceptuelle persiste, qui implique de développer, par exemple, de grandes ontologies sur les principaux jeux de métadonnées, pour permettre aux requêtes de cheminer automatiquement entre diverses représentations des ressources.

5. Métadonnées en XML, des applications

5.1 Les archives ouvertes et le protocole OAI-PMH

Le mouvement du libre accès à l'information scientifique et technique a pour objectif de permettre à la communauté scientifique un accès plus large et plus rapide aux résultats de la recherche. Parti d'initiatives isolées au début des années 1990, dont celle de Paul Ginsparg et de la « base de preprints » arXiv ou celle de Stevan Harnad avec CogPrints, ce mouvement a pris de l'ampleur à travers la mobilisation de groupes de chercheurs et les prises de position nationales et internationales en sa faveur.

Ce principe suppose un modèle économique et social alternatif, où la communauté scientifique se réapproprie le processus éditorial, avec des modalités techniques et juridiques nouvelles. Même si l'évolution est plus lente que ne l'espéraient les pionniers, le libre accès a pris sa place aux côtés de l'édition traditionnelle : le DOAJ, Directory of Open Journals[45] recense pratiquement 2500 revues en accès libre ; il y a plus de 1200 archives ouvertes[46], et un des moteurs de recherche les plus connus, OAIster[47], indexe environ 9,8 millions de documents issus de plus de 700 archives scientifiques. L'enquête SHERPA-ROMEO[48] classe les éditeurs commerciaux selon la possibilité pour « leurs » auteurs de publier en

---

[43] Dans les pages MARC, MODS, EAD répertoriées ci-dessus
Et pour des feuilles de style METS vers IMS-CP, voir http://iu.berkeley.edu/creatingcontent/
[44] Chez Archivistics, http://www.archiviststoolkit.org/, ou dans le logiciel français PLEADE, éditée par AJLSM, ; pour les outils de transformation seuls, voir http://projets.ajlsm.com/ead-chan/

[45] DOAJ, http://www.doaj.org/
[46] Voir http://gita.grainger.uiuc.edu/registry
[47] OAIster, http://oaister.umdl.umich.edu/o/oaister/
[48] Voir http://www.sherpa.ac.uk/romeo.php

accès libre ; Elsevier, le plus gros éditeur international en Sciences Technologies et Médecine, appartient à la catégorie la plus ouverte, celle qui accepte la publication des pre-prints et des post-prints validés par la revue, sans délai d'embargo.

Les modalités organisationnelles se sont dégagées progressivement au sein de l'Open Archive Initiative[49]. La distinction entre revue en accès libre et archive ouverte a été formalisée : comme la revue traditionnelle, la revue en accès libre doit avoir un processus de validation des articles avant publication, alors qu'une archive peut contenir d'une part des articles scientifiques non validés, d'autre part des résultats de recherche de tous types et formats. Le concept d'archive thématique versus archive institutionnelle a été précisé plus tard ; la première regroupe des publications du même domaine scientifique, alors que la seconde publie l'ensemble des résultats des chercheurs d'un même organisme ou d'un ensemble d'organismes. En France, ce rôle est dévolu à HAL[50], développé et maintenu par le CCSD, Centre pour la Communication Scientifique Directe du CNRS, pour l'ensemble des EPST.

Pour permettre l'interopérabilité de ces serveurs d'archives dispersés, la convention de Santa-Fé (1999) a étudié diverses possibilités, dont la norme Z39.50, et a choisi de développer un protocole asynchrone simple, le protocole OAI-PMH, fonctionnel depuis 2001. Ce protocole est basé sur :

- la définition de deux types d'acteurs, les « fournisseurs de données », qui exposent des entrepôts de métadonnées compatibles avec le protocole (en général créés avec des outils libres comme EPrints, DSpace, CDSWare ou Fedora), et les « fournisseurs de services » ou « moissonneurs », qui interrogent régulièrement les entrepôts et répondent aux requêtes utilisateurs en exposant les données avec une certaine valeur ajoutée (les plus connus sont OAIster, ARC et CiteBase) ;
- l'exposition d'enregistrements (*record*) de métadonnées minimaux contenant des éléments Dublin Core simple en XML., sachant que les moissonneurs peuvent également interpréter et indexer tout document XML, par exemple MODS, METS ou EAD ;
- la définition de verbes de requête permettant aux fournisseurs de service de communiquer régulièrement avec les fournisseurs de données.

Une question restait en suspend, celle de l'interopérabilité des archives ouvertes avec l'ensemble du Web. En effet, les entrepôts de métadonnées font partie du « Web invisible » : ce sont en général des bases de données impliquant des requêtes dans un langage spécifique, et la réponse est une page dynamique. Un projet d'interfaçage de ces entrepôts avec le Web, dp9, n'a pas été réellement suivi par les producteurs d'archives ouvertes. Aujourd'hui, d'une part, les moteurs savent mieux interroger les bases de données, d'autre part, certains outils, comme Google ou le

---

[49] Open Archive Initiative, OAI, http://www.openarchives.org ;
Voir aussi « Libre accès à l'information scientifique et technique, http://openaccess.inist.fr
[50] HAL, Hyper Articles en Ligne, http://hal.ccsd.cnrs.fr

répertoire Yahoo, ont passé des accords avec les moissonneurs pour utiliser leurs index. Les archives ouvertes sont donc largement accessibles.

Les archives ouvertes sont une réalisation fondamentale en termes de stratégie alternative de publication et de diffusion, d'engagement de différents pays et institutions, et le protocole OAI-PMH constitue l'exemple phare de l'efficacité de métadonnées standardisées et pourtant très simples ; nous verrons dans l'exemple 5.4 que son adoption dépasse le domaine des publications scientifiques.

5.2 Bibliothèques numériques

Les publications montrent que les grandes bibliothèques numériques utilisent largement les standards MARC-XML, MODS, EAD et METS en interne et exposent leurs données en OAI-PMH, des réalisations nord-américaines (Eden 2004) aux projets européens (Bibliothèque Numérique Européenne, projet DRIVER visant à impulser et fédérer des bibliothèques scientifiques numériques européennes), en passant par le système d'information et le portail de l' ABES-SUDOC, Agence Bibliographique de l'Enseignement Supérieur, les réalisations de la Bibliothèque Nationale de France, et celles d'autres pays, aussi bien en Amérique du Sud qu'au Japon.

L'exemple de la British Library est instructif[51]. Suite à des développements de projets divers échelonnés dans le temps, souvent en partenariat, et soutenus par des métadonnées diverses et locales, cet organisme rencontrait des difficultés en matière d'interopérabilité externe et de maintenance du système d'information ; il a choisi de mettre en place pour tous les services un profil d'application Dublin Core intégrant des éléments MODS et quelques éléments d'un jeu national pour les documents administratifs. Ce profil a été élaboré parallèlement à celui du DCMI pour les bibliothèques, DC-Lib[52], avec des participants communs aux deux projets.

Par ailleurs, les systèmes d'information documentaires proposent presque tous aujourd'hui des modules d'export et de moissonnage OAI-PMH dans le cadre de leur fonction portail ; beaucoup intègrent des imports et exports XML de métadonnées standards et locales, et quelques-uns des bases compatibles XML.

5.3 La mutualisation des services culturels et patrimoniaux

Dans ce domaine où l'activité de numérisation est importante, les projets sont souvent nés du même constat : les nombreux services en ligne sont multiformes dans leurs interfaces comme dans leurs métadonnées, et les utilisateurs de plus en plus variés et transversaux. Pour leur éviter des recherches un peu aléatoires sur des sites divers, il est indispensable de fédérer l'accès aux services au travers

---

[51] Dublin Core Application Profiles at the British Library – a Case Study, Clayphan R, DCMI 2005, tutorial 5

[52] Library Application Profile, Clayphan R. and all, http://dublincore.org/documents/2004/09/10/library-application-profile/, septembre 2004

d'interfaces communes, déclinables selon les besoins : archives ouvertes, portails, catalogues…, couverture régionale, nationale ou internationale, objectif institutionnel, thématique ou autre…

Aux Etats-Unis, le RLG a développé un guide de recommandations pour les institutions (archives, musées, bibliothèques) alimentant la RLG Cutural Material Database, un projet fédératif de mise en ligne de corpus patrimoniaux ; les formats adoptés sont Dublin Core, MARC, MODS, EAD et METS[53].

Le projet français de mutualisation du patrimoine culturel numérisé impose à chaque organisme participant d'utiliser MARC-XML, MODS ou EAD, selon ses habitudes de travail, et prévoit un export de ces données en OAI-PMH, pour constituer une ou plusieurs archives ouvertes[54]. Il s'agit donc d'un nouvel usage de l'OAI-PMH, différent de l'accès aux résultats de la recherche, et il est probable que d'autres réalisations s'approprieront ce protocole simple et fédérateur.

Le projet européen MINERVA/MICHAEL fédère une quinzaine de pays et a repris les principes techniques du projet français[55]. La plateforme-portail STRABON[56], une de ses premières réalisations, fédère des sites sur des localisations archéologiques peu connues, proposant chacun un fonds d'images structuré, avec des documents sur le contexte et des possibilités de navigation diverses ; pour chaque site, les métadonnées des images ont été enrichies par des spécialistes au travers de logiciels de gestion d'images, puis exportées pour créer des enregistrements compatibles OAI-PMH, lesquels sont moissonnés pour le portail. Cette application utilise uniquement des outils libres, préexistants ou créés dans le cadre du projet.

5.4 Thèses électroniques françaises, TEF

Le projet TEF[57] a pour objectif de valoriser les thèses en augmentant leur diffusion et de créer une chaîne éditoriale unique, pour le document et ses métadonnées. Il a fait l'objet de textes réglementaires depuis 2000 et d'un

---

[53] Descriptive Metadata Guidelines for RLG Cultural Materials, http://www.rlg.org/en/page.php?Page_ID=214

[54] Numérisation du patrimoine culturel, informations techniques, http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/technique.htm
Réalisations, des catalogues pour l'instant, voir :
- Portail culture.fr http://www.culture.fr
- Catalogue des fonds culturels numérisés, http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm

[55] MICHAEL (Multilingual Inventory of Cultural Heritage in Europe), http://www.michael-culture.org/

[56] STRABON, Système d'Information Multilingue et Multimédia pour le Patrimoine Culturel et le Tourisme euro-méditerranéen, http://www.strabon.org/portal/

[57] Thèses électroniques françaises, http://www.abes.fr/abes/documents/tef/recommandation.html

partenariat très interactif entre les institutions concernées et l'AFNOR. Les métadonnées doivent respecter la richesse de description actuelle définie par l'ABES-SUDOC, tout en devenant interopérables pour des applications Web hors des bibliothèques. Elles intègrent le jeu Dublin Core qualifié, complété par des éléments MODS et ETD-MS (projet international de description des thèses), ainsi que des éléments propres pour le suivi administratif et le statut juridique des thèses en France ; ces éléments sont encapsulés dans un fichier METS pour chaque thèse et ses versions. La plupart des valeurs des éléments est contrôlée, et correspond à des espaces de nom reconnus.

Le *work-flow*, du dépôt du sujet à la soutenance sera géré au travers de la plateforme de saisie STAR (Signalement des thèses, Archivage et Recherche), qui termine sa période de test dans diverses universités pilotes ; celle-ci permet de créer les métadonnées au fur et à mesure, mais pas encore d'éditer la thèse en XML avec un modèle unique pour l'ensemble des universités. Le *work-flow* intègre la diffusion de la thèse soutenue et/ou de sa description sur divers portails et services, dont le portail du SUDOC et l'archive ouverte HAL et son archivage pérenne au CINES.

5.5 Le CN2SV, EAD et les métadonnées images

Le CN2SV[58], Centre National de Numérisation des Ressources Visuelles CNRS, né fin 2005, a pour mission la mise en place d'une structure d'informatisation, de conservation numérique et de mise à disposition par le web d'archives de laboratoire, de chercheurs et de scientifiques, ainsi qu'une activité de veille et d'expertise sur ce sujet (Pouyllau 2006) ; l'accent est mis sur la coopération entre chercheurs, documentalistes, bibliothécaires et informaticiens, indispensable pour organiser la gestion des données dans une plate-forme de préservation de ressources numériques. Techniquement, il est adossé au Centre Alexandre Koyré/CRHST, Centre de Recherche en Histoire et Technologies des Sciences. Une dizaine de corpus issus de l'histoire des sciences et des aires culturelles est accessible aujourd'hui.

Le CN2SV utilise exclusivement des métadonnées standards. Les propriétés (descriptives du contenu, techniques et administratives) de chaque image sont enrichies, puis extraites par programmation DOM pour alimenter le fichier EAD descriptif de la collection ; une attention particulière est portée aux questions d'archivistique (normes, intérêt et politique de sélection des fonds, veille sur les pratiques des différentes communautés scientifiques), aux droits d'auteurs et aux différents niveaux d'accès aux documents. La conservation repose sur le modèle organisationnel OAIS et sur des fichiers METS intégrant EAD et MIX. Pour la diffusion, le CN2SV est fournisseur de contenu OAI-PMH ; les enregistrements de granularité variée, créés automatiquement à partir du fichier EAD, permettent une exploitation diversifiée des ressources au travers de partenariats (portails de communautés scientifiques par exemple). Des fonctionnalités intéressantes sont proposées sur le site du CN2SV, comme une représentation graphique par nuages de

[58] CN2SV, http://www.cn2sv.fr

mots-clés qui permet d'affiner les recherches, et la possibilité pour chaque utilisateur autorisé d'ajouter des commentaires aux images, pour enrichir la description.

Faute d'outil adapté à ses besoins, le CN2SV a développé une chaîne complète de traitement des « instruments de recherche » EAD basée sur trois outils destinés à être proposés sous licence libre ; il explore parallèlement les nouveaux logiciels publiés.

5.5 Les métadonnées à l'INIST

En quelques années, pour continuer à remplir leurs missions, tous les services de l'INIST, Institut de l'Information Scientifique et Technique, ont été impliqués dans la production de métadonnées :

- portails dédiés à l'édition électronique de revues et congrès[59] ou de monographies, notamment rapports de recherche[60], implémentés sous DSpace, outil de gestion pérenne de corpus numériques compatible OAI-PMH ; contrairement à HAL, une grosse partie de l'activité porte sur de la mise en ligne rétrospective, et les métadonnées sont au maximum issues de signalements existants reformatés automatiquement ;

- évolution des notices bibliographiques d'un format UNIMARC-texte à un format UNIMARC-XML exploitant largement les possibilités d'arborescence et de liens entre éléments, notamment pour faciliter les usages infométriques ;

- création d'un format d'échange simplifié, EXODIC[61], profil d'application Dublin Core, utilisé en entrée dans le cadre de partenariats avec des laboratoires ou des éditeurs de revues pour alimenter les bases de données, et en sortie vers certains serveurs traditionnels et vers google et Google Scholar, qui indexent et proposent ainsi les 12,5 millions de références disponibles à l' INIST ;

- portail terminologique TermSciences[62], en association avec des partenaires, extérieurs ; cet outil est destiné à valoriser et mutualiser les ressources terminologiques des organismes publics de recherche et d'enseignement supérieur, vers un référentiel commun. Basé sur XML, il est conforme à la norme Terminologic Mark-up Framework, ISO 16642.

5.7 OpenURL

La norme OpenUrl permet d'accéder de manière contextualisée à des ressources Web dans un but d'une interopérabilité contrôlée des services documentaires. A partir de métadonnées (titre, auteur, ISSN…) accompagnant en arrière plan la citation d'une ressource, un résolveur de liens recherche les URL de localisation de la ressource, et propose préférentiellement celles qui permettent à l'utilisateur d'accéder au texte intégral selon ses droits, en terme d'abonnement à une revue par

---

[59] I-revues, http://irevues.inist.fr
[60] LARA, Libre accès aux rapports scientifiques et techniques, http://lara.inist.fr
[61] EXODIC, Exchange object description for INIST citations, http://exodic.inist.fr/
[62] TermSciences, http://www.termsciences.fr/

exemple. La version 0.1 a été achetée et mise en oeuvre par Ex-Libris, éditeur de logiciels documentaires, et ce protocole a repris un second souffle en 2003 avec sa version 1.0, approuvée par l'ISO : réorganisation des « entités » participant au protocole, possibilité d'exposer les métadonnées en XML et intégration de nombreux espaces de noms standards portant soit sur des jeux de métadonnées standards comme Dublin Core, MARC-XML, MODS, soit sur des identifiants uniques comme le DOI ou l'ISSN.

6. RDF : vers le Web Sémantique

6.1 Principes

RDF, Ressource Description Framework, est né en 1997 à l'initiative du W3C et repose sur un schéma XML spécifique. Il a été élaboré par une communauté professionnelle très large (représentants des bibliothèques, des standards du Web, experts des documents structurés, de la représentation de la connaissance, acteurs du domaine programmation objet et modélisation). Ce langage permet de construire le Web sémantique, qu'on peut définir comme un Web ouvert et décentralisé, formel et compréhensible par les machines, permettant l'interconnection structurelle des données (Euzenat 2004).

Conçu sur le modèle des langages de programmation objet, RDF offre d'abord un cadre formel de description des ressources au niveau structurel et syntaxique, sous la forme de triplets associant l'URI (Uniform resource identifier[63]), de la ressource, ses propriétés et les valeurs correspondantes ; ce cadre accueille tout jeu sémantique de métadonnées. RDF Schema[64] y ajoute un niveau de structure supplémentaire en définissant un certain nombre de propriétés et de classes d'objets pouvant être utilisés dans les modèles de documents. Enfin, un langage de requête pour l'environnement RDF, s'appuyant sur le protocole HTTP, a été développé en 2004, RDF Data Query Language ou RDQL.

Ce langage permet de relier structurellement les descriptions - ou parties de descriptions - de ressources entre elles. Par exemple, une œuvre, d'abord décrite dans son contexte de production, sera reliée à l'ensemble des informations produites à son sujet, qu'elles portent sur le contenu, les acteurs ou les droits, ainsi qu'aux ressources terminologiques liées et à d'autres données factuelles éventuelles ; la masse d'information disponible sur le Web prend alors un réel sens structurel et cognitif.

---

[63] L'URI, ou mieux, le PID, Persistent identifier, correspond à un certain nombre de projets complémentaires, dont un méta-projet du W3C, l'URN, Unique resource name, visant à faire correspondre un identifiant unique et stable aux localisations de la ressource

[64] http://www.w3.org/TR/rdf-schema/

Ce projet intéresse les producteurs de jeux de métadonnées ; Dublin Core ou LOM, par exemple, disposent d'un schéma RDF, et MPEG21 comprend une partie RDF. Il a suscité également la création de jeux sémantiques complémentaires.

Bien que les exemples d'application se multiplient, il est difficile de savoir si le web de demain sera réellement RDF, comme le souhaite le W3C ; l'implémentation de ce standard est complexe, sa progression lente, les outils réellement orientés RDF encore rares[65], et d'autres initiatives proches sur le plan conceptuel, mais syntaxiquement un peu différentes se sont développées en parallèle.

6.2 Quelque exemples

6.2.1 Descriptions de personnes

Nous avons parlé plus haut des vCard ainsi que du projet FOAF, qui permet entre autres de relier une œuvre à son contexte de production, d'identifier des réseaux professionnels et leurs évolutions, et de pratiquer la recherche d'information par « associations » de ressources diverses.

6.2.2 Droits des oeuvres

Différents formats courants de ressources, textuelles, audio, images et multimedia disposant d'une description en XML intègrent volontiers une description des droits en RDF ; Creative Commons, par exemple ; cette initiative propose également des outils d'intégration et un travail collaboratif aux développeurs [66].

Pour les enregistrements OAI-PMH, en XML simple jusqu'à récemment, l'Open Archive Initiative recommande depuis décembre 2004[67] de préciser les droits d'usage en intégrant par exemple une licence Creative Commons en RDF dans l'élément Dublin Core « rights ».

6.2.3 Ontologies et taxonomies

Les ontologies sont des représentations dynamiques des connaissances dans un domaine. Les taxonomies relèvent de l'activité de classification.

Les ontologies sont développées soit sous RDF, notamment avec OWL, Ontologies Writing Language, soit avec des langages et concepts proches, comme Topics Maps ou SKOS.

---

[65] Voir par exemple, Redland RDF Applications Framework, http://librdf.org/
[66] http://wiki.creativecommons.org/Embedding_Specifications et http://wiki.creativecommons.org/Category:Developer
[67] http://www.openarchives.org/OAI/2.0/guidelines-rights.htm

Ces ressources terminologiques ont de fortes implications industrielles, et sont souvent développées dans un cadre réservé. La société Ontopia, qui travaille dans ce domaine, a mis en accès libre un exemple sur l'opéra italien sous Topics Maps [68].

6.2.4 RSS[69] et syndication de sites

La syndication est une technique permettant d'afficher automatiquement de l'information provenant d'autres sites, éventuellement triée, filtrée et reconfigurée, avec mises à jour automatiques. Les sites producteurs éditent un fichier, le canal ou fil RSS, décrivant leurs nouveautés, fils que l'utilisateur visualise dans un lecteur (ou agrégateur) de RSS[70], ou dans un navigateur comme Firefox. Cette technique permet donc de suivre facilement l'évolution des sites disposant de canaux, et représente une alternative à la consultation régulière des sites, , humaine ou automatique, ainsi qu'à un certain nombre de listes de diffusion. Le nombre des canaux RSS disponibles augmente très rapidement actuellement, et leur existence sur un site fait partie des critères de qualité chez nos voisins anglo-saxons

Il existe deux grandes versions de RSS, toutes deux XML, l'une en RDF et l'autre de syntaxe plus simple, chacune ayant évolué avec le temps. Elles intègrent entre autres la possibilité de décrire le contenu en Dublin Core. Un autre schéma, Atom, développé en 2004, visait à unifier la syntaxe et à donner des possibilités supplémentaires de description de ressources non textuelles. Toutes ces versions sont proches dans le principe et la syntaxe, et sont lisibles par les mêmes outils d'affichage RSS. Sur le modèle d'Atom, les schémas RSS de bases intègrent aujourd'hui des modules complémentaires pour la description et la manipulation des documents audio, images, video[71].

Aujourd'hui, au-delà d'une information standard sur les nouveautés d'un site, RSS est utilisé pour intégrer une description de contenu orientée usages, et les utilisations se sont diversifiées. Certains sites proposent des fils thématiques. De plus en plus, les bases de données bibliographiques et factuelles proposent des fils RSS personnalisés pour les mises à jour régulières des recherches enregistrées sur leurs serveurs. Les amateurs de musique, vidéo, émissions radiophoniques l'utilisent largement pour repérer, filtrer, télécharger et classer automatiquement les nouveautés mises en ligne... voire créer automatiquement de nouveaux médias thématiques ! La dernière nouveauté est le RSS bidirectionnel, où divers utilisateurs deviennent aussi producteurs, par exemple pour gérer des agendas partagés. Les

---

[68] Ontopia, http://www.ontopia.net/, onglet 'topics Maps'

[69] RSS : à l'origine "RDF - Resource Description Framework - Site Summary", souvent maintenant "Real Simple Syndication", parfois "Rich Site Summary"

[70] Un exemple d'agrégateur libre et en ligne : Bloglines, http://bloglines.com

[71] voir http://en.wikipedia.org/wiki/Podcasting
ou Media RSS, http://search.yahoo.com/mrss/mrss

RSS représentent ainsi la première application XML/RDF largement adoptée sur le Web[72].

7. Questions et perspectives

Ces dix dernières années ont mis en lumière le rôle des métadonnées dans la préservation et la réutilisation des ressources, et permis de progresser dans l'adoption des standards ; un certain nombre de questions restent d'actualité.

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 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 ?

Concernant la qualité des métadonnées disponibles, plusieurs études, certes un peu anciennes (citées par Day 2005), montrent que le résultat laisse souvent à désirer sur le plan de l'exhaustivité et de l'interopérabilité, entre différentes archives OAI-PMH, pour le multimédia ou les données de la recherche. La standardisation des valeurs et pratiques représente une partie de la réponse. Mais, pratiquement, il n'est pas toujours facile de savoir ce qui sera utile à d'autres demain, tout en restant dans un investissement raisonnable.

Une question liée est celle de « la subjectivité cachée et du biais culturel » mis en œuvre lors de l'implémentation des métadonnées au temps T pour une communauté précise, alors que les contextes d'utilisation futurs seront probablement très différents. Le modèle OAIS insiste sur la description des connaissances de la communauté, et la nécessité parfois d'ajouter des données pour aider d'autres communautés à interpréter les données (Day 2005).

Il faut enfin souligner l'importance de fichiers de métadonnées pérennes et accessibles, eux aussi. Le choix de XML est un élément de réponse, et les fichiers correspondants sont théoriquement moins vulnérables que les ressources. Le modèle OAIS envisage également cette question. Il rappelle que les métadonnées aussi peuvent être appelées à migrer, lors d'évolution des standards par exemple, et que cette migration doit être accompagnée et documentée. Dans ce modèle, les métadonnées sont de vrais objets numériques, qui appartiennent aux mêmes

[72] Différentes conférences de Hervé Lecrosnier, actuellement à l'EBSI, évoquent ces usages

packages que la ressource, mais peuvent aussi faire l'objet d'une modélisation séparée, dans d'autres bases de données liées par exemple.

Le chantier multiforme ouvert autour des métadonnées a déjà des fondations solides, mais il n'est pas prêt de s'achever en termes d'évolutions et d'usages!

Références

American Council of Learned Societies Commission for Cyberinfrastructure for the Humanities and Social Sciences, *Our Cultural Commonwealth*, rapport, décembre 2006, http://www.acls.org/cyberinfrastructure/OurCulturalCommonwealth.pdf

ARTIST, Le Henaff D., Ducloy J., Ducasse J-P., Grivel L., Foulonneau M., Nicolas Y., Métadonnées pour une cyberinfrastructure de la recherche : le cas des thèses françaises, mai 2006, http://artist.inist.fr/article.php3?id_article=332

Day Michael, Installment on « metadata », *Digital Curation Manual, version 1.1*, novembre 2005, http://www.dcc.ac.uk/resource/curation-manual/chapters/metadata/metadata.pdf

Dhérent C., Faire un répertoire ou un inventaire simple en EAD, version du 18 février 2003, http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/repertoireEAD.html

Dhérent C. (modérateur), *Des métadonnées pour bien utiliser les ressources électroniques*, Journée d'information AFNOR/CG46, 7 juin 2005, Bibliothèque nationale de France, http://www.bnf.fr/pages/infopro/journeespro/no-Afnor2005.htm [11 documents]

Eden B. L. (theme editor), *MARC and metadata : METS, MODS and MARC XML : current and future implications, Part 1*, Library Hi Tech, 2004, vol. 22, n°1, p. 6-112, http://www.emeraldinsight.com/0737-8831.htm [10 articles]

Eden B. L. (theme editor), *MARC and metadata : METS, MODS and MARC XML : current and future implications, Part 2*, Library Hi Tech, 2004, vol. 22, n°2, p. 119-180, http://www.emeraldinsight.com/0737-8831.htm [7 articles]

Euzenat J., Troncy R., *Web sémantique et pratique documentaire, Publier sur Internet*, Séminaire INRIA, 27 septembre – 1° octobre 2004, Aix-les-bains, ADBS éditions, p. 157-188

Higgins Sarah, What are metadata standards, http://www.dcc.ac.uk/resource/standards-watch/what-are-metadata-standards/, août 2006

Hillmann D., "Using Dublin Core", novembre 2005, http://dublincore.org/documents/usageguide/

Jantz R., Giano M-J., Architecture and technology for trusted digital repositories, D-Lib Magazine, juin 2005, Vol 11 n° 6, http://www.dlib.org/dlib/june05/jantz/06jantz.html

McCown F., Chan S.,Nelson M-L, Bollen J., The Availability and Persistence of Web References in D-Lib Magazine, 5th International Web Archiving Workshop, 22-23 septembre 2005, Vienne (Autriche) http://www.iwaw.net/05/papers/iwaw05-mccown1.pdf

National Information Standards Organisation (NISO), Understanding Metadata, 2004, http://www.niso.org/standards/resources/UnderstandingMetadata.pdf [NISO, 2004-1]

NISO Workshop: Metadata Practices on the Cutting Edge. Washington, May 20 2004, [13 documents], http://www.niso.org/news/events_workshops/MD-2004_agenda.html [NISO, 2004-2]

Peccatte P./ Soft Experience, Métadonnées: une initiation - Dublin Core, IPTC, EXIF, RDF, XMP", mis à jour mars 2006, http://peccatte.karefil.com/software/Metadata.htm

Pouyllau S. Pouyllau D., Mouton M-D., Melka F., L'archivage des données numériques pour la recherche par le Centre National pour la Numérisation de Sources Visuelles (Centre de Ressources Numériques du CNRS), *Les Rencontres 2006 des Professionnels de l'IST*, Nancy, 18-20 juin 2006
http://hal.archives-ouvertes.fr/docs/00/09/61/10/PDF/CN2SV-presentation.pdf,
http://hal.archives-ouvertes.fr/docs/00/09/61/10/PDF/CN2SV-CNRS-poster2006.pdf

Robial M., "Synthèse : RSS et syndication de contenu", *Liste ADBS-info*, mai 2004, http://sympa.adbs.fr/wws/arc/adbs-info/2004-05/msg00028.html

W3C, "RDF primer, W3C recommandation", février 2004, http://www.w3.org/TR/rdf-primer/