Panorama : des métadonnées pour les ressources électroniques

Catherine Morel-Pair

Panorama : des métadonnées pour les ressources électroniques

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 maillon essentiel pour l'interopérabilité de l'information et sa gestion. Des applications sont développées par de nombreux acteurs de différents domaines pour tous les types de ressources numériques; à côté du jeu d'éléments Dublin Core, central pour l'interopérabilité, il existe des ensembles complémentaires tout aussi importants, qui répondent chacun à des besoins particuliers. Leur implémentation se fait aujourd'hui principalement en XML, et l'adoption du format RDF permettra d'atteindre une réelle interopérabilité entre ces ressources. Ce document est un essai de synthèse sur le contexte d'apparition des métadonnées, les principaux jeux d'éléments et leurs formats d' implémentation, avec des exemples d'applications.*

ABSTRACT. *Metadata are structured data sets about physical or digital resources. It is an essential way towards information interoperability and management. Many applications are developed by occupational actors in various fields. Today, Dublin Core metadata set is three R'S for interoperability, and other important complementary sets answer other needs. XML language is the current metadata implementation format up-to-now, and RDF format will reach a real resources interoperability. This document wants to give a review of metadata in context, main sets and implementation formats, with applications examples.*

MOTS-CLÉS : *métadonnées, ressources électroniques, Dublin Core, XML, RDF, interopérabilité, partage d'information, projets, synthèse.*

KEYWORDS : *metadata, numeric resources, Dublin Core, XML, RDF, interoperability, information sharing, projects, review.*

1. Introduction

Une métadonnée (du Grec, “méta”, ce qui dépasse, englobe) est une donnée à propos d’une autre donnée. En sciences de l’information, les métadonnées sont des ensembles de données structurées décrivant des ressources physiques ou numériques, ou, sur un plan plus fonctionnel, “de l’information structurée qui décrit, explique, localise la ressource et en facilite la recherche, l’usage et la gestion” [14].

Si ce terme est utilisé depuis longtemps dans certains domaines ou formats de ressources (données géospatiales, images, audiovisuel) , il n’est apparu que depuis une dizaine d’années dans les métiers de l’information et de la conservation, où d’autres termes correspondent à des concepts proches, catalogage et signalement par exemple. Aujourd’hui, cependant, les spécificités du document électronique en matière de cycle de vie, de gestion, de droits d’usage et de recherche d’information et les besoins d’interopérabilité entre ressources font des métadonnées un élément incontournable dans le paysage numérique.

Après un tour d’horizon de ce contexte, nous situerons les initiatives et leurs acteurs ; nous détaillerons le jeu d’éléments Dublin Core, central pour l’interopérabilité, puis l’intérêt du langage XML pour les métadonnées ; nous passerons en revue d’autres jeux d’éléments complémentaires, à orientation bibliographique ou créés pour d’autres objectifs, et citerons des projets utilisant ces formats ; nous expliciterons l’intérêt de RDF pour l’interopérabilité, et les applications qui en sont issues aujourd’hui ; nous terminerons par quelques idées générales sur l’implémentation de métadonnées en pratique.

2. Des besoins nouveaux, de description, gestion et interopérabilité

2.1 Ressources électroniques et fonctions des métadonnées

Par rapport à l’objet d’information traditionnel, géographiquement situé, et conçu selon une chaîne de production et de diffusion spécifique, le document électronique est un objet fragile, et qui a un cycle de vie spécifique : diversité des origines, formats et contenus, évolution fréquente du contenu, accessibilité plus large, facilité technique d’utilisations diverses, pérennité très dépendante de machines et logiciels eux-mêmes peu pérennes.

Mais, et surtout, au-delà de spécificités du document isolé, la question est celle d’une masse exponentielle de ressources informationnelles distribuées, hétérogènes etmouvantes, et de son impact tant sur les pratiques de recherche d’information que de production. Sur le Web, en particulier, la “découverte” (*discovery*) de ressources devient plus riche et plus transversale, mais aussi plus aléatoire quant à la pertinence des résultats, et il importe d’aller vers le Web sémantique. Du côté production, des ressources très diverses (ou leurs signalements) peuvent contribuer à créer de nouvelles ressources, et être intégrées dans différents contextes, (annuaires et nouvelles collections virtuelles par exemple), ce qui implique, outre le respect des droits, une certaine interopérabilité entre ces ressources, au minimum interopérabilité de leurs descriptions.

Les métadonnées standardisées constituent une réponse à ces nouveaux besoins des ressources et collections numériques ; elles permettent en effet [4, 14 et divers] :

- l'amélioration de la pertinence et de l'exhaustivité des recherches, le tri et le filtrage des résultats,
- la traçabilité : historique des mises à jour, des versions successives, autres formats, sources …
- la description de la propriété intellectuelle et des droits d'usage,
- l'évaluation ou la labellisation des ressources, produits et services,
- la représentation des ressources selon différents niveaux de granularité, de la collection au "segment" de ressource, intégrant la structure de l'ensemble,
- l'échange automatisé de données : génération de portails ou catalogues communs, syndication de sites…
- l'analyse statistique à des fins de bibliométrie, infométrie, travaux terminologiques,
- une meilleure gestion informatique des ressources, à court et long terme (pérennité de l'archivage et de l'accessibilité),
- la description de personnes physiques : signatures électroniques, réseaux sociaux, profils utilisateurs (annuaires, tels LDAP, certificats, choix de préférences sur un site… et gestion, ou personnalisation des accès),
- la description des acteurs, procédures et outils de production, préconisée dans le cadre du Record Management pour assurer la fiabilité, la pérennité et la qualité des données.

Nous nous attacherons ici aux métadonnées liées aux ressources. Selon une typologie devenue classique, il existe dans ce cadre trois grandes catégories complémentaires, qui rejoignent et regroupent les fonctions décrites ci-dessus :

- métadonnées descriptives du contenu "intellectuel" de la ressource : titre, sujet, description libre…,
- métadonnées instancielles, ou techniques et de structure : caractéristiques physiques et informatiques décrivant l'instance de la ressource (format, date …), sa structure ou sa place dans une hiérarchie, les outils nécessaires à sa consultation,
- métadonnées administratives, portant d'une part sur la propriété intellectuelle, la responsabilité, les droits d'utilisation et les sources utilisées, d'autre part sur la préservation de la ressource (archivage et accessibilité pérennes).

2.2 Autour de l'interopérabilité

L'interopérabilité peut être définie comme 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" [14]. Dans un monde de plus en plus numérique, c'est un objectif important, auquel les métadonnées participent en complémentarité avec d'autres processus.

Sur un plan technique, l'interopérabilité se réalise à 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.

Le tableau suivant résume le positionnement en ces termes de différents standards bien connus.

	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

Ainsi, les métadonnées sont toujours implémentées dans un langage structuré, (X)HTML, XML, RDF[1], et échangées par des protocoles. Selon l'usage attendu, les éléments sont présents soit dans la ressource elle-même (documents HTML, métadonnées natives des images…), soit dans un fichier associé ; un "enregistrement" (*"record")* correspond à la description d'une ressource dans un format particulier.

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'autorités standardisés, ontologies). Dans ce but, de plus en plus de producteurs de jeux de métadonnées éditent des documents, registres de métadonnées décrivant chaque élément (caractéristiques, liens avec les autres éléments, usage, valeurs attendues, parfois *"mapping"* avec d'autres jeux…), et des*"guidelines"* ou recommandations de bonnes pratiques. Pour les valeurs, il convient dans tous les cas d'utiliser au maximum des outils reconnus et accessibles en ligne, codes ISO ou W3C, vocabulaires, thesaurus, classifications, ontologies et listes d'autorité existantes ou à venir.

[1] HTML : HyperText Mark-up Language,
XML : eXtensible Mark-up Language,
XHTML: structuration plus stricte du HTML selon les règles XML
RDF: Resource Description Framework

3. Des acteurs et des initiatives, un essai de typologie …

Le Consortium W3C, World Wide Web Consortium, composé surtout d'industriels de l'informatique et des TIC, est l'instance internationale permanente chargée de l'avenir du Web. Il travaille essentiellement, et activement, autour de l'élaboration et de la diffusion des protocoles et des formats syntaxiques et structurels, de HTML à RDF et au Web sémantique ; le site www.w3.org est incontournable sur ces aspects.

Dès les premières versions de HTML, et depuis, il existe des métadonnées, incluses dans l'en-tête du fichier ; elles appartiennent aux trois grandes catégories décrites, de façon opérationnelle et précise pour une page Web. En pratique, cependant, elles souffrent d'une standardisation insuffisante des usages et valeurs des éléments, et sont sous-utilisées, tant par les producteurs de sites que par les outils de recherche.

Pour pallier ces limites, le jeu d'élément Dublin Core a été créé en 1995 par le DCMI, Dublin Core Metadata Initiative[2], groupe international de professionnels du milieu documentation-bibliothèques et d'autres origines (musées, IETF[3]…) réuni à Dublin dans l'Ohio. Ce jeu correspond à une description "généraliste" s'appliquant à tout type de ressource, dans le but d'améliorer la recherche d'information, et avec l'objectif d'une large utilisation, en particulier sur le Web. Le DCMI, devenu instance permanente, travaille toujours à l'évolution et à l'adaptation de ce jeu, au travers de groupes thématiques ouverts et d'un workshop annuel.

D'autres acteurs de différents métiers ont élaboré des jeux sémantiques plus spécifiques, répondant à leurs besoins particuliers de description et de gestion des ressources, jeux en général orientés préférentiellement vers une ou deux des catégories citées. Si certains ont eu une vie relativement éphémères, beaucoup d'autres sont en voie de normalisation ou sont reconnus comme des standards.

Dans ce qui peut apparaître comme un foisonnement d'initiatives, on peut essayer une typologie qui situe ces standards selon leurs contextes d'utilisation.

Certains jeux s'appliquent à un domaine de connaissances ou d'action très spécifique, comme FGDC[4] pour la description géospatiale, HIDDEL[5] pour l'évaluation des sites santé, Creative Commons[6] pour les droits d'utilisation des ressources, vCcard[7] pour la signature électronique, FOAF[8] pour les personnes physiques, leur activité et leurs réseaux sociaux…

---

[2] Dublin Core Metadata Initiative, http://www.dublincore.org
[3] Internet Engineering Task Force, http://www.ietf.org/
[4] Federal Geographic Data Committee, http://www.fgdc.gov/metadata/metadata.html
[5] Health Information Disclosure, Description and Evaluation Language, basé sur les éléments du Net Scoring, http://www.medcircle.org/
[6] Voir http://creativecommons.org
[7] Voir http://www.w3.org/TR/2001/NOTE-vcard-rdf-20010222/
[8] The Friend of a Friend (FOAF) Project, http://www.foaf-project.org/

Certains autres sont liés à des formats spécifiques de ressources : IPTC, XMP et MIX pour l'image ; MPEG7 et MPEG21 pour le multimédia, les éléments de l'en-tête des documents structurés en XML selon des DTD telle Docbook, Erudit ou TEI[9]…

D'autres encore sont orientés métier :

- MARC-XML, UNIMARC-XML (BiblioML par exemple) et MODS pour la description bibliographique[10], avec MADS pour les données d'autorité
- EAD pour les fonds archivistiques et collections spécialisées, avec EAC pour les données d'autorité[11],
- ONIX pour les métiers du livre[12],
- IEEE-LOM et IMS Metadata pour les ressources éducatives[13].

Il existe enfin des "méta-formats", METS ou IMS-CP[14], qui décrivent de manière très structurée une collection d'objets numériques quelconques et éventuellement dispersés, et font appel à d'autres jeux pour décrire l'objet de base.

A l'exception de quelques-uns, qui peuvent être vus comme "concurrents", MARC-XML et MODS par exemple, ces jeux sont surtout complémentaires en matière d'objectifs, d'objet décrit ou de métier ; chacun a un intérêt particulier et un rôle à jouer dans le paysage, et tous ces standards sont utilisés largement aujourd'hui, seuls ou associés entre eux ; ainsi, pour répondre à tous les objectifs qu'elles se fixent, les grandes réalisations autour des métadonnées font en général appel à plusieurs jeux d'éléments standards.

A côté de ces initiatives, un cadre de réflexion plus générique s'impose, qui permet d'aller plus loin dans la modélisation et l'interopérabilité, en tirant le meilleur parti des initiatives actuelles.

---

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

[10] MARC : Machine Readable Cataloging, http://www.loc.gov/standards/ ;
MODS, Metadata Object Description Schema, http://www.loc.gov/standards/ ;
BiblioML, représentation XML des formats bibliographiques et autorités Unimarc, développée en France par le ministère de la Culture et la société AJLSM, http://www.biblioml.org/fr/

[11] EAD: Encoding Archive Description, EAC: Encoding Archives Context, voir http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html et http://www.loc.gov/standards/

[12] Sur le site d'EDItEUR, groupe pour le commerce électronique du livre et des séries, http://www.editeur.org/

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

[14] METS, Metadata for Encoding and Transmission Standards, http://www.loc.gov/standards/
l'IMS Global Learning Consortium, http://www.imsglobal.org/ propose à la fois un jeu d'éléments orienté objets d'apprentissage, IMS Metadata, proche de LOM, et un format de description de collections, IMS Content Package (IMS-CP)

Les organismes normalisateurs et "standardisateurs" jouent un rôle important dans le processus, notamment le comité technique ISO TC46 au niveau international. Pour la France, son homologue et représentant, le Comité Général AFNOR CG 46, chapeaute la commission de Normalisation CN 357, qui comprend des groupes d'experts travaillant sur différentes thématiques : description bibliographique, codes langues et pays, métadonnées des thèses et d'autorité, DTD EAD-EAC, adaptation du jeu d'éléments LOM. Le NISO, National Information Standards Organization américain "identifie, développe, maintient et publie des standards techniques pour gérer l'information dans un monde évolutif et de plus en plus numérique" en lien avec le monde industriel, et s'intéresse de près aux métadonnées [14].

De nombreux organismes et groupes de travail, internationaux ou nationaux, réfléchissent à l'usage, à l'implémentation et l'interopérabilité des métadonnées, chacun dans son domaine d'activité et développent des correspondances certains jeux d'éléments. L'OCLC s'investit particulièrement, tant pour une réflexion conceptuelle que pour le développement d'outils[15].

Enfin, la question des métadonnées est loin d'être étrangère aux acteurs de la recherche scientifique, en sciences de l'information évidemment, mais dans bien d'autres domaines divers, des applications géospatiales à la télémédecine, en passant par l'analyse automatique des images, le séquençage des documents audio-visuels ou l'amélioration de l'accès aux ressources Web pour des groupes particuliers comme les handicapés, par exemple…

4. Dublin Core au centre de l'interopérabilité

4.1 Eléments, qualificatifs et spécifications d'encodage

Les premiers travaux du DCMI ont abouti à la création d'un jeu de base de 15 éléments, normalisé ultérieurement par l'ISO (norme 15836-2003).

Ce module de base a été enrichi par :

- des qualificatifs (*"refinements")* précisant les éléments
- des "schémas d'encodage", soit syntaxiques (modèles d'implémentation des éléments en XML ou RDF), soit sémantiques (vocabulaires et codes recommandés pour les valeurs).

L'ensemble Dublin Core continue aujourd'hui à évoluer, avec l'ajout d'éléments plus spécifiques et de qualificatifs.

L'implémentation se fait en Dublin Core simple, sans qualificatifs, ou en Dublin Core qualifié, intégrant ces derniers. Un guide d'usage détaillé est présent sur le site officiel [10], et renvoie à d'autres documents plus techniques.

---

[15] Golby C-J., Smith D., Childress E., "Two Path to Interoperable Metadata", OCLC Research Publications, http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf

Le tableau suivant[16] présente les 15 éléments de base avec leur définition et une équivalence avec le format UNIMARC, bien connu des bibliothécaires et documentalistes.

Dublin Core	UNIMARC, zone, sous-zone
Title *Titre du document, à priori titre principal* *Un qualificatif : alternative (autres titres)*	200 $a, Titre propre 200 $e, Sous-titre et compléments du titre 510 à 517, Autres formes de titre 540, 541, Titre ajouté/traduit par le catalogueur
Creator *Nom de la personne, de l'organisation ou du service à l'origine de la rédaction du document*	700 $a, Nom de personne - Responsabilité principale 701 $a, Nom de personne - Autre responsabilité principale 710 $a, Collectivité – Responsabilité principale 711 $a, Collectivité - Autre responsabilité principale 200 $f, Première mention de responsabilité
Subject *Sujet, mots-clés, phrases de résumé, ou codes de classement*	606, Nom commun – Vedette matière 607, Nom géographique – Vedette matière 610, $a Indexation, vocabulaire libre 675, Classification UDC 676, Classification DDC 680, Classification LCC 686, Autres Classifications
Description *Résumé, table des matières, ou texte libre* *Qualificatifs : abstract et tableOfContents*	330 $a, Résumé ou extrait
Publisher *Nom de la personne, de l'organisation ou du service à l'origine de la publication du document*	210, $c Nom de l'éditeur, du diffuseur, etc… 210, autres sous-zones pour autres données (adresse…)
Contributor *Nom d'une personne, d'une organisation ou d'un service qui contribue ou a contribué à l'élaboration du document*	701 $a $b, Nom de personne - Autre responsabilité principale 702 $a $b, Nom de personne – Responsabilité secondaire 711 $a, Collectivité - Autre responsabilité principale 200 $g, Mention de Responsabilité suivante
Date	210 $d, Date de publication, diffusion, etc.

[16] Sources :
OpenWeb Group, http://www.openweb.eu.org/articles/dublin_core/
UK Office for Library and Information Networking, University of Bath, http://www.ukoln.ac.uk/metadata/interoperability/dc_unimarc.html,
Les métadonnées des theses électroniques françaises, http://www.abes.fr/abes/documents/tef/recommandation/tef_01.pdf
Manuel UNIMARC, format bibliographique, 4° edition, version française, UBCIM Publications, 2002

Date d'un évènement dans le cycle de vie du document *Huit qualificatifs*	100 $a, Données générales de traitement (autres zones pour mentions plus spécifiques, thèses, congrès…)
Type *Nature ou genre du contenu*	105 $a, Données codées : monographie 106 $a, Données codées : texte – présentation physique 336 $a, Note, type de ressource électronique 608 $j Vedette matière forme, genre ou caractéristiques physiques, subdivision forme
Format *Format physique ou électronique du document* *Qualificatifs : extend et medium*	135 $a, Données codées – ressources électroniques 230 $a, Ressources électroniques, définition et volume des fichiers 856 $q et $s, Adresse électronique et mode d'accès, type de format et taille du fichier
Identifier *Identificateur non ambigu, Système de référencement standardisé (URI, ISBN…)*	010 $a, ISBN et 011 $a, ISSN 012 à 073, Autres numéros standardisés 856 $u, Adresse électronique et mode d'accès, identificateur électronique normalisé (URI, ex.URL) 856 $f, Nom du/des fichier(s)
Source *Ressource dont dérive le document en totalité ou en partie*	324 $a, Note sur le document original Voir aussi les correspondances avec certains qualificatifs de Relation, tableau suivant
Language *Langue du document*	101 $a, Langue du document
Relation *Ressource liée, logiquement ou techniquement* *Dénomination formelle recommandée* *Douze qualificatifs*	Eléments du bloc 4XX surtout, voir détails dans le tableau suivant
Coverage *Portée du document, couverture temporelle, spatiale ou administrative* *Qualificatifs : spatial et temporal*	606 $z, Nom commun, vedette matière, sub-division chronologique 607, Nom géographique, vedette matière 610 $a, Indexation, vocabulaire libre
Rights *Informations sur le statut des droits de la ressource ou lien vers le détenteur des droits* *Trois qualificatifs (dont 2 récents)*	300 $a, Note générale

Pour la description des qualificatifs, voir le registre du site Dublin Core, ou la page de l'Open Web, en Français [12]. A titre d'exemple, sont donnés en annexe 2 les qualificatifs de l'élément "relation" avec leur définition et une correspondance avec les zones et sous-zones UNIMARC.

4.2 Implémentation

Différentes syntaxes permettent d'implémenter les éléments Dublin Core :

- HTML et XHTML, d'intérêt limité, hors volonté de standardisation, car non utilisé par les moteurs actuels,

- XML, utilisé en particulier dans l'enregistrement OAI-PMH minimal, et dans des projets d'échange et de mutualisation de données,

- RDF, au niveau d'interopérabilité plus général de ce format.

4.3 Intérêt et limites

Dublin Core fait l'objet d'un large consensus et d'une large utilisation aujourd'hui grâce aux atouts suivants[17] :

- sa création dans un contexte international et multidisciplinaire ;
- sa sémantique simple et "commune", facilement compréhensible, particulièrement pour les éléments de base ;
- son extensibilité (compatible avec d'autres jeux d'éléments, évolutivité) et sa flexibilité (grande souplesse d'implémentation) ;
- son adoption dans différents domaines, métiers et pays, et dans des applications non prévues initialement ou des domaines industriels connexes ;
- son évolutivité au travers de groupes de travail ouverts ;
- la volonté du DCMI, de diffuser et faire adopter ce modèle ; le site officiel, www.dublincore.org, est très riche en tutoriels et recommandations, exemples, modèles et outils, et reflète bien l'activité de ce groupe d'acteurs ;
- la normalisation des 15 éléments de base.

Cependant, ce jeu est orienté essentiellement vers la description d'un objet électronique simple en vue de la recherche d'information ; il est insuffisant pour la description de collections, la gestion administrative et technique, limité pour la description d'objets physiques, et peu contraignant en matière d'architecture.

Les groupes de travail du DCMI ont pour but de faire évoluer ces limites ; aujourd'hui par exemple, certains travaillent sur les thèmes suivants : architecture, descriptions de personnes, domaines environnement et éducation, documents images et multimédia, collections, bibliothèques, préservation…

[17] Adapté de *Susan Haigh, Flash Réseau n°63,* Services de technologie de l'information, Bibliothèque nationale du Canada, Décembre 1999

5. XML et métadonnées

5.1 Intérêt de XML

XML, eXtensible Markup Language ou langage extensible de balisage, est un sous-ensemble de la norme internationale SGML avec des apports du langage hypertexte HTML. Il est très utilisé, tant pour encoder les ressources que pour implémenter des métadonnées destinées à l'échange. En effet :

- c'est un format pérenne, car indépendant des logiciels et utilisant UNICODE
- il permet une structuration logique du contenu du document,
- il existe des modèles standards et partageables de documents, DTD (Document Type Description) et, plus récemment, schéma XML. Plus proche du document XML, plus souple quant à l'évolutivité, le schéma XML permet en outre de définir strictement les valeurs attendues et d'intégrer de manière non ambiguë des éléments d'autres schémas, en faisant référence explicitement à leurs espaces de nom[18] ;
- les ressources XML peuvent être utilisées de multiples façons grâce au processus de transformation XSLT basé sur les feuilles de style XSL et aux mécanismes XPath et XQuery : présentations diverses du même document, génération de "nouvelles" ressources par extractions de données particulières ou intégration de ressources hétérogènes, analyse de corpus …
- XML est enfin plus simple à implémenter que SGML, dont il reprend 10% des éléments seulement.

Ainsi, une structuration XML des métadonnées leur donne de l'efficacité et de la valeur : pérennité et accessibilité du contenu, enregistrements facilement convertisables pour d'autres utilisations, dont le partage de données (coopérations, alimentation de portails et catalogues, d'archives ouvertes, analyses statistiques) De plus, si la ressource elle-même un document XML, la création des métadonnées est facile à automatiser.

.

5.2 Dublin Core en XML

L'implémentation des composants Dublin Core en XML est décrite en détail sur le site officiel [11]. Le schéma proposé est simple, avec un seul niveau d'arborescence, en Dublin Core simple comme en Dublin Core qualifié. En Dublin Core simple, il a été adopté en particulier pour les enregistrements de métadonnées des entrepôts OAI-PMH.

Certaines applications, comme TEF, Thèses Electroniques Françaises, disposent d'un modèle à deux niveaux de hiérarchie, où les qualificatifs sont des éléments-fils des éléments de base.

---

[18] "L'espace de nom, ou namespace est le nom d'un organisme qui a mis au point un schéma XML. Un espace de nom est identifié par un identificateur uniforme de ressource (adresse URL ou nom URN) […]. Plus généralement, tout ensemble fermé de noms peut être considéré comme un espace de noms. (Glossaire RCIP, http://www.rcip.gc.ca/Francais/Normes/glossaire.html, traduit et adapté de *Dublin Core Metadata Glossary*, ébauche finale, 24 février 2001, library.csun.edu/mwoodley/dublincoreglossary.html)

5.3 XML et autres jeux de métadonnées

Beaucoup d'autres jeux de métadonnées sont implémentés en XML aujourd'hui.

5.3.1 Jeux à orientation bibliographique

Ce sont essentiellement MARC-XML, des applications UNIMARC en XML comme BiblioML, MODS et ONIX.

MARC et ses dérivés (MARC21, UNIMARC, InterMARC…) correspondent au format de référence de la description bibliographique. Nés dans les années 60 pour normaliser la description des documents et faciliter l'échange et l'exploitation de celle-ci par les ordinateurs, maintenus par la Library of Congress, ils constituent des jeux de métadonnées essentiellement descriptives, permettant un signalement extrêmement précis et standardisé des documents.

Ces formats complexes, peu lisibles par l'homme et implémentés sous de multiples versions nationales ou locales sont surtout pratiqués par les professionnels des bibliothèques - centres de documentation, et peuvent difficilement être utilisés de manière généralisée, dans le contexte d'une multitude de ressources électroniques d'origine diverses ; de plus, s'ils décrivent bien tous les objets physiques, ils sont limités par rapport à la description et la gestion complètes du document électronique.

L'utilisation de 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 [6, 7, 17]. L'implémentation des données d'autorité liées à MARC se fait au moyen du jeu d'éléments MADS, également en XML, et issu des différentes listes d'autorité liées à MARC.

Le jeu d'élément MODS, né en 2003 à l'initiative de la Library of Congress, est directement issu de MARC, dans une volonté d'adaptation à l'environnement actuel : schéma XML avec codage UNICODE, verbalisation orientée utilisateurs des noms des zones et sous-zones, proche de Dublin Core et de ONIX, éléments adaptés à la description des ressources numériques, possibilités de descriptions simples utilisant un minimum d'éléments et pouvant être complétées ultérieurement… [6]

Un exemple de description d'un même objet en MARC-XML, MODS et Dublin Core est présenté en annexe 3.

ONIX, enfin, développé par l'industrie du livre pour gérer de façon unique le flux d'information entre tous les acteurs concernés, offre à la fois une description bibliographique et physique précise de l'ouvrage et une description de sa vie sociale et économique (tirages et diffusion, campagnes de marketing, prix littéraires …). Un des objectifs d'EDItEUR est qu'il devienne le format d'entrée des bibliothèques, et il est associé à des outils de conversion ONIX vers MARC et UNIMARC.

5.3.2 Autres jeux portant sur des objets numériques simples

Hors du monde strictement bibliographique, nombreux aussi sont les jeux de métadonnées implémentés en XML.

On peut citer citer des jeux couvrant des domaines spécifiques comme FGDC, très utilisé dans les applications géospatiales, cartographiques et socio-démographiques ou LOM, d'un grand intérêt pour la description des objets éducatifs dans le cadre du e-learning. Signalons une spécificité de ce dernier, la génération en cours d'utilisation de nouvelles métadonnées sur le processus d'apprentissage (interactivité entre l'utilisateur et l'outil, enregistrement des résultats de tests).

Pour les images[15], de nombreux formats propriétaires ont été créés au cours du temps, les uns très techniques comme ceux des formats jpg, tif, gif …, d'autres plus descriptifs comme celui d'Adobe-Photoshop. Tous sont basés sur la norme IPTC, standard d'échange initialement développé pour la presse et adopté par de nombreuses applications et matériels. Les standards récents s'écrivent en XML :

- XMP, développé par Adobe-Photoshop intègre les éléments Dublin Core ;
- le standard IPTC lui-même a évolué depuis cette année vers IPTC-Core, basé sur XMP ;
- MIX, maintenu par la Library of Congress, contient une description riche des caractéristiques de l'image et de son acquisition, destiné à gérer durablement et de façon non propriétaire des collections d'images.

Les métadonnées des en-têtes de documents XML standards, TEI par exemple, permettent une description précise du document à orientation bibliothéconomique (exemple simple en annexe 4). Il s'agit surtout de gérer et d'archiver dans un fonds un document et ses évolutions, et d'échanger avec des partenaires spécifiques ; l'en-tête, partie prenante du document, est générée au moment de la création et de la publication, et, si elle est suffisamment riche, elle peut servir de source unique pour les signalements ultérieurs.

5.3.3 Description et gestion de collections

La DTD EAD est apparue il y a une dizaine d'années ; elle formalise en XML la norme ISAD(G) sur les instruments de recherche archivistiques. La description archivistique repose sur le respect des fonds et la structuration en fonction de la logique de production; elle peut avoir un grand nombre de niveaux, se fait du général au particulier avec héritage, et sans redondance d'un niveau à l'autre. Aujourd'hui, EAD décrit surtout des collections d'objets physiques ; dans le cadre de la numérisation des fonds, elle commence à décrire des collections d'objets numériques, comme cela était prévu dans sa conception [8].

On trouvera également en annexe 5, un exemple simple de la structure d'un document EAD.

Les données d'autorité peuvent être implémentées dans le format EAC, formalisé un peu plus tard, également en XML.

EAD est bien implanté dans le milieu archivistique de nombreux pays, d'autant plus que c'est le premier outil numérique standard de mise en œuvre de la norme ISAD(G) ; il est également utilisé pour certaines collections spécialisées, y compris en France.

D'autres jeux, XML toujours, sont orientés vers la description de collections d'objets numériques variés et éventuellement distribués, comme METS, ou IMS-CP. METS apporte une description physique et logique de la collection, très structurante et très opérationnelle pour la gestion, la recherche et la préservation Il fait appel à d'autres jeux d'éléments sémantiques, comme Dublin Core et MIX, pour les métadonnées portant sur l'objet de base.

5.4 Des applications

Depuis quelques années, les réalisations en matière de métadonnées se multiplient ; les exemples suivants, glanés au cours de lectures, de congrès ou d'échanges avec différents acteurs, ne prétendent pas à l'exhaustivité mais simplement à concrétiser la réalité du phénomène.

5.4.1 Archives ouvertes et protocole OAI-PMH

Le protocole OAI-PMH d'interopérabilité des archives ouvertes repose sur un enregistrement minimal Dublin Core simple en XML ; même si l'évolution est plus lente que ne l'espéraient les pionniers du mouvement, les archives ouvertes représentent une réalisation fondamentale en termes d'interopérabilité et de potentiel pour l'avenir, de stratégie alternative de publication et de diffusion, d'engagement de différents pays et institutions dans le mouvement, et sont un des grands exemples de l'efficacité de métadonnées standardisées. A côté de l'enregistrement minimal en Dublin Core, dont un exemple est donné en annexe 1, les "moissonneurs", moteurs spécialisés OAI-PMH, comme OAIster, ARC ou CiteBase peuvent intégrer d'autres enregistrements : MODS, MARC-XML et METS.

5.4.2 Bibliothèques et patrimoine culturel

Les projets suivants sont souvent nés d'un constat : l'existence de nombreux services en lignes produits par des organismes proches dans leurs missions, ou parfois par le même organisme, services multiformes dans leurs conceptions comme dans leurs signalements, et cela dans des contextes d'utilisation de plus en plus variés et transversaux ; d'où des besoins d'homogénéiser l'accès aux services, ou au moins leur description pour faciliter la génération d'interfaces communes, qui seront déclinées selon les besoins, nationales ou internationales, locales ou thématiques : archives ouvertes, portails, catalogues, "inventaires" ; quand ces services sont gérés par le même organisme, s'y ajoute un objectif de simplification de la maintenance.

Ces projets visent l'interopérabilité, et notamment la compatibilité OAI-PMH ; ils utilisent tous le jeu Dublin Core, simple ou qualifié, et empruntent à d'autres jeux standards cités auparavant, MARC (et dérivés) en XML ou MODS, parfois METS ou EAD, soit pour créer d'autres enregistrements, soit dans un seul enregistrement, selon un "profil d'application" particulier ; ils incluent quelques éléments locaux répondant à des besoins spécifiques, notamment de gestion.

Il faut d'abord citer les projets et réalisations de bibliothèques ou groupes de bibliothèques, de l'American Memory, réalisé par la Library of Congress, qui contient plus de 5 millions de ressources au projet TEL de Bibliothèque Numérique Européenne, en passant par le portail et

système d'information ABES-SUDOC, les réalisations de la Bibliothèque Nationale de France sur certains fonds, le projet global de la British Library, ou ceux d'autres pays, dont les publications commencent à faire état, bibliothèques sud-américaines, japonaises …

Il faut citer ensuite les projets (et réalisations) autour de la numérisation et de la mutualisation du patrimoine culturel, qui représentent une activité importante : projet RLG[19] aux Etats-Unis, projet français de mutualisation du patrimoine culturel numérisé[20] et projets d'autres pays européens, fédérés aujourd'hui dans le projet européen MICHAEL[21].

5.4.3 D'autres exemples

Le projet TEF de thèses électroniques françaises[22] a pour objectif de valoriser les thèses en augmentant leur diffusion et de créer une chaîne éditoriale unique. Les métadonnées doivent respecter la richesse de description actuelle, qui a fait ses preuves, tout en étant interopérables, pour des applications hors des bibliothèques. Elles intègrent les éléments Dublin Core qualifié complétés par des éléments MODS et ETD-MS (projet européen de description des thèses), ainsi que des éléments propres portant sur le suivi administratif et le statut juridique des thèses ; une feuille de style permet de générer un enregistrement OAI-PMH ou UNIMARC.

Quelques éditeurs de revues en accès libre (sur des sites universitaires ou personnels) ont une exigence de standardisation du contenu et de la description, et cela permet d'automatiser l'intégration des collections dans une archive ouverte, par exemple.

Un dernier exemple, enfin : en peu de temps, tous les départements d'un organisme comme l'INIST ont été impliqués dans la production de services et de "notices" intégrant des métadonnées standardisées, ou par des travaux terminologiques dépassant le cadre de l'Institut.

6. Identifiants uniques et pérennes

Dans un contexte d'interopérabilité des ressources, la notion d'URI, un identifiant unique et pérenne définissant chaque d'entre elles, prend tout son intérêt. Elle mérite donc un petit détour avant de parler de RDF, où elle est fondamentale.

Au-delà de la simple URL, insuffisamment stable, on recense différentes initiatives d'URI, qui impliquent la mise en oeuvre de registres et/ou de "résolveurs de liens" :

- le PURL, pour Persistent URL, géré par l'OCLC,
- les ISSN, ISBN électroniques

---

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

[20] Sur le projet : Numérisation du patrimoine culturel, informations techniques, http://www.culture.gouv.fr/culture/mrt/numerisation/fr/technique/technique.htm
Sur les réalisations : Portail culture.fr http://www.culture.fr
Catalogue des fonds culturels numérisés, http://www.culture.gouv.fr/culture/mrt/numerisation/fr/f_02.htm

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

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

- le système Handle, utilisé, entre autres, par les archives ouvertes sous DSpace et par les éditeurs pour les DOI (Digital Object Identifier),
- l'URN, Uniform Resource identifier Name, standard du W3C pour la normalisation des espaces de nom, qui permet d'intégrer de manière non ambiguë les différents autres types d'URI dans une description.

Le projet ARK[23], Archive Resource Key, permet de relier la ressource, ses métadonnées et les engagements de l'éditeur sur la pérennité, et automatise la recherche de la localisation actuelle sans registre, à partir de l'identifiant seul ; cette initiative semble intéressante, mais encore peu implémentée, au moins en Europe.

7. RDF

7.1 Principes et intérêt

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 un ensemble de professionnels très large (communautés des bibliothèques, des standards du Web, 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, et donc réellement interopérable [9].

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 de la ressource, ses propriétés et les valeurs correspondantes [4, 5, 19], cadre qui accueille tout jeu sémantique existant. RDF Schema[24] y ajoute un niveau de structure supplémentaire en définissant un certain nombre de propriétés et de classes d'objets pouvant être utilisés dans les descriptions. 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 [9].

La recherche et l'échange d'information sont ainsi facilités, et des applications existent aujourd'hui , avec des annuaires comme Dmoz.org ou CISMef[25], le grand répertoire mondial du Web, ou des bases de données musicographiques ou filmographiques[26] [9].

Dans ce contexte, certains producteurs de jeux de métadonnées actuellement implémentés en XML, ont déjà créé un schéma spécifique RDF utilisable (Dublin Core par exemple), ou le développent pour leurs versions futures (LOM par exemple). Les formats récents de métadonnées sur les ressources audiovisuelles, comme MPEG7 et MPEG21 sont également écrits en RDF.

---

[23] Voir "A Dozen Primers on Standards", Computer Libraries, 24, 2, février 2004, http://www.infotoday.com/cilmag/feb04/primers.shtml

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

[25] Annuaire mondial collaboratif DMoz, Open Directory Project, http://www.dmoz.org
Catalogue et Index des sites Médicaux Francophones CISMef, http://www.chu-rouen.fr/cismef/

[26] Par exemple Gracenote, www.gracenote.com et IMDb, www.imdb.com

De plus, ce langage permet de relier structurellement les descriptions – ou parties de descriptions - de ressources entre elles, grâce leurs URI. Par exemple, toute œuvre peut ainsi être reliée à son contexte de production et d'usage, à l'ensemble des descriptions produites ultérieurement à son sujet, aux outils terminologiques du domaine, et la masse d'information disponible sur le Web prend alors un sens structurel et cognitif.

La puissance de RDF a suscité la création de jeux sémantiques complémentaires, indispensables pour mettre en œuvre toutes les fonctions des métadonnées.

7.2 Des jeux de métadonnées et des applications RDF

7.2.1 Descriptions de personnes

Les vCard[27], cartes de visite électroniques standardisées par le W3C, sont utilisées pour authentifier un document ou une transaction, et implicitement par beaucoup de logiciels de messagerie.

Le projet FOAF[28], centré sur les personnes, leurs réseaux et leurs œuvres, 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 association".

7.2.2 Droits des oeuvres

Les licences Creative Commons[29] représentent une alternative au copyright pur en permettant de définir des droits d'utilisation diversifiés. Cette réalisation, initiée par un éminent professeur de droit américain pour le partage d'œuvres artistiques au départ, et lancée officiellement en France en 2004, a débouché en 2005 sur le projet Sciences Commons, qui doit définir des cadres d'application et des licences spécifiquement adaptées au contexte des ressources scientifiques.

Le site officiel, http://creativecommons.org, propose par simples formulaires, d'une part la recherche d'œuvres en fonction des droits, d'autre part le choix de différentes licences d'usage ; la licence sélectionnée est présentée en texte, langage simple et juridique, et en RDF ; il reste à la relier à l'œuvre concernée.

Pour les enregistrements OAI-PMH, en XML simple jusqu'à récemment, l'Open Archive Initiative recommande depuis décembre 2004[30] l'intégration de licences Creative Commons en RDF dans l'élément Dublin Core "rights" (voir un exemple en annexe 6).

7.2.3 Ontologies et taxonomies

Les ontologies sont des représentations dynamiques des connaissances dans un domaine ; elles permettent de définir des classes ou concepts (nantis de propriétés), des objets ou termes

[27] Representing vCards objects in RDF, http://www.w3.org/TR/vcard-rdf
[28] The Friend of a Friend (FOAF) Project, http://www.foaf-project.org/

[30] http://www.openarchives.org/OAI/2.0/guidelines-rights.htm

appartenant à ces classes, et d'implémenter des relations riches et diverses entre ces objets. 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 de RDF, comme Topics Maps ou SKOS.

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 [31].

7.2.4 RSS[32] et syndication de sites

La syndication est une technique permettant d'afficher automatiquement dans une page 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 en général leurs nouveautés, fils que l'utilisateur peut visualiser dans un lecteur (ou agrégateur) de RSS[33], ou dans un navigateur comme Mozilla Firefox. Cette technique permet donc de suivre facilement l'évolution des sites disposant de canaux et, à terme, pourrait remplacer un certain nombre de listes de diffusion [2, 16].

Le nombre des canaux RSS disponibles augmente très rapidement actuellement, et leur existence sur un site est d'ores et déjà un critère de qualité chez nos voisins anglo-saxons.

Il existe deux grandes versions de RSS, toutes deux XML, l'une issue de RDF et l'autre de syntaxe plus simple, chacune ayant évolué avec le temps ; toutes deux sont proches dans le principe et la syntaxe, et lisibles par les outils d'affichage RSS cités.

L'implémentation reste en évolution :

- un autre schéma, Atom, a été développé en 2004, et vise à unifier la syntaxe et à donner des possibilités supplémentaires de description de ressources non textuelles.

- des modules complémentaires aux schémas RSS de base, déjà utilisés par divers communautés ou organismes, sont en cours de développement pour répondre à des besoins particuliers de description et de manipulation de documents spécifiques, audio, images, video[34].

7.3 Perspectives

Après le temps du "tissage" et de la croissance du Web, des premières pages statiques aux frames et aux pages dynamiques diverses, impliquant divers protocoles additionnels et évolutifs pour améliorer l'interopérabilité, arrive le temps du "démêlage" par la représentation et le lien structurel, et le consortium W3C soutient fortement RDF pour le Web du troisième millénaire.

---

[31] Pour comprendre les principes des Topics Maps et explorer la démonstration, http://www.ontopia.net/, onglet 'topics Maps'

[32] RSS : à l'origine "RDF - Resource Description Framework - Site Summary", souvent maintenant "Real Simple Syndication", parfois "Rich Site Summary", développé pour la syndication de sites

[33] Un exemple d'agrégateur libre : Bloglines, http://bloglines.com

[34] Dans le cadre de Podcasting, voir http://en.wikipedia.org/wiki/Podcasting ou Media RSS, voir http://tools.search.yahoo.com/mrss/mrss.html et http://tools.search.yahoo.com/mrss/

L'adoption de RDF à large échelle demande cependant un "saut technologique" important de la part des acteurs du Web, car ce format est complexe. Si des outils d'implémentation et/ou de recherche de fichiers RDF existent depuis plusieurs années pour des applications spécifiques, ils commencent seulement à se développer dans un cadre plus générique[35], et demandent encore beaucoup de compétences techniques ; la question reste donc ouverte…

8. Implémenter des métadonnées

8.1 Une démarche projet

Pour aller vers l'interopérabilité et la durée, chaque organisme professionnel peut, et doit, se poser la question d'un signalement standardisé des ressources d'information dont il a la charge.

Cependant, c'est une activité consommatrice de temps et de moyens humains ; il ne s'agit pas de suivre un effet de mode, mais bien d'une démarche de gestion de projet dans le cadre du système d'information de l'organisme.

Il faut d'abord s'interroger sur les objectifs visés, en terme d'environnement, (système d'information, projets et missions de l'établissement, projets voisins), de types de ressources, de besoins d'usage de ces métadonnées, pour choisir parmi les standards existants ou créer un profil d'application spécifique.

Il faut ensuite définir la méthode d'implémentation et de validation, choisir les outils adaptés, et définir les acteurs impliqués dans la chaîne.

L'implémentation de métadonnées peut se faire de deux manières :

- par la création de nouvelles descriptions, au travers par exemple de l'en-tête d'un document TEI ou de formulaires de soumission dans une archive ouverte ; dans les acteurs potentiels, il ne faut pas oublier le créateur de la ressource, qui la connaît mieux que quiconque, mais l'expérience montre, d'une part qu'il n'est pas toujours motivé par cette activité, d'autre part qu'un professionnel de l'IST garde un rôle important pour compléter et valider la description.
- surtout, elle peut être effectuée automatiquement et en temps réel à partir du cœur de l'activité documentaire de chaque organisme, par conversion des données dont elle dispose pour d'autres usages.

8.2 Des outils

Libres ou commerciaux, ils sont de plus en plus nombreux, mais ne dispensent pas toujours d'une application "maison" . On en distingue quatre classes [14] :

- des éditeurs de texte spécialisés, HTML et XML, qui possèdent des fonctions d'aide à la saisie et à la validation,

[35] Par exemple, Redland RDF Applications Framework, http://librdf.org/

- des formulaires de saisie ; soit présents sur le Web (soumission dans les archives ouvertes et annuaires collaboratifs, génération de licences Creative Commons…), soit implémentés localement,
- des outils de conversion depuis un signalement existant, principalement scripts et feuilles de style XSL : UNIMARC vers MARC, MARC - MODS - EAD vers Dublin Core sur le site de la Library of Congress[36].; EAD vers Dublin Core[37] sur le site d'AJLSM; TEF vers UNIMARC et enregistrements OAI-PMH sur le site de l'ABES ; ONIX vers MARC et UNIMARC sur le site EDItEURS...

  Il faut souvent adapter ces outils ou en créer de nouveaux, en fonction des spécificités des données disponibles, d'objectifs et de choix particuliers d'équivalences.
- des outils d'extraction des métadonnées à partir du contenu de la ressource, surtout sur des documents structurés XML, à nouveau essentiellement scripts et feuilles de style.

Le site officiel Dublin Core, rubrique "Tools and Software" propose des exemples de chaque type d'outil, intéressants dans un but de démonstration et de formation.

Les plate-formes éditoriales libres qui génèrent des documents structurés en XML, comme Lodel[38] ou SDX[39], possèdent en général plusieurs de ces fonctions, et créent entre autres des métadonnées compatibles OAI-PMH. Les Systèmes de gestion de Contenu, *CMS ou Content Management Systems*, comme Spip, proposent une édition de fil RSS automatique, et un fichier de métadonnées qui sera intégré dans les différentes pages du site.

Progressivement, enfin, dans des outils commerciaux de Gestion Electronique de Documents (GED) orientés création de portails, comme Archimed en France, apparaissent des modules qui permettent d'une part de créer des enregistrements compatibles OAI-PMH, d'autre part de "moissonner" les entrepôts OAI-PMH du Web.

Il reste encore beaucoup à faire, sur un plan technique, mais aussi conceptuel (la sémantique des éléments, la granularité, et les valeurs recommandées différent souvent entre jeux) ; le développement d'outils intégrés et modulaires de création et de conversion de métadonnées est, selon le NISO, un challenge pour réussir la montée en charge de métadonnées standardisées dans les années à venir [14]. Le prototype ERROLs de l'OCLC est un premier outil de ce genre : implémenté en METS, accessible au travers d'un Web Service, il recense (et donne accès à) un certain nombre de tableaux d'équivalence et d'outils de conversion, pour l'homme ou la machine, dans un souci du contexte : lien avec les jeux source et cibles dans leurs

[36] Voir Library of Congress, standards : http://www.loc.gov/standards, pages MARC, MODS, EAD (tableaux de correspondances et feuilles de style)
Pour des feuilles de style METS vers IMS-CP, voir http://iu.berkeley.edu/creatingcontent/ par exemple

[37] AJLSM est une société d'études, de développement et de formation autour de l'information numérique Ces outils ont été réalisés dans le cadre du projet de numérisation des ressources du Centre Historique des Archives Nationales (CHAN), et sont intégrés par ailleurs dans la plate-forme PLEADE, commercialisée par AJLSM ; voir http://projets.ajlsm.com/ead-chan/

[38] LODEL CMS, logiciel d'édition électronique, http://www.lodel.org/

[39] SDX, un moteur de recherche et un environnement de publication pour documents XML, http://adnx.org/sdx/

différentes versions, équivalences assumées par une communauté particulière, extensibilité du contenu[40].

9. Conclusion

Les réflexions et expériences autour des métadonnées datent d'une bonne dizaine d'années ; aujourd'hui, le paysage est riche, bien balisé en standards et applications, et commence à se stabiliser. Dublin Core intervient comme cœur d'interopérabilité et les autres jeux d'éléments standardisés sont indispensables pour répondre à tous les besoins. Des passerelles et outils existent déjà, qui ne cessent de se développer, en quantité et en qualité.

De "nouvelles" questions peuvent maintenant être traitées :

- celle d'une utilisation des éléments de métadonnées homogène entre producteurs de données, conforme à des guides de bonne pratique, et celle des valeurs données aux éléments utilisés, pour une réelle interopérabilité ; mais cette problématique est bien connue des professionnels bibliothèque-documentation ...
- celle de la pérennité des données, et de la pérennité de leur accessibilité ; des projets importants ont démarré sur tous les continents[41], et suivent le modèle OAIS, Open Archive Information Systems, normalisé par l'ISO.

Pour les organismes gérant des ressources électroniques, un projet d'interopérabilité par des métadonnées standardisées est non seulement un gage de visibilité et de normalisation, mais aussi une occasion de se former progressivement aux technologies et pratiques actuelles, avec la certitude que, même si les standards évoluent, un signalement complet réalisé en format structuré aujourd'hui sera évolutif et utilisable demain.

Bibliographie

1 - Chartron G, "Normes et standards du document numérique", janvier 2000, http://web.ccr.jussieu.fr/urfist/presse/standard/coursintro.htm

2 - Cottin S., "Tout savoir sur le RSS", septembre 2003, http://www.servicedoc.info/article.php3?id_article=125

[40] ERROLs : Extensible Repository Resource Locators
Voir : Golby C-J., Young J-A., Childress E., "A Repository of Metadata Crosswalks", D-Lib Magazine, décembre 2004, http://www.dlib.org/dlib/december04/godby/12godby.html
Et une version de démonstration : http://errol.oclc.org/schemaTrans.oclc.org.html

[41] Notamment :
- PREMIS (OCLC/RLG) et Digital Preservation Coalition (OCLC/ Université d'Oxford)
- En Europe, dans le cadre du réseau d'excellence sur les bibliothèques numériques DELOS
- Au Royaume Uni, le projet CEDAR
- En Australie, le projet de la National Library

3 - Desrichard Y.,“Vers la convergence des formats bibliographiques ? ONIX, application XML du monde de l’édition”, Bulletin des Bibliothèques de France, 2004, t. 49, n° 3, p. 55-63,
http://bbf.enssib.fr/bbf/html/2004_49_3/2004-3-p55-desrichard.xml.asp

4 - Dubost K., “RDF et les métadonnées”,
http://www.w3.org/2000/Talks/1019-rdf-poly/Overview.html

5 - Dubost K., “Spécification du modèle et la syntaxe du cadre de description des ressources (Resource Description Framework ou RDF)”, 2002 ,
http://www.la-grange.net/w3c/

6 - 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]

7 - 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]

8 – 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

9 - 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

10 - Giuliani E., “Les métadonnées : de la convergence à la normalisation”, Dossiers de l'audiovisuel, 2000, n°. 93, p. 29-31

11 - Hillmann D., “Using Dublin Core”, août 2003,
http://dublincore.org/documents/2003/08/26/usageguide/

12 - Jacquet C., Dublin Core et les métadonnées,
http://www.openweb.eu.org/articles/dublin_core/

13 - Lupovici C., “Identification et metadonnees : diversite des standards”, Information, documentation, transfert des connaissances, 1999, p. 184-190

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

15 - Peccatte P./Soft Experience,"Métadonnées: une initiation - Dublin Core, IPTC, EXIF, RDF, XMP", mars 2004,
http://peccatte.karefil.com/software/Metadata.htm

16 - 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

17 - Roumieux O., “La famille Marc : les métadonnées des bibliothécaires”, Archimag, 2001, n°. 143, p. 36-38

18 - Sibille C., la DTD EAD (Encoded Archival Description), avril 2004,
http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html

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

20 - Wikipedia, the free encyplopedia, recherche metadata, http://en.wikipedia.org/wiki/Metadata

Annexes

Annexe 1 : Dublin Core simple dans un enregistrement OAI-PMH

```
<record>
    […] <header>
            <identifier>oai:archivesic.ccsd.cnrs.fr:sic_00000025</identifier>
            <datestamp>2002-05-17</datestamp>
            <setSpec>sic_docu</setSpec>
            […]
        </header>
        <metadata>
            <oai_dc:dc
xmlns:oai_dc=http://www.openarchives.org/OAI/2.0/oai_dc/
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:xsi=http://www.w3.org/2001/XMLSchema-instance
xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/
http://www.openarchives.org/OAI/2.0/oai_dc.xsd">

<dc:creator>Gallezot, Gabriel</dc:creator>
<dc:creator>Chartron, Ghislaine</dc:creator>
<dc:creator>Noyer, Jean-Max</dc:creator>
<dc:description>Dans une démarche constructive, […] communauté.</dc:description>
<dc:title>Une archive ouverte des publications en InfoCom</dc:title>
<dc:language>fr</dc:language>
<dc:subject>Information retrieval</dc:subject>
<dc:subject>Electronic publishing</dc:subject>
<dc:subject>Knowledge management</dc:subject>
<dc:subject>Hypertext, hypermedia</dc:subject>
<dc:date>2002-03-25</dc:date>
<dc:identifier>http://archivesic.ccsd.cnrs.fr/sic_00000025.en.html</dc:identifier>
<dc:type>Text</dc:type>

</oai_dc:dc></metadata></record>
```

Source : @rchiveSIC, archive en Sciences de l'Information et de la Communication gérée par le CCSD, Centre pour la Communication Scientifique Directe

Annexe 2 : Dublin Core qualifié : l'élément "relation" comme exemple

Les ressources sur les équivalences UNIMARC–Dublin Core qualifié sont rares, contrairement aux équivalences vers le Dublin Core simple, et aucune à ce jour n'émane d'un organisme normalisateur. A titre d'illustration, l'élément "relation", le plus riche, possède treize qualificatifs, à remplir de préférence par valeurs de type URI ; le tableau suivant établit des correspondances pour douze d'entre eux avec des zones UNIMARC appartenant principalement au "bloc des liens" 4XX , avec la sous-zone $u pour des valeurs de type URI [42].

Qualificatif Dublin Core	Equivalence UNIMARC
isFormatOf La ressource décrite a le même contenu intellectuel que la ressource référencée ici, mais un **format/support** différent, et lui est postérieure	452, Autre édition, support différent
hasFormat Réciproque : la ressource est décrite antérieure à ressource référencée ici, a le même contenu et un **format/support** différents	452, Autre édition, support différent
isversionOf La ressource décrite est une version, édition, adaptation de la ressource référencée ici (modification du **contenu intellectuel**)	451, Autre édition , même support 454 $a, Traduit de 455 $a, Reproduction de
hasVersion Réciproque : la ressource décrite a d'autres versions, éditions ou adaptations de **contenu** différent	451, Autre édition, même support 453, Traduit sous le titre 456, Reproduit comme
isReplacedBy La ressource est caduque et remplacée par la ressource référencée ici	442, Remplacé par et autres 44X pour séries continuées par (suite, absorption…)
Replaces Réciproque : la ressource remplace la ressource référencée	432, Remplace Autres 44X pour séries (suite, absorption…)
requires La ressource a besoin de la ressource référencée pour être	488, autres oeuvres liées,

[42] Source :
UNIMARC – Dublin Core, Bibliothèque Nationale de Russie,
http://www.rba.ru:8101/rusmarc/soft/dc_unimarc.html
Les métadonnées des theses électroniques françaises,
http://www.abes.fr/abes/documents/tef/recommandation/tef_01.pdf
Manuel UNIMARC, format bibliographique, 4° edition, version française, UBCIM Publications, 2002

correctement présentée, transmise, ou pour assurer sa cohérence	337 $aRequires (ou 311 $aRequires) : Notes, Configuration requise
isRequiredBy Réciproque : la ressource est nécessaire à l'interprétation, l'exécution, la transmission de la ressource référencée ici	488, autres oeuvres liées 311$aIs required by , Note sur les zones de lien
isPartOf La ressource fait partie, physiquement ou logiquement, de la ressource référencée ici	461, Niveau de l'ensemble 462,Niveau du sous-ensemble
hasPart Réciproque : la ressource décrite contient, physiquement ou logiquement, la ressource référencée	463, Niveau de l'unité matérielle
references La ressource décrite fait référence ou points sur la ressource référencée ici	488, Autres oeuvres liées 321$aAccompagnies , Note sur les index, extraits et citations publiés séparément
isReferencedBy Réciproque : la ressource décrite est citée ou pointée par la ressource référencée ici	488, Autres oeuvres liées 321$aIs referenced by , Note sur les index, extraits et citations publiés séparément

NB : le dernier qualificatif de relation, conformsTo, fait référence à un standard auquel la ressource est conforme et n'a pas d'équivalent UNIMARC.

Annexe 3 : MARC-XML, MODS et Dublin Core qualifié

Cette présentation de quelques éléments (identifiant, titre, auteur, description et sujet/mots-clés…), a pour but de visualiser succinctement l'écriture de chaque format, et ne prétend aucunement rendre compte des spécificités, richesses et limites de chacun.

(en italique et entre < !-- -->, sont des commentaires d'aide à la lecture)

MARC-XML

```
<!-- espace de nom : répertoire accessible en ligne contenant le schéma XML -->
<collection xmlns="http://www.loc.gov/MARC21/slim">

  < -- un enregistrement de métadonnées sur une ressource -->
   <record>
     (…)
   < !-- ISSN, et disponibilité-->
      <datafield tag="020" ind1=" " ind2=" ">
         <subfield code="a">0152038655 :</subfield>
         <subfield code="c">$25.95</subfield>
      </datafield> (…)
   <!-- createur (personne) et information associée -->
      <datafield tag="100" ind1="1" ind2=" ">
           <subfield code="a">Saint Exupery, Antoine</subfield>
           <subfield code="d">1900-1944.</subfield>
      </datafield> (…)
   <!-- titre principal, et information associée -->
      <datafield tag="245" ind1="1" ind2="0">
         <subfield code="a">Le Petit Prince</subfield>
      </datafield>
   <!-- titre, variantes -->
      <datafield tag="246" ind1="3" ind2="1">
         <subfield code="a">Little Prince</subfield>
      </datafield> (…)
   <!-- notes générales -->
      <datafield tag="500" ind1=" " ind2=" ">
         <subfield code="a">CD-Rom inclus</subfield>
      </datafield>
   <!-- résumé de type 'review'-->
      <datafield tag="520" ind1=" " ind2=" ">
        <subfield code="a">Enfance, poesie.(…) dans le désert.</subfield>
      </datafield> (…)
   <!-- mots-clés, LCSH -->
```

```
        <datafield tag="650" ind1=" " ind2="0">
            <subfield code="a">Aventure</subfield>
            <subfield code="x">Juvenile poetry.</subfield>
        </datafield>…
    </record> (…)
</collection>
```

MODS, Metadata Object Description Standard

```
< !-- élément racine et espace de nom -->
<mods xsi:schemaLocation="http://www.loc.gov/mods/
http://www.loc.gov/standards/mods/mods.xsd">
  <!-- element titre -->
      <titleInfo>
          <title>Le Petit Prince</title>
          <title type="translated" xml:lang="en">Little Prince</title>
      </titleInfo>
  <!-- element auteur, type personne -->
      <name type="personal">
           <namePart>Saint Exupery, Antoine</namePart>
           <namePart type="date">1900-1944</namePart>
              <role>
                 <text>creator</text>
              </role>
      </name> (…)
  < ! -- éléments résumé et note -->
       <abstract> Enfance, poesie.(…) dans le désert </abstract>
       <note>CD-ROM inclus</note>
  < !-- élément audience, public visé -->
        <targetAudience>jeunesse</targetAudience>
  < !-- mots-clés type LCSH -->
        <subject authority="lcsh">
            <topic>Aventure</topic>
            <topic>Juvenile poetry</topic>
        </subject>…
  < !-- identifiant, type ISBN -->
      <identifier type="isbn">0152038655</identifier>
    (…)
</mods>
```

Dublin Core qualifié

```
< !-- élément racine, avec espaces de noms -->
<metadata xmlns:dc=http://purl.org/dc/elements/1.1/
          xmlns:dcterms=http://purl.org/dc/terms/
          xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">

  < !-- élément et qualificatif titre -->
    <dc:title>Le Petit Prince</dc:title>
    <dcterms:alternative xsl:lang="en">Little Prince</cdterms:alternative>
  <!-- créateur -->
    <dc:creator>Saint Exupery Antoine,  1878-1967</dc:creator>
  < !-- résumé, qualificatif de l'élément 'description' -->
    <dcterms:abstract> Enfance, poesie.(…) dans le désert</dcterms:abstract>
   < !-- format, type MIME si possible -->
    <dc:format>applis/pdf</dc:format>
   < !-- public visé -->
    <dcterms:audience>jeunesse</dcterms:audience>
   < !-- sujet, mots-clés, de type LCSH -->
    <dc:subject xsi:type="dcterms:LCSH">Aventure</dc:subject>
    <dc:subject xsi:type="dcterms:LCSH">Junevil poetry</dc:subject> (…)
   < -- identifiant de type URI -->
    <dc:identifier xsi :type="URI">http://[URL]</dc:identifier>
   < !-- lien vers la version imprimée -->
    <dcterms :isVersionOf
xsi:type="ISBN">0152038655</dcterms :IsVersionOf>
      (…)
</metadata>
```

Annexe 4 : en-tête TEI[43]

TEI est un modèle de document XML initialement développé par des linguistes. Cette discipline implique un niveau d'exigence élevé dans la structuration fine des constituants logiques du document, et la DTD TEI est un outil utilisé dans de nombreux domaines aujourd'hui. L'en-tête du document, teiHeader, contient un ensemble de métadonnées de description et de gestion.

```
<teiHeader status="new" type="text">

 <!-- description bibliographique du fichier -->
    <fileDesc>
    <!-- titre du document, auteur, responsabilité -->
      <titleStmt>
        <title>Introduction</title>
        <author>
          <name>Joëlle Weill</name>
        </author>
      </titleStmt>
      (…)
    < !-- taille du fichier -->
      <extend>(taille)</extend> (…)
     </fileDesc>

<!--"profil", informations sur le contenu  -->
     <profileDesc>
    <!-- langue -->
        <langUsage default="NO">
          <language id="fr-FR">ISO fr-FR</language>
        </langUsage>
    <!-- mots-clés -->
        <textClass default="NO">
          <keywords>
            <term>
              <title>Conserver, restituer, créer</title>
              <ref target="weill" targOrder="U"/>
            </term>
            <term>
              <title>Le jardin et le temps</title>
              <ref target="weill2" targOrder="U"/>
            </term>
            (…)
          </keywords>
```

[43] Source : "Séminaire Barbirey – Quel avenir aujourd'hui pour les jardins anciens", http://www.seminairebarbirey.culture.gouv.fr/

```
                    </textClass>
          </profileDesc>
          <encodingDesc><!-- Objectifs, principes et choix techniques de l'encodage
-->
          (…)
    <!-- modifications apportées au texte -->
          <revisionDesc> <!-- suite d'éléments <change> décrivant la date, le
responsable et le type de la modification  -->
          </revisionDesc>
  </teiHeader>
```

Annexe 5 : document EAD [8]

Le document EAD comporte 2 segments :

- un ensemble de métadonnées issu de TEI portant sur le document EAD lui-même,

- un ensemble d'information sur le contenu de l'unité documentaire décrite (fonds, collection, …) ; les niveaux hiérarchiques de description différents s'imbriquent et ont souvent les mêmes balises.

```
< !-- elements obligatoires -->
<ead>
        < !-- en-tête du document EAD, métadonnées TEI -->
        <eadheader>
                <eadid>        < !-- identifiant du fichier  -->
                <filedesc>        < !-- description du fichier -->
                        <titlestmt> < !-- mentions de titre dont auteur du fichier -
->
                                <titleproper>   < !-- titre propre -->
        < !-- description de l'unité, fonds, collection…,avec attribut de niveau -->
        <archdesc>
                <did>            < !-- identification et description de l'unité -->
```

```
<archdesc> contient souvent de plus :
                <bioghist>       < !-- biographie/histoire institutionnelle -->
                <scopecontent>   < !-- présentation du contenu  -->
                <controllaccess>     < !-- points d 'accès contrôlé -->
                <dsc> < !-- description des composants, regroupe des éléments
<c> par exemple -->
                        <c> < !-- description de composants logiques de niveaux
différents définis grâce à un attribut ; élément récursif jusqu'à l'item de base,
contenant toujours un <did> -->
 … et <did> < !-- contient souvent : -->
                <repository>      < !-- organisme responsable de l 'accès
intellectuel  -->
                <origination>     < !-- origine  -->
                <unittitle>       < !-- intitulé  -->
                <unitdate>        < !-- date  -->
                 <unitid>         < !-- identifiant  -->
                <physdesc>       < !-- description physique -->
                <abstract>        < !-- résumé -->
                <physloc>         < !-- localisation physique -->
                <container>       < !-- unité de conditionnement -->
                <dao>             < !-- objet archivistique numérique  -->
```

Annexe 6 : RDF, un exemple

Exemple d'intégration d'une licence Creative Commons dans l'élément Dublin Core "rights" d'un enregistrement OAI-PMH[44].

Jusqu'à 2004, le schéma OAI-DC, lié au protocole OAI-PMH, prévoyait uniquement une structure XML simple, et la valeur de l'élément "rights" était une chaîne de caractère, mention de copyright ou URI sur le détenteur des droits. Fin 2004, une recommandation demande de renseigner cet élément avec une licence Creative Commons en RDF.

La licence suivante permet la reproduction, la distribution et la création de travaux dérivés de l'œuvre, et impose pour ces usages la déclaration du nom du créateur et l'apposition de la même licence.

```
<!-- element OAI-DC 'rights', espaces de noms et schémas des sous-éléments  -->
 <rights xmlns="http://www.openarchives.org/OAI/2.0/rights/"
     xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
     xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/rights/
       http://www.openarchives.org/OAI/2.0/rights.xsd">

  <!-- element décrit dans le  schema rights.xsl  -->
  <rightsDefinition>
        <!-- element RDF, espace de noms RDF CC -->
         <rdf:RDF xmlns="http://web.resource.org/cc/"
          xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
               <!-- elements Creative Commons - RDF -->
               <License rdf:about="http://creativecommons.org/licenses/by/2.0/">
         <permits rdf:resource="http://web.resource.org/cc/Reproduction"/>
         <permits rdf:resource="http://web.resource.org/cc/Distribution"/>
          <requires rdf:resource="http://web.resource.org/cc/Notice"/>
          <requires rdf:resource="http://web.resource.org/cc/Attribution"/>
          <permits rdf:resource="http://web.resource.org/cc/DerivativeWorks"/>
                 </License>
             </rdf:RDF>
       </rightsDefinition>
 </rights>
```

[44] Source : http://www.openarchives.org/OAI/2.0/guidelines-rights.htm