



<http://www.adbs.fr/metadonnees-mutations-et-perspectives-46545.htm>

Présentation de l'ouvrage

Alors que la majorité des ressources documentaires sont maintenant en ligne, la question de l'accès à ces textes, sons, images et données se pose de façon toujours plus aiguë. Pour que la recherche d'information gagne en pertinence et en précision, pour que l'accès aux ressources numériques soit facilité, les index, thésaurus, taxonomies, ontologies et autres formes de langages documentaires coexistent dans un web qui devient de plus en plus sémantique.

Si le terme métadonnée s'est imposé ces dernières années, il ne s'agit pas simplement d'un glissement de vocabulaire. Créées par des humains (auteurs du document ou médiateurs) ou des machines, les métadonnées permettent de décrire, mais aussi de structurer et d'organiser un document et l'information qu'il contient. La notion même de document en est bouleversée.

Les mutations récentes et les perspectives d'évolution de ces métadonnées constituaient le thème du séminaire « IST et informatique » proposé par l'INRIA en 2008 pour faire le point sur ce qui constitue le cœur de métier des spécialistes de l'information et de la documentation : la description des documents et la représentation des connaissances ; et pour s'interroger sur l'impact des changements en cours sur leurs pratiques et leurs métiers.

Représentation et accès à l'information : transformation à l'œuvre

Sylvie Dalbin

Assistance et techniques documentaires (ATD)



Ces dernières années ont été marquées par un intérêt grandissant pour les métadonnées et l'utilisation d'outils documentaires associés (vocabulaires contrôlés, classifications) qui paraissent appartenir aux traditions du secteur des archives, musées, bibliothèques ou services documentaires, ou « secteur de l'info-doc » dans la suite de ce chapitre. Largement porté par de nouveaux acteurs opérant en dehors de nos cercles professionnels, cet engouement peut nous laisser penser à un simple élargissement de l'audience de ces métadonnées ou de leurs usages dans la Société de l'Information.

Nous savons aujourd'hui qu'il n'en est rien et que nous avons à conduire des transformations, parfois profondes, de nos méthodes et des outils et pratiques métiers qui en découlent. Mais, au-delà des aspects fonctionnels ou techniques, c'est notre regard même sur les documents et les utilisateurs de ces ressources qui doit évoluer.

Ce premier chapitre va nous permettre d'introduire le phénomène des métadonnées, apparu de façon visible au sein du secteur de l'info-doc au début des années 1990. Nous tenterons d'identifier, à travers quelques exemples, sur quels plans techniques ou fonctionnels nos méthodes, pratiques ou outils traditionnels doivent à la fois être étendus sur l'axe des contenus des ressources et des usages, et transformés dans leurs fondements par rapport à une approche strictement centrée sur la notice bibliographique. Car, même si les missions ou les finalités restent identiques ou si celles-ci s'étoffent pour répondre à de nouveaux publics et de nouveaux usages, il semble impossible d'appliquer nos méthodes les plus traditionnelles à ce nouveau contexte.

La première partie de ce chapitre présente la notion et le processus de création de métadonnées qui structurent aujourd'hui les applications et les dispositifs d'information. La deuxième partie fournit aux praticiens des pistes pour faire évoluer le regard qu'ils portent sur les métadonnées, afin d'exploiter pleinement les technologies les plus récentes qui font l'objet des chapitres suivants et d'être ainsi mieux armés pour répondre aux besoins et s'approprier les pratiques d'aujourd'hui et de demain.

1. Métadonnées : processus de création et administration

Repartir des définitions fournies par des dictionnaires va nous éclairer à la fois sur la portée des métadonnées et sur le processus de leur création.

◆ 1.1 Métadonnées : l'aboutissement d'une démarche

1.1.1 Métadonnée = méta + donnée

Le vocable métadonnée apparaît dès 1969 dans le contexte du développement d'un produit informatique associé à un MetaModel par J. E. Myers¹. Il est dès lors utilisé dans le secteur de l'informatique, dans celui des statistiques ainsi que, dès le début des années 1980, dans le domaine de l'informatique décisionnelle [6].

« *Data about data are referred to as metadata.*² » Cet énoncé nous informe d'une filiation entre métadonnée et donnée. Mais quel lien établir entre les deux? Un retour aux sources de chacun de ces deux termes nous permettra de mieux cerner cette notion et d'en préciser le périmètre.

Une donnée est une représentation d'un élément de connaissance.

En langue générale et par extension de la définition fournie par les mathématiciens, une donnée est « ce qui est connu et admis, et qui sert de base à

1 Voir: Jane Greenberg, « Understanding metadata and metadata schemes », *Cataloging & Classification Quarterly*, 2005, vol. 40, n° 3-4, p. 17-36, <http://ils.unc.edu/mrc/publications> (page 19).

2 James Martin, *Strategic data planning methodologies*, Englewood Cliffs, New Jersey, Prentice-Hall, 1982, p. 127. Cité par [6].

un raisonnement, à un examen ou à une recherche³ ». « Si un homme raisonne mal, c'est qu'il n'a pas les données pour raisonner mieux », écrivait Diderot en 1771 (*Sur le livre de l'Esprit*).

Les normes nous fournissent des définitions fonctionnelles plus précises: une métadonnée est une « représentation réinterprétable d'une information, sous forme conventionnelle convenant à la communication, à l'interprétation ou au traitement. Nota 1: Les données peuvent être traitées par des moyens humains ou automatiques. Nota 2: Par réinterprétable, on entend que la représentation n'est, en principe, pas utilisable en l'état⁴ ».

Pour l'informatique, une donnée est une « représentation d'une information (élément de connaissance) sous une forme conventionnelle destinée à faciliter son traitement.⁵ » L'informatique réduit la portée du terme en définissant un environnement applicatif précis: le développement de systèmes informatiques.

La première définition précise l'usage de la donnée: support d'un raisonnement, examen, recherche. La deuxième ajoute deux caractéristiques non formulées dans les définitions fournies par des dictionnaires de langue générale. Celle, tout d'abord, d'être une représentation d'un élément de connaissance; la connaissance étant individuelle (humaine), seule une représentation de celle-ci, une donnée, peut être partagée. Une autre caractéristique de la notion de donnée nous fait pénétrer dans le monde du codage. En effet, la donnée s'expose sous une forme conventionnelle, c'est-à-dire qu'elle se donne à voir sous une forme codifiée, cette codification étant souvent normalisée.

Notons d'emblée que la notion de donnée est bien antérieure à son usage dans le domaine de l'informatique et que cette définition fournit la portée réelle du terme. Enfin, parmi les caractéristiques citées, celle de la représentation d'un élément de la connaissance est sûrement la plus importante mais la moins revendiquée. Nous y reviendrons.

Méta est un préfixe grec (μετα) qui peut exprimer (1) la succession (métacarpin), (2) le changement comme dans métamorphose, (3) la participation ou encore (4) le dépassement ou englobement (métalangue)⁶. Le préfixe méta nous signale la présence d'une relation forte entre deux choses, l'une introduite par ce préfixe, l'autre identifiée par l'unité qui le suit: -carpin,

3 Portail lexical CNRTL: www.cnrtl.fr/definition/donnee (consulté en mai 2008).

4 XP X50-435. Septembre 1995. *Management des systèmes - Gestion documentaire - Concepts généraux* (norme expérimentale).

5 Arrêté du 22 décembre 1981. FranceTerme, termes recommandés au *Journal officiel de la République française*, <http://franceterme.culture.fr>

6 Voir: Danielle de Clercq, *Étymons grecs et latins du vocabulaire scientifique français*, ITINERA ELECTRONICA, Université catholique de Louvain, 2000, p. CXXI, pot-pourri.fltr.ucl.ac.be/itineria/ebook/etymons.pdf

-morphose, -langue. Le deuxième indice fourni précise que les « choses dont on parle », les référents, sont de nature identique : une métalangue est aussi une langue, une métadonnée est une donnée. Le terme méta est utilisé pour indiquer l'autoréférence et pour désigner un niveau d'abstraction supérieur.

Nous retiendrons de ce bref examen que :

- les données, qu'elles soient analogiques ou numériques, se doublent d'une information sur elles-mêmes appelée métadonnée, destinée à renseigner sur elles et à anticiper leur utilisation ;
- une métadonnée est une représentation de la donnée [voir ci-dessous la section 1.1.3] : parlant de métadonnée, nous nous situons déjà à un deuxième niveau de représentation, oubliant bien souvent le premier niveau de représentation ;
- la métadonnée, tout comme la donnée qu'elle représente, est codifiée, ce qui suppose le choix de catégories d'outils adaptés à cette opération.

Les distinctions que nous venons de faire entre donnée et métadonnée vont nous permettre de clarifier les rapports entre les vocables rencontrés : élément de données et métadonnées, répertoires, référentiels ou registres de métadonnées, schémas de métadonnées ou schémas d'encodage. Ce travail nous permettra de mieux comprendre leurs apports respectifs et leur complémentarité.

1.1.2 Des métadonnées : pour quoi faire ?

Représentant ou décrivant le monde réel physique ou virtuel, une métadonnée ou un ensemble de métadonnées peuvent être créées pour représenter à peu près tout ce qui fait le monde et y donner accès. Les métadonnées ne sont donc pas propres au secteur de l'info-doc ni à celui de l'informatique. Elles sont utilisées dans de très nombreux environnements pour représenter aussi bien un gène qu'un ouvrage, une facture qu'une personne, un produit ou une matière, une ville, une projection géographique, une action ou une activité comme un suivi de production, une date de validité, une mesure du taux de pollution chimique de l'air, etc.

Les métadonnées font partie du monde des outils de repérage et d'accès à l'information définis comme « des voies d'accès à l'information, *via* le document et son support, à travers des outils qui sont capables de fournir les références du document, voire les informations elles-mêmes⁷ ».

Intuitivement on sent que pour représenter une telle diversité de données sources, les métadonnées ne peuvent prendre exactement la même forme.

⁷ Sylvie Fayet-Scribe, *Chronologie des supports, des dispositifs spatiaux, des outils de repérage de l'information*, http://biblio-fr.info.unicaen.fr/bnum/jelec/Solaris/d04/4fayet_1tab.html

Les métadonnées associées à des enregistrements de signaux physiologiques intégrés à des dossiers médicaux portent sur des noms d'« auteurs » et des dates de création de ces enregistrements, mais elles portent également sur l'identification des personnes traitées, les techniques utilisées pour produire le signal à partir d'une nomenclature normalisée, son format, son support ou autres caractéristiques intrinsèques à l'enregistrement, le contexte de la production de ces données, les données résultats, les signaux eux-mêmes qui représentent l'information physiologique, éventuellement les commentaires des techniciens sur la mesure ou l'analyse des résultats obtenus, etc.

Mais, aussi différentes qu'elles soient et quelle que soit la nature de l'information qu'elles représentent, toutes ces métadonnées constituent autant de points pour rechercher, exploiter, naviguer, gérer, diffuser ou produire à nouveaux des données, par des humains ou d'autres agents automatiques.

Cet exemple permet d'isoler quelques catégories de métadonnées de natures différentes :

- des métadonnées pour décrire et identifier la donnée source, ses caractéristiques, le contexte de sa production ;
- des métadonnées d'administration des données représentées, incluant la gestion des droits d'accès à ces données ;
- des métadonnées structurelles entre les éléments de données ou au sein des données que, selon leur fonction, on peut associer à la première catégorie ou à la deuxième ;
- des méta-métadonnées, c'est-à-dire des métadonnées sur les métadonnées produites (qui les a créées, quand, etc.).

1.1.3 Les étapes clés conduisant aux métadonnées

L'étude des étapes conduisant aux métadonnées va nous fournir des éléments à même de cerner le territoire concret des métadonnées, et par contre-coup celui des données elles-mêmes.

A. Conception métier

Quelles sont les données à représenter ?

Une première phase regroupe les étapes liées aux données et à leur relation avec les « choses de la vie » à représenter ou à décrire. La démarche comporte une étape initiale essentielle de modélisation de la réalité avant de dessiner le modèle qui sera exploité dans une deuxième phase d'ingénierie informatique.

◆ Étape 1 – Modéliser la réalité

Un modèle est une abstraction de la réalité, un objet, un événement, un projet, un concept, etc.

Exemples :

« Partage des données bibliographiques et d'autorité⁸ »

Dans le domaine biomédical, favoriser la guérison en décrivant précisément les maladies

Développer l'administration électronique⁹

Cette réalité perçue est analysée et épurée, le modèle ne conservant que certaines caractéristiques représentatives, sélectionnées par les concepteurs en fonction des objectifs assignés au projet.

Exemples :

Le modèle FRBR fournit un « cadre conceptuel où [sont] identifiées et clairement définies les entités pertinentes pour les utilisateurs de notices bibliographiques¹⁰ »

La communauté médicale a développé dans les années 1940 un « modèle conceptuel biomédical » des conséquences des maladies¹¹.

Le modèle est donc une vue subjective mais jugée pertinente de la réalité. Le contexte de la modélisation, contexte défini par les objectifs du projet, est un élément essentiel de la compréhension des modèles.

Les modèles et la modélisation constituent des aides à l'élaboration et à la structuration des idées ; ce sont des supports au raisonnement. En effet, les facultés de compréhension de l'homme ont leurs limites face à la complexité. En restreignant le problème étudié et en le formalisant, la modélisation et les modèles permettent de se concentrer sur les caractéristiques formellement identifiées intégrant les besoins et contraintes spécifiques à son environnement. Ils facilitent également les échanges entre personnes différentes (chercheurs, professionnels, développeurs, etc.) en donnant une vision « externalisée » de l'objet étudié, une simulation du système étudié.

Modèles et modélisation sont utilisés depuis de nombreuses années dans des domaines variés : l'économie, la pédagogie, la médecine, l'écologie, l'architecture ou, bien sûr, le secteur de l'info-doc.

Dans cette étape initiale du processus global qui conduit à la production et à l'administration de métadonnées, le modèle « fonctionne à un niveau conceptuel » : il n'atteint pas le niveau de décomposition et de description requis par un modèle de données informatiques.

« Le modèle élaboré pour la présente étude incarne, autant que faire se peut, une conception "globale" de l'univers bibliographique ; il est censé être indépendant de tout corpus de règles de catalogage ou de toute application des concepts qu'il expose. » [FRBR¹²].

Il s'agit bien d'un cadre conceptuel, d'une énonciation de principes directeurs. Nous pourrions aussi parler de modèle métier pour le distinguer du

modèle de données utile à la mise en œuvre informatique dont il sera question plus loin.

◆ *Étape 2 – Formalisation du modèle conceptuel*

Les cadres conceptuels élaborés dans l'étape 1 offrent des canevas qui guident la construction d'un dispositif et des bases des systèmes d'information à associer; ils permettent de les documenter. Si le niveau de précision de ce cadre conceptuel de l'univers à représenter est variable et peu utilisable directement par des développeurs informatiques, il est par contre toujours possible voire souhaitable d'utiliser à ce stade des outils de modélisation aptes à aider les praticiens dans cette démarche.

C'est dans ce cadre que s'utilise à l'heure actuelle le modèle entité-relation.

Le modèle entité-relation [figure 1]

Une entité est un objet pourvu d'une existence propre. Ce peut être: un ou des individus (un auteur, une équipe, une société, un gène, etc.), une chose concrète ou abstraite (publication, pays, etc.) ou un événement (commande, manifestation, prescription, etc.).

Une relation est une association fonctionnelle entre deux entités ou classes d'entités.

On parle d'attribut pour désigner les données élémentaires sur une entité (date ou lieu de naissance d'une personne) ou sur une relation. Un des attributs de l'entité est un identifiant (numéro SIRET pour l'entité entreprise, numéro ISBN pour l'entité publication, DOI, etc.) permettant de nommer ou d'identifier, de façon non ambiguë, une instance (valeur) de l'entité.

Nous reviendrons sur l'importance de ces identifiants [voir notamment le chapitre 5].

La cardinalité (dimension ou degré de la relation) est le nombre d'entités impliquées dans cette relation. La relation peut ne faire intervenir qu'une seule entité: elle est dite réflexive. Il est nécessaire de préciser pour chaque entité la cardinalité minimum (0 ou 1) et maximum (n).

8 *Les principes internationaux de catalogage de l'IFLA* (Appel à commentaire), introduction, avril 2008, www.ifla.org/VII/s13/icc/principles_review_200804.htm

9 Site officiel « Les ressources de l'administration publique »: www.synergies-publiques.fr

10 Groupe de travail IFLA sur les spécifications fonctionnelles des notices bibliographiques, *Spécifications fonctionnelles des notices bibliographiques [FRBR]: rapport final*. Édition française établie par la Bibliothèque nationale de France, 2001, p. 9/124. www.ifla.org/VII/s13/wgfrbr/finalreport.htm

11 Extrait de: Patrick Fougeyrollas, « Changements sociaux et leurs impacts sur la conceptualisation du processus de handicap », *Réseau international CIDIH et facteurs environnementaux*, 1998, vol. 9, n° 2-3, p. 7-13, www.med.univ-rennes1.fr/sisrai/art/modele_conceptuel.html

12 FRBR (voir la note 10).

FIGURE 1 – FORMALISATION ENTITÉ-RELATIONS-ATTRIBUTS

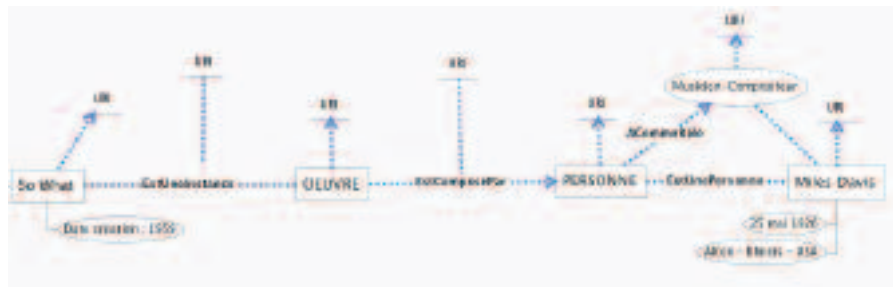
Dans une approche traditionnelle, le typage des relations n'est pas intégré au modèle fourni au système informatique.

Table Œuvre	Code_Personne	Table Personne	Code_Personne
So What	2	Miles Davis	2
So What	4	Carla Bley	4

Les relations ne sont pas "visibles" entre ces deux tables.



Dans les nouvelles approches entités/relation, les entités, les relations et leurs attributs sont explicités et formalisés.



Le modèle CRM formalise une relation « *has former or current owner* » (et son contraire « *is former or current owner of* » (P51) », entre deux entités: Physical Thing (E18) et Actor (E39).

Une personne qui a composé une œuvre sera formellement distinguée de celle qui interprète cette œuvre, grâce à un typage des relations [voir le schéma de la figure 1].

Concernant la cardinalité: dans une bibliothèque, un usager (entité) emprunte (relation) de 0 à n livres (entités); mais le livre objet physique (entité) ne peut être emprunté que par 0 ou 1 lecteur.

Ces formalismes ne sont pas suffisants pour l'implantation d'une application informatisée, mais ils permettent de s'assurer d'une compréhension fine et partagée entre acteurs du métier. Un des obstacles rencontré par les praticiens de l'info-doc est la difficulté, pour ceux qui n'ont pas participé à cette schématisation, d'interpréter ce formalisme et, au-delà, d'avoir une réelle compréhension des concepts et relations métiers sous-jacents. Un investissement est certainement à faire non pas dans l'élaboration même de ces formalismes mais dans leur interprétation et leur utilisation par un public plus large de praticiens et de formateurs métiers.

◆ *Étape 3 – Développement de référentiels métiers*

Pour concevoir un dispositif (un catalogue de bibliothèque, un système de surveillance des maladies, une base de connaissance sur la création musicale, etc.) en s'appuyant sur un modèle conceptuel préétabli, les différentes communautés métiers développent des règles et des outils pour produire les données: règles de sélection et d'acquisition des ressources à traiter, règles de traitement, vocabulaires d'indexation, identifiant normalisé, normalisation ou codification des valeurs, etc. Cet outillage peut porter d'autres dénominations: vocabulaires contrôlés, nomenclatures, listes d'autorité, composants sémantiques, schémas d'encodage (terme employé pour le vocabulaire contrôlé dans le monde des métadonnées) des données et enfin référentiels métiers.

Ces vocabulaires et règles ont existé de tous temps, mais l'informatique en a multiplié la nature et le nombre. Avec le déploiement de la logique des métadonnées, on assiste à un effort visible de mutualisation et de rationalisation de référentiels métiers communs: archives, bibliothèques, musées, mais aussi dans des secteurs comme la photographie, la pédagogie, etc.

Le modèle biomédical prend appui sur un schéma international, la *Classification internationale des maladies* (CIM)¹³, et des règles associées.

Le modèle bibliographique français se structure autour d'un schéma conceptuel normalisé au niveau international, l'ISBD (International Standard Bibliographic Description); il exploite selon des règles précises un référentiel terminologique, RAMEAU.

L'industrie photographique a mis sur pied un schéma de description et de codage des données photographiques, l'IPTC, mis à la disposition de la profession. Ce modèle inclut des vocabulaires contrôlés comme les genres, des catégories et des sujets¹⁴.

¹³ Voir la note 11.

¹⁴ Voir le site officiel de l'International Press Telecommunications Council (IPTC) : www.iptc.org/NewsCodes

Le Réseau européen du patrimoine, HEREIN, se réfère au thésaurus multilingue éponyme¹⁵.

Quant à la production de données au sein du Réseau européen sur l'environnement EIONET, elle exploite, entre autres, le thésaurus GEMET.

Le cadre politique et stratégique de l'administration électronique propose, en réponse à la loi n° 2005-102 du 11 février 2005, un *Référentiel général d'accessibilité des administrations*, le RGAA¹⁶.

Ces activités d'élaboration de modèles conceptuels et d'outils associés (étapes 1 à 3) utilisables par les praticiens métiers ne se présentent pas sous un format directement exploitable par des automates ou d'autres systèmes d'information. Il est nécessaire de leur faire subir une série de traitements propres au monde des systèmes d'information, traitements que nous allons aborder maintenant.

B. Conception informatique **Des données acceptables par des automates**

◆ *Étape 4 – Développement du modèle de données informatique*

Nous retrouvons dans cette étape la notion de modèle, mais cette fois-ci dans une approche plus formelle exploitant des règles et outils propres aux systèmes informatisés.

Le modèle de données informatique est désigné comme un modèle abstrait par rapport au schéma technique et au *binding* informatique établis en aval. Il spécifie, dans l'environnement XML, les concepts utilisés, la nature des éléments de données et leur agencement. Il décrit une structure informationnelle abstraite de manière formelle.

C'est dans ce cadre que la notation UML¹⁷, langage graphique de modélisation des données et des traitements, est de plus en plus fréquemment utilisée. En effet, la présentation du modèle sous une forme strictement rédactionnelle peut prêter à des divergences de compréhension au moment de sa mise en œuvre informatique, créant ainsi des divergences dans la réalisation d'applications prenant pourtant comme point de départ le même modèle conceptuel. Cette notation permet de décrire sans ambiguïté les éléments du système d'information et leurs interrelations. Le formalisme apporté

¹⁵ Réseau européen du patrimoine, HEREIN, www.european-heritage.net/sdx/herein/thesaurus/showdoc.xsp?doc=howto

¹⁶ www.synergies-publiques.fr/rubrique.php?id_rubrique=202

¹⁷ Unified Modeling Language (UML). Voir : G. Booch, J. Rumbaugh, I. Jacobson, *UML, le guide de l'utilisateur*, Eyrolles, 2000. Voir aussi : Laurent Piechocki, *UML, le langage de modélisation unifié*, <http://uml.free.fr/index-cours.html>

ici par un langage graphique comme UML vise clairement à lever toute ambiguïté aussi bien sur les concepts, les entités que sur les relations mises en œuvre dans le modèle de données.

De façon pragmatique si les ressources ou les compétences pour utiliser une notation comme UML manquent, un tableur ou un outil de visualisation graphique peut être exploité. L'idée est qu'une représentation sous une forme schématisée et si possible normalisée impose un travail plus rigoureux et assure un rendu plus proche de la réalité qu'un modèle exprimé en langage naturel.

Mais le passage entre modèle conceptuel orienté métier et modèle conceptuel orienté données informatiques reste le point d'achoppement dans cette démarche complexe qui va d'une réalité à une application. Il est nécessaire qu'informaticiens et praticiens se retrouvent autour d'une même table pour contrôler leur compréhension réciproque du modèle à implanter jusque dans les moindres détails, sans omettre de revenir aux étapes précédentes lorsque subsistent des ambiguïtés.

◆ **Étape 5 – Schéma des données informatiques et encodage des métadonnées**

La démarche classique de tout projet informatique aboutit, dans cette cinquième étape, à un schéma¹⁸ des données sous un format adapté aux activités informatiques. Si la forme que prend un schéma informatique diffère selon l'environnement technique dans lequel on se trouve (tables des bases de données relationnelles ou schéma normalisant le balisage des informations et de leur arborescence dans un document XML), le schéma et sa documentation restent dans tous les cas la concrétisation du modèle conceptuel et l'ultime outil manipulé par des humains. Ce formalisme est *a priori* orienté vers les publics développeurs, ce qui rend ces outils difficiles à comprendre pour les acteurs métier.

Ici aussi une grande ambiguïté subsiste. À cette étape du processus – qui se poursuivra par les travaux de développement proprement dit des programmes et des interfaces –, il s'agit d'un encodage à caractère informatique, sous-entendu compréhensible par la machine. Mais nous retrouvons les deux mêmes niveaux que dans les étapes « métiers » :

- l'encodage que nous considérons comme strictement technique, qui structure les données en vue de préparer les applicatifs. Un premier niveau de sémantique pour les automates apparaît à ce niveau ;
- plus récemment est apparu un autre type d'encodage dans l'environnement XML. Celui-ci a prétention à préserver et consolider la sémantique des

¹⁸ Un schéma est « une représentation des constituants fondamentaux d'un objet complexe, incluant les relations fonctionnelles existant entre ces constituants » [CNRTL, www.cnrtl.fr].

métadonnées définie à l'étape précédente. Cet encodage constitue un élément clé des dispositifs dans l'optique de l'interopérabilité entre communautés, entre modèles ou entre schémas.

Dans le monde du web et du web ouvert, les nouveautés de ces dix¹⁹ dernières années portent incontestablement sur l'arrivée :

- de l'environnement XML pour la mise en forme que nous appellerons ici technique. Les langages de schémas – XML-Schéma W3C XML (XSD), Schematron ou RelaxNG – sont alors utiles dans la production et la validation des documents XML et la génération d'applications²⁰ ;
- d'un vocabulaire RDF pour l'encodage des métadonnées [voir le chapitre 5].

La génération d'un document XML conforme au schéma informatique précédemment élaboré (souvent nommé *binding*) constitue un outil adapté à une implantation en machine [voir l'exemple ci-contre]. En général, parce qu'ils doivent pouvoir évoluer rapidement en fonction des contraintes techniques, ces schémas ne sont pas normalisés mais fournis à titre informatif. Ils sont proposés dans des référentiels avec d'autres outils de développement (feuille de style, outils de conversion, etc.) et la documentation associée.

La famille des **formats MARC**, développés dans les années soixante, correspond à des formalisations informatiques du modèle bibliographique avec comme finalité d'être lues par des machines ; les règles d'encodage utilisées sont propres à ces applications bibliographiques.

Le programme MARCXML²¹ a pour finalité le développement dans un environnement XML d'un cadre de travail dédié aux données MARC (MARC21). Le référentiel propose de nombreux composants comme des schémas, des feuilles de style et des outils logiciels. L'environnement XML permet d'aller plus loin que le strict format MARC ou Unimarc grâce aux possibilités d'imbrication des éléments rendant possibles d'autres types de regroupement ; le système de nommage XML offre plus de souplesse aux utilisateurs. Quant au projet MODS²² de schéma de métadonnées pour les données bibliographiques, il permet d'utiliser au mieux les possibilités de XML.

Le processus ne s'arrête pas à cette étape et se poursuit par les travaux concrets de développement informatique que nous ne traitons pas dans ce chapitre.

19 La spécification XML 1.0 a été publiée en 1998 et la première version de RDF date de 1997.

20 Éric van der Vlist, *Quel langage de schéma XML choisir pour chaque usage ?* Présentation faite à la réunion SparklingPoint, 14 mars 2003, <http://xmlfr.org/documentations/articles/030314-0001>

21 www.loc.gov/standards/marcxml

22 MetadataObject Description Schema (MODS) : www.loc.gov/standards/mods

23 *Les FRBR, qu'est-ce que c'est ?* Manuel, 10 mars 2005, www.figoblog.org/document594.php [l'avis d'une bibliothécaire à l'arrivée de ce modèle].

EXTRAIT DU BINDING XML PROPOSÉ SUR LE SITE DE LOM-FR

```

</lom:title>
<lom:language>fre</lom:language>
<lom:description>
  <lom:string language="fre">Ce matériel pédagogique est spécifiquement destiné à des
  étudiants étrangers non francophones et orientés vers l'apprentissage du français langue
  étrangère dans les spécialités scientifiques. Un parcours guidé d'environ 50 exercices de
  français exploite un exposé oral de 24 mn, enregistré et médiatisé (vidéo et diapositives
  synchronisée). </lom:string>
</lom:description>
<lom:keyword>
  <lom:string language="fre">gestion de projet</lom:string>
</lom:keyword>
<lom:keyword>
  <lom:string language="fre">FLE</lom:string>
</lom:keyword>
<lom:keyword>
  <lom:string language="fre">Filipe</lom:string>
</lom:keyword>
<lom:aggregationLevel>
  <lom:source>LOMv1.0</lom:source>
  <lom:value>3</lom:value>
</lom:aggregationLevel>
<lomfr:documentType>
  <lomfr:source>LOMFRv1.0</lomfr:source>
  <lomfr:value>texte</lomfr:value>
</lomfr:documentType>
<lomfr:documentType>
  <lomfr:source>LOMFRv1.0</lomfr:source>
  <lomfr:value>image en mouvement</lomfr:value>

```

◆ 1.2 Du domaine normatif au domaine applicatif

Si la démarche globale et la typologie des outils de modélisation, de structuration et de schématisation présentées dans la section précédente semblent justifiées et cohérentes par rapport aux finalités des applications, il faut bien avouer que toutes ces étapes ne sont pas toujours effectuées ni même prévues. Ou bien encore la cohérence d'ensemble n'est pas toujours au rendez-vous, les étapes finales d'implantation pouvant se mettre en place avant la formalisation totale des étapes de modélisation²³.

1.2.1 Complexité de la mise en œuvre d'applications

Plusieurs facteurs se conjuguent pour complexifier la situation et interagir dans ce processus.

Plusieurs modèles et schémas en jeu

Les applications au plus près des utilisateurs finals des ressources doivent bien souvent articuler plusieurs schémas voire plusieurs répertoires de métadonnées.

En effet, le cas d'une bibliothèque, d'un centre d'archives, d'un musée ou d'un dépositaire qui n'utiliserait qu'un seul schéma normalisé au niveau international – MARC, MODS ou OAIS – constitue un cas particulier. La réalité n'est pas aussi uniforme que cela, et le développement d'applications embrassant dossiers d'affaires ou d'archives (RM), documents techniques ou pédagogiques (LOM-FR), publications (Unimarc simplifié), fonds audiovisuels ou photographique (IPTC), objets de musée ou données métiers comme des observatoires, obligent concepteurs et praticiens à bien préciser le sens de leur propre schéma pour trouver les points de convergence.

Car l'objectif est à la mesure des enjeux: il s'agit de prendre en compte les spécificités des différentes catégories de publics et ressources au sein d'une structure d'ensemble sémantiquement cohérente, sans réduire les potentialités d'exploitation de ces réservoirs aux seules métadonnées Dublin Core, fussent-elles qualifiées. Sur cet aspect, l'étude du recouvrement entre un schéma spécialisé comme celui des ressources pédagogiques LOM-FR et le Dublin Core montre que les données spécialisées non couvertes ou mal identifiées par ce dernier sont pour les utilisateurs de ces informations d'une grande efficacité dans le repérage et surtout la réexploitation des ressources.

Plus que la multiplication des schémas eux-mêmes, les difficultés éprouvées au cours de ce travail d'architecture de l'information viennent des micro-mondes représentés au sein de chacun de ces schémas. Surtout lorsqu'ils constituent un référent de longue date au cœur d'une communauté, ceux-ci sont élaborés sur la base de l'existant. Et, s'ils évoluent, ils le font rarement en tenant compte d'autres besoins en proximité, tant ce travail est déjà complexe. Mais, en réduisant trop fortement le périmètre de certains éléments de données ou en les enfermant dans un format trop spécifique, ces schémas rendent l'articulation avec d'autres schémas difficile (encodage ou sémantique variés) voire impossible (donnée absente), dans un périmètre d'usage pourtant proche.

Les exigences liées à l'interopérabilité

En terme d'architecture d'information, les choix de fusion, intégration ou unicité logique mais aussi physique des données qui ont prévalu jusque-là semblent difficiles à maintenir dans le contexte du web. Il s'agit plutôt de qualifier les données sur un périmètre donné et de faciliter l'interopérabilité [7] entre applications.

« L'interopérabilité est un état qui existe entre deux applications quand, pour une tâche spécifique, une application peut accepter les données d'une autre application pour effectuer cette tâche, de manière appropriée et satisfaisante, sans que cela ne nécessite l'intervention d'un opérateur extérieur. [17] »

Mais il ne suffit pas de respecter des normes générales (Dublin Core) ou d'utiliser les mêmes outils (XML) pour assurer l'interopérabilité entre applications et la fluidité des données.

La contrainte de l'existant

De nombreux projets ont à prendre en charge un existant conséquent. Nous ne sommes plus seulement dans un univers de recherche développant de nouveaux outils de représentation, de modélisation ou d'encodage, mais bien dans celui d'applications et de données d'exploitation où le développement de nouveaux systèmes s'effectue avec l'obligation de préserver la continuité de service. Ce contexte est le quotidien bien réel de bon nombre de lieux d'information documentaire. Mais en définitive il est également le quotidien de tout lieu où l'information est la matière première et le moteur des activités: information de l'Administration et des échanges avec les administrés, matériaux (corpus) des chercheurs, dossiers médicaux ou dossiers d'affaires, données de production, observatoires divers, etc.

Adopter un esprit de liberté

Enfin les méthodes de conduite de projet et l'esprit avec lequel ces travaux sont conduits doivent prendre en compte certaines spécificités des nouveaux environnements informatiques XML.

Les principes de développement des années 1960 à 1990 restaient très binaires: allait-on suivre la norme ou pas, avec en toile de fond le principe que toute métadonnée produite était alimentée manuellement et si possible « livre en main »! Le contexte technique de ces années ramenait la démarche à trois étapes clés sur les cinq exposées précédemment, les modèles métiers et les schémas techniques applicatifs étant faiblement dissociés. Il était très difficile, lorsqu'une norme était suivie, d'imaginer des compléments applicatifs et encore moins une articulation harmonieuse avec un autre environnement normatif.

Quant à l'alimentation d'une métadonnée par récupération ou production automatique²⁴, elle reste encore aujourd'hui difficile à imaginer pour certains d'entre nous. Partager un réservoir de notices ou coproduire des métadonnées

²⁴ Notons que la capture automatique d'éléments bibliographiques de documents sur la base de feuilles de style ou de macrocommandes est possible avec les outils bureautiques depuis vingt ans.

avec des producteurs ou des automates ne relève pas de la même démarche collaborative de travail : dans la première, on est totalement maître de l'unité produite ; dans l'autre cas, cette unité est coproduite sans que l'on exerce la moindre action sur les parties qui ne seraient pas de sa responsabilité, si ce n'est celle d'ajouter d'autres métadonnées à celles fournies en amont pour normaliser les points d'accès.

Cette approche traditionnelle de production de métadonnées est dans une certaine mesure battue en brèche dans l'environnement XML. L'approche normative reste valable et prend même encore plus de poids, mais elle doit s'articuler avec une approche ouverte sur les besoins et l'environnement de travail des publics et sur les possibilités applicatives. Par approche ouverte, nous faisons référence tout d'abord à l'utilisation conjointe de plusieurs normes ou de normes avec des éléments applicatifs non normalisés, mais aussi à la nécessité d'accepter, pour des raisons évidentes de performance, le principe de la coproduction des métadonnées. Des métadonnées sont produites conjointement avec la ressource elle-même ; d'autres métadonnées sont ajoutées tout au long de la vie de la ressource en intégrant celles liées aux usages et à la gestion.

Dans ce contexte, la logique de portail ou simplement de services doit reposer sur le développement d'interfaces et de langages de recherche orientés utilisateurs, exploitant des réservoirs de métadonnées récupérés automatiquement. Les fils RSS et les applications composites (*mashup*) témoignent de ces nouvelles possibilités qui seules peuvent satisfaire à la fois les exigences des utilisateurs et les contraintes économiques liées aux volumes et flux d'information.

1.2.2 Avec quels outils produire des métadonnées ?

Nous vivons une époque charnière où la production de métadonnées oscille entre une saisie postérieure à la production des données assurée par des spécialistes de cette activité et la récupération de métadonnées produites à la source et éventuellement codifiées dans une autre étape. Quels sont les outils utilisables pour ces deux catégories d'activités ?

La production de métadonnées peut être préalable à celle même de la ressource lorsqu'il existe un plan de production documentaire ; elle peut être concomitante, ce qui suppose des outils d'aide à la production (feuille de style ou éditeur adapté), ou postérieure à celle-ci.

Ces métadonnées sont produites par des automates, par des humains ou par un travail partagé entre des automates et des humains. Des travaux sont conduits sur des applications générant automatiquement des métadonnées de ressources numériques [2].

Enfin, les métadonnées peuvent se trouver dans un fichier autonome ou dans une base de données; être associées ou non à la ressource, chaque métadonnée ayant alors comme fonction d'identifier la ressource de référence; ou encore être totalement intégrées à la ressource (*embedded*).

Matériellement de nombreuses solutions sont possibles. Des appareils, équipements ou logiciels ont la capacité de fournir les métadonnées associées aux données produites au moment même de leur production grâce à diverses programmations: feuilles de style ou macro avec des éditeurs bureautiques, données de géolocalisation ou date et lieu avec les caméras ou appareils photo, caractéristiques des valeurs enregistrées par un capteur, systèmes d'écriture reconnus par le système, etc.

Ces mêmes métadonnées peuvent être saisies par un humain avec un logiciel de type éditeur de texte ou des applications dédiées comme le logiciel LomFrPad²⁵, ou avec tout système de gestion de bases de données plus classique dont, bien sûr, les applications documentaires. Il existe également des éditeurs convertisseurs en ligne comme Educaméta²⁶, développé par le Scéren.

Mais il ne suffit pas de produire les métadonnées, il faut pouvoir les exploiter et les manipuler au sein d'applications. Pour cela la panoplie des outils s'étoffe chaque jour: outils de diffusion, de fouille de données, bases de connaissances, référentiels terminologiques, etc., sont autant d'outils de production et d'acquisition de contenus, de gestion et d'organisation, d'accès ou de communication des données et des métadonnées [voir les chapitres 2 et 6].

◆ 1.3 Administration des métadonnées et des registres de métadonnées

Pour optimiser et sécuriser les flux de données et de métadonnées – ces dernières étant associées ou non aux données sources –, des communautés de praticiens s'efforcent de normaliser des ensembles de métadonnées, de les documenter et de les codifier suivant des règles et des vocabulaires communs pour qu'elles puissent être partagées et réexploitées dans des contextes variés. Le maintien de registres de métadonnées constitue un élément indispensable pour garantir la protection du patrimoine informationnel dans des contextes où l'on doit échanger fréquemment et avec de fortes contraintes d'interopérabilité.

25 Site du LomFrPad : <http://correlyce.tech.fr/lomfrpad>
 Article : www.normetic.org/Evaluation-d-editeurs-de.html
 Site de LOMPAD : <http://helios.licef.ca:8080/LomPad/en/index.htm>
 26 www.educameta.cndp.fr

1.3.1 Jeu de métadonnées, registre ou référentiel de métadonnées, profil d'application

Plusieurs vocables sont utilisés dans des contextes sensiblement différents. Nous trouvons : jeu de métadonnées, registre de métadonnées, profil d'application, référentiel.

Lorsque les outils propres à la production et à l'administration d'un ensemble d'éléments de données – jeu de métadonnées, règles, vocabulaire d'encodage et éventuellement tableaux de concordance – sont documentés et mis à la disposition de communautés, sans être directement exploités au sein d'une application, on parlera de registre de métadonnées (*metadata registry*). Ce terme peut comporter une dimension réglementaire ou obligatoire qui n'existe pas dans le terme de schéma de métadonnées.

On parlera de profil d'application ou profil applicatif (AP, *application profile*) d'un schéma donné lorsque, au sein d'une organisation, d'une communauté d'utilisateurs ou d'une application précise, sont déclarés non seulement un jeu de métadonnées dérivé du schéma source mais aussi la codification des éléments de données et des valeurs, la normalisation des vocabulaires de codification (on parle aussi de schémas d'encodage) utilisés pour désigner ces valeurs, ainsi que des règles d'utilisation pour le développement d'un schéma technique. On peut citer le profil d'application pour les ressources pédagogiques, LOM-FR pour la France ou Normetic pour le Québec²⁷, ces deux profils s'appuyant sur un même modèle largement utilisé dans ce secteur, le LOM étant en cours de normalisation au sein de l'ISO. Ces profils d'application documentent les métadonnées et leur utilisation, facilitent leur réutilisation et la création d'applications interopérables.

Le terme référentiel, quant à lui, vient de son usage en informatique où il désigne « un ensemble de bases de données contenant les “références” d'un système d'information » : « Le référentiel, c'est la colonne vertébrale d'un système d'information.²⁸ » Son orientation contraignante et formalisée parce qu'applicative est parfois fortement atténuée, lorsque le registre ou le schéma proposé dans le référentiel est mis à disposition sans qu'une application soit réellement mise en œuvre. C'est le cas du terme référentiel appliqué à la mise à disposition de vocabulaires contrôlés. On rencontre également le mot de référentiel pour désigner un entrepôt de métadonnées de ressources, c'est-à-dire pour une base de références de ressources décrites suivant le schéma de métadonnées en référence.

²⁷ Normetic, GTN-Québec : www.normetic.org

²⁸ Michel Volle, *Qu'est-ce qu'un référentiel?* 15 juillet 2001, www.volle.com/travaux/referentiel.htm

Nous garderons le terme de référentiel dans un cadre applicatif où sont associés les outils et les données d'encodage, ainsi que les règles, bonnes pratiques ou directives. Pour renforcer cet aspect contraignant, certains le nomment « registre référentiel »²⁹.

1.3.2 Documentation des métadonnées et des registres de métadonnées

L'utilisation de schémas de données, comme de tout outil informatique, suppose que ceux-ci soient correctement spécifiés et documentés afin de répondre aux exigences de fonctionnement et aux besoins d'évolution technique ou fonctionnelle.

Plusieurs motifs rendent cette documentation indispensable. Tout d'abord la multiplication des systèmes d'information comme supports aux activités humaines rend indispensable l'interopérabilité entre applications et par contrecoup impose une transparence sur les applications et les métadonnées, ces dernières représentant, *in fine*, les données. De plus, la grande mobilité des technologies de l'Internet et de leurs usages suppose de bien maîtriser les applications et la structure des données pour évaluer rapidement l'impact de ces évolutions. Dans le même ordre d'idée, la composition des systèmes intégrant des schémas normalisés et, pour certains, réglementaires impose une grande réactivité face à des changements imposés de l'extérieur. La documentation de métadonnées nombreuses que comporte un référentiel ne peut que faciliter ce suivi.

Il n'en reste pas moins que la documentation des systèmes informatiques – celle intégrée aux programmes à l'attention des développeurs ou celle externe sous la forme de manuels pour les administrateurs ou utilisateurs – a toujours constitué un problème en raison des coûts de conception et de maintenance.

Nous venons de citer deux catégories de documentation, pour systèmes et administrateurs « informatiques » et pour administrateurs et gestionnaires « métiers », toutes les deux sous des formes différentes et toutes les deux utiles à deux catégories d'utilisateurs.

Le développement informatique entre dans l'ère du développement agile qui préconise de « produire un logiciel entièrement testé et qui fonctionne plutôt qu'une documentation claire³⁰ ». De plus, dans le contexte du web sémantique, un système d'encodage des métadonnées comme RDF, en préservant la sémantique de celles-ci, constitue en soi une documentation sur les données,

29 Le registre référentiel de l'administration électronique du Québec : www.msg.gouv.qc.ca/fr/administration/registre.asp

30 Laurent Desmons, *Qu'est ce que le développement agile ?* 2004, www.dotnetguru.org/articles/dossiers/devagile/DeveloperAgile.htm

ce qui fait dire à certains spécialistes que cet autoréférencement pourrait être suffisant.

Mais ici il s'agit de la documentation conceptuelle et de sa traduction dans un format, et pas nécessairement de la documentation de tous les développements. Et, en attendant que tous ces systèmes s'autoréferencent et qu'ils produisent une documentation exploitable, une documentation précise de ces référentiels reste indispensable pour les administrateurs et gestionnaires des applications et des registres de métadonnées.

La norme ISO 11179 (Metadata registry ou MDR) a été développée dès la fin des années 1990 à cette fin. Elle est complétée par des procédures d'enregistrement de référentiels (ISO/IEC TR 20943-1:2003) et des directives pour faciliter l'interopérabilité entre registres de métadonnées (ISO/IEC 20944). Ces normes sont proposées par le groupe ISO sur les métadonnées, ISO/IEC JTC1 SC32³¹.

La norme MDR fournit un cadre et des règles pour décrire les métadonnées; elle vise à normaliser et à enregistrer ces éléments de données en vue de faciliter l'interopérabilité et la coopération entre applications et entre jeux de métadonnées différents³², et ceci sans qu'il soit nécessaire de développer un ensemble de traducteurs entre ces jeux de métadonnées.

La norme ISO 11179 « permet de mettre en place un réel système de gestion des métadonnées, c'est-à-dire un système formel qui fournit l'information d'autorité sur la sémantique et la structure de chaque élément. Pour chaque élément, le registre en donne la définition, les qualificatifs qui lui sont associés, ainsi que les correspondances avec des équivalents dans d'autres langues ou d'autres schémas. »

Cette norme est utilisée dans plusieurs institutions publiques et gouvernementales. Citons, par exemple, les travaux sur l'administration électronique en France³³ ou au Canada³⁴.

Elle est en cours de révision dans le cadre d'un projet nommé eXtended MetaData Registry (XMDR) Project³⁵, dont l'objectif est de prendre en compte les spécificités de nouvelles catégories de métadonnées et de compléter la

31 Groupe de travail ISO qui définit les normes internationales pour les métadonnées et technologies associées : <http://metadata-standards.org>

32 Nicolas Delestre, Yolaine Bourda, *Utilisation de la norme ISO11179 pour améliorer l'interopérabilité entre les différents schémas de métadonnées pédagogiques*, http://halshs.archives-ouvertes.fr/docs/00/02/75/65/PDF/Delestre_Bourda.pdf

33 Référentiels sur Synergie, le site de l'administration électronique de la France: www.synergies-publiques.fr/rubrique.php?id_rubrique=1

34 Présentation de la norme par le Secrétariat du Conseil du Canada: www.tbs-sct.gc.ca/im-gi/meta/mdregistry/mdregistr-fra.asp

Le Registre référentiel ebXML-OASIS: www.msg.gouv.qc.ca/admin_electro/registre_referentiel_ebxml-oasis.html

35 http://xmdr.org/content_survey.html

description soit avec d'autres éléments de données soit par l'ajout de règles de normalisation et de schémas d'encodage intégrés à la norme.

Les limites de la norme ISO 11179 dans sa forme actuelle ont rendu nécessaire l'utilisation d'autres prescriptions. Par exemple, le projet ISO pour les ressources pédagogiques, MLR, anciennement LOM, s'appuie sur la norme ISO 9735-2:2002 sur l'échange de données électronique pour la définition de certaines catégories de métadonnées, en particulier pour les métadonnées composites.

Il s'agit donc de documenter chaque élément de donnée du registre selon un même format :

Identifier – identifiant unique assignée à l'élément

Version – version de l'élément de donnée

Name – nom donné à l'élément de donnée

Definition – énoncé qui présente clairement le concept et la nature essentielle de l'élément de donnée

Registration Authority – entité autorisée à enregistrer l'élément de donnée

Language – langue dans laquelle l'élément de données est spécifié

Type de présence – condition de présence ou d'absence: obligatoire, conditionnel et optionnel

Datatype – type de données ou règles canoniques de représentation de l'élément

Comment – commentaire d'usage

Dans le cas où des métadonnées de nature différente sont utilisées, il est indispensable de définir d'autres attributs spécifiques :

Indicateur de structure – élément de donnée composite, autonome ou composant d'un élément composite

Indicateur d'occurrence – dans le cas d'un élément composite, l'ordre des éléments composants est signifiant ou non

Order relation semantics – dans le cas où l'ordre est signifiant

Une documentation s'appuyant sur ces attributs constitue un document de référence qui facilite les échanges entre interlocuteurs au sein d'un projet ou entre plusieurs projets, ou lors d'études comparées avec d'autres schémas pour préparer des évolutions applicatives.

1.3.3 Enregistrement du schéma de métadonnées

La mention de « registre » fait référence à la fonction d'enregistrement. De fait, une partie de la norme ISO 11179 traite des « agences d'enregistrement » qui auraient pour mission de certifier la conformité d'un registre de métadonnées au même titre que pour les identifiants normalisés³⁶. Cette fonction d'enregistre-

36 Un exemple: l'agence d'enregistrement pour l'ISAN, www.france-isan.org/9/presentation.html

EXEMPLE EXTRAIT DU MLR – ISO 19788-2 : 2008 (PAGE 34)

Attribute	Attribute value (string)
Identifier	ISO_IEC_19788-2:2008::DES0092
Name	{{(Contribution Person, eng), (Personne participant à la création de la ressource, fra)}}
Definition	{{(Person that has participated in the making of the learning resource, eng), (information concernant la personne ou l'organisation ayant contribué à la ressource, fr)}}
Presence type	3 (optional)
Structure indicator	2 (component)
Linguistic indicator	1 (linguistic)
Repeatability indicator	10 (any number of occurrences)
Ordered indicator	1 (significant)
Order relation	{{(Contribution Person should be ordered as most relevant first, eng)}}
Lexical space	[[PERSON]] (see section 7.2.4)
Canonical representation rules	PRS0007 (see section 7.3.1)
Example	begin:vcard version:3.0 fn:The British Museum n;;;; org:The British Museum; adr,type=work,type=pref::;Great Russel Street;London;;WC1B 3DG;UK tel,type=work:+44 (0) 20 7323 8000 email,type=internet,type=pref,type=wprk:information@thebritishmuseum.ac.uk end:vcard
Comment	-

ment a été prise en compte dès le début des travaux sur les schémas de métadonnées dans le cadre de différents projets européens, successivement DESIRE, SCHEMAS puis CORES³⁷. Depuis la fin de ces études, en 2003, les archives sont conservées par l'Institut de recherche anglais, l'UKOLN³⁸.

Au sein du Dublin Core, la possibilité d'enregistrer les schémas d'encodage utilisables dans le registre Dublin Core ou dans un profil d'application avait été instaurée mais non maintenue. Des travaux se poursuivent au sein d'un groupe de travail *ad hoc* sur ce thème³⁹.

Ces différents outils d'administration et de gestion optimisent l'administration d'un registre ou référentiel de métadonnées, et facilitent les extensions à partir

d'autres schémas ou microschémas. Que ce soit par l'utilisation de normes qualité propres à un secteur ou une organisation particulière ou par celle de la norme ISO 11179 ou de son successeur, l'élaboration de cette documentation ne peut que faciliter le suivi à long terme et la qualité globale du dispositif.

1.3.4 Quelques mots sur la qualité des données et des métadonnées

Dans l'univers des métadonnées, on peut identifier plusieurs axes « qualité ».

◆ **La qualité des données elles-mêmes.** Celle-ci nous est familière; elle renvoie à la pertinence des règles de codification et à la qualité des schémas d'encodage (vocabulaires contrôlés), au suivi des règles et aux compétences de ceux qui sont en charge de cette activité. Elle permet d'assurer la fiabilité et l'exhaustivité des données. Elle se mesure par des contrôles opérés, par exemple, sur la présence et la qualité des éléments définis comme obligatoires dans les règles de production des données.

◆ **La qualité « interne » du schéma de métadonnées** par rapport au modèle conceptuel représenté: est-il fidèle aux données à représenter? Intègre-t-il l'ensemble des entités, relations et attributs prévus? Sa structure est-elle en cohérence avec les relations définies entre les entités du modèle? Le niveau de spécificité des données est-il bien représenté par les éléments de données et leur format? Le schéma est-il cohérent, c'est-à-dire est-il applicable dans différents contextes relevant du même domaine avec les mêmes objectifs de qualité des données, par exemple est-il applicable avec la même précision quel que soit le type de média?

◆ **La qualité « externe » du schéma de métadonnées.** Si l'on se donne des critères économiques à travers une exigence forte d'interopérabilité ou d'évolutivité, le schéma, les règles, les outils devront être: hospitaliers à d'autres environnements, bien spécifiés et documentés, extensibles et paramétrables. Par exemple, le respect des normes techniques ou sémantiques assure un certain niveau d'interopérabilité; le niveau de granularité des métadonnées peut diminuer les coûts de développement d'interfaces de recherche *ad hoc*⁴⁰.

◆ Ce qui est plus rarement pris en compte est **la qualité du modèle initial** par rapport aux objectifs: présente-t-il une image complète et juste de la réalité modélisée? Permet-il d'atteindre les objectifs visés: identifier des ressources,

37 Site du projet CORES (clôt): www.cores-eu.net

38 www.ukoln.ac.uk

39 Rubrique Registry de DCMI: <http://dublincore.org/groups/registry>

40 Voir aussi le travail effectué au sein du Groupe sur la préservation de l'information numérique (PIN): Claude Huc, *Les critères d'évaluation des formats de représentation de l'information: synthèse et usage*, 2006, http://vds.cnes.fr/pin/presentations/2006/critere_perennite.pdf

sécuriser l'archivage des ressources, gérer les droits et l'accès, faciliter le repérage et l'exploitation de celles-ci par les publics cibles...

La qualité est complexe à définir car elle doit englober l'ensemble des points évoqués. C'est donc un projet en soi qu'il convient d'engager assez tôt⁴¹.

Le contrôle qualité par échantillonnage sur les données et leurs usages⁴² reste encore trop peu pratiqué dans la réalité. Et, lorsque des données d'exploitation sont relevées, elles ne sont pas toujours analysées dans un objectif de suivi et d'évaluation. La formalisation des modèles et des schémas devrait faciliter la définition précise de critères à développer et l'établissement de processus de contrôle qualité.

Tout comme la documentation de ces référentiels, on ne peut s'attendre, pour l'avenir, qu'à des améliorations dans le domaine du contrôle qualité.

2. Métadonnées : de l'importance des modélisations

La première section de ce chapitre nous a permis de nous familiariser avec le monde des métadonnées et de la modélisation, ainsi qu'avec les différents outils à développer dans ce contexte. Mais cette approche peut donner le sentiment qu'il suffit, pour obtenir des référentiels de qualité, de suivre une démarche de gestion de projet et d'appliquer des règles de façon rigoureuse. Nous voudrions, dans cette seconde partie, montrer à travers quelques exemples que l'univers des métadonnées suppose tout à la fois :

- une posture ouverte, bien sûr sur les technologies nouvelles, en particulier l'environnement XML, mais plus encore sur les représentations et les modèles qui s'entrechoquent lors de projets d'interopérabilité (2.1.);
- une approche nouvelle des ressources numériques et de leurs usages (2.2.).

41 Voir aussi [8, p. 10].

42 Un exemple d'examen d'une application en vue de contrôler l'usage d'un référentiel de métadonnées : le cas du « contenu du moteur de recherche du Site du Canada », www.tbs-sct.gc.ca/im-gi/mwg-gtm/docs/2005/status-etat/status-etat01-fra.asp

◆ 2.1 Affiner son micromonde et s'ouvrir à d'autres mondes

Avant d'étudier ou d'exploiter un schéma de métadonnées, il est nécessaire d'analyser avec attention les modèles sous-jacents, souvent implicites, qui structurent les composants proposés.

Les parties introductives des documentations des schémas, des cadres conceptuels ou des référentiels fournissent des éléments sur les entités prises en charge et les publics destinataires de ces outils.

Le cadre conceptuel du FRBR (Functional Requirements for Bibliographic Records) modélise les fonctions de la notice bibliographique à l'attention des usagers de ce type de notice: « Répondre aux besoins des utilisateurs et [...] prendre plus efficacement en compte la diversité des besoins générés par des supports variés, ainsi que la diversité des contextes d'utilisation des notices bibliographiques. [...] tracer [...] le contour des fonctions que remplit la notice bibliographique, en prenant en compte les différents supports, les différentes utilisations, et les différents besoins des utilisateurs.⁴³ ». Un « cadre conceptuel où soient identifiées et clairement définies les entités pertinentes pour les utilisateurs de notices bibliographiques, les attributs de chacune de ces entités, et les types de relations qu'elles entretiennent entre elles.⁴⁴ »

La recommandation TEF (Thèses électroniques françaises) définit un jeu de métadonnées descriptives et de gestion pour les thèses: « Modélisation des métadonnées de thèse – un jeu de métadonnées pour les thèses électroniques soutenues en France – métadonnées connues par l'établissement de soutenance et indispensables aux autres acteurs en charge d'une mission de diffusion, de signalement ou de conservation.⁴⁵ »

Le profil d'application du LOM (Learning Object Metadata) modélise des ressources pédagogiques à l'attention des publics finals de la ressource: « Répondant aux besoins des acteurs français du secteur éducatif. [...] Une ressource pédagogique est définie comme "toute entité, numérique ou non, qui peut être utilisée pour l'éducation, la formation ou l'apprentissage" [...] A pour objectif de faciliter la recherche, l'évaluation, l'acquisition et l'utilisation des ressources pédagogiques par les apprenants, les enseignants ou les logiciens.⁴⁶ »

Le schéma DDI (Data Documentation Initiative) décrit toute la documentation d'enquête à l'attention des communautés de producteurs, utilisateurs ou archivistes: standard international de documentation

43 FRBR, p. 7-8 (voir la note 10).

44 FRBR, p. 9.

45 Groupe Afnor CG46/CN357/GE5, *Les métadonnées des thèses électroniques françaises, TEF*, 2^e éd., mars 2006, www.abes.fr/abes/documents/tef/recommandation/tef.pdf. Page 12.

46 NF Z76-040 Décembre 2006. *Technologies de l'information pour l'éducation, la formation et l'apprentissage - Profil français d'application du LOM (LOMFR) - Métadonnées pour l'enseignement*. Page 6.

des données d'enquêtes, il « permet la description d'enquêtes et de sondages en sciences humaines et sociales, de la description du projet à la description détaillée de chaque variable.⁴⁷ »

Le contexte et la finalité de ces quatre modèles expliquent aisément le périmètre des outils développés. Mais, bien que tous ces exemples de schémas ou modèles relèvent du champ du « document », nous identifions des variations importantes entre eux. Ces variations portent principalement sur :

- l'objet central pris en charge: des notices bibliographiques accompagnées des points d'accès traditionnels (FRBR), des notices étendues à des données de gestion (TEF), des ressources pédagogiques (LOM), l'ensemble des éléments d'une documentation d'enquête incluant les données d'enquête, elles-mêmes d'une granularité plus fine (DDI);
- les publics cibles: des utilisateurs en proximité forte avec l'objet modélisé, qui *a priori* ont une bonne compréhension du modèle proposé;
- le statut de ce public par rapport au « contenu » informationnel. Sur ce point les variations sont les plus grandes: plutôt gestionnaires ou intermédiaires (FRBR, TEF), producteurs des données ou utilisateurs finals (LOM, DDI) avec une prise en compte des besoins de conservation dans ces deux cas.

Nous voudrions insister sur l'importance de bien caractériser les représentations auxquelles se réfèrent les modèles ou schémas: les utilisateurs finals font largement des choix distincts de ceux faits par des gestionnaires en définissant des métadonnées représentant les « contenus », complémentaires aux métadonnées usuelles de nature bibliographique. Ce point est plus précisément abordé dans la partie 2.2.

De plus, l'étude de ces schémas montre l'importance des traces du passé. Bien souvent il s'est agi d'informatiser des méthodes et outils existants. La présence de ce passé se justifie dans un environnement où les missions restent identiques (dépôt légal, préservation) et où les contraintes de reprise des existants sont très fortes, ce qui est le cas dans de nombreux dispositifs documentaires. Dans ce contexte, il n'est pas toujours aisé de faire évoluer les modèles pour répondre aux nouveaux besoins et s'ouvrir aux technologies actuelles.

Pour étayer notre propos, considérons trois éléments précis: les listes d'autorité, la langue et les thèses.

47 www.centre.quetelet.cnrs.fr/documentation.php

48 Présentation des ISBD : www.bnf.fr/pages/infopro/normes/no-isbd.htm

49 « La bibliographie est une technique de communication qui se propose de rechercher et de rendre disponibles à des utilisateurs des références de documents ou des informations rapidement utilisables. » « Faire une bibliographie, c'est faire du service public. » Marie-Hélène PrévotEAU, Jean-Claude Utard, *Manuel de bibliographie générale*, nouv. éd, Éd. du Cercle de la librairie, 2005. Cité par Catherine Éloi, *Bulletin des bibliothèques de France*, 2006, t. 51, n° 3, p. 130.

2.1.1 Listes d'autorité : simples listes contrôlées ou bases de connaissances ?

Le monde du livre et des bibliothèques a déployé très tôt des formats techniques normalisés d'échange de l'information bibliographique informatisée, à l'origine pour réaliser des échanges entre professionnels de la bibliographie, les agences bibliographiques nationales: le format MARC (1965, Library of Congress) ou Unimarc en France (1977, IFLA). Ces schémas s'appuient sur le modèle des ISBD (descriptions bibliographiques internationales normalisées) qui « constitue un ensemble de spécifications normatives, validées au niveau international, pour la description bibliographique des documents existant dans les bibliothèques⁴⁸ ».

Ce modèle est centré sur l'objet publié et sa gestion une fois celui-ci publié (localisation, entrée, sortie). Il s'agit d'établir la description de l'objet et de placer celui-ci au sein d'une collection ou d'un fonds. Dans ce modèle, les contenus informationnels des objets sont très faiblement représentés et l'attention est portée sur la normalisation de la description de l'objet. C'est dans ce contexte que se situent les listes d'autorité.

Ce sont des listes de termes normalisés, soit de mots matières soit de noms propres, obligatoirement et nécessairement utilisés dans le catalogage et l'indexation. Ces termes normalisés, organisés au sein de fichiers d'autorité de structure simple, servent à contrôler les variantes du nom d'une entité ou l'ensemble des valeurs d'une zone donnée: noms de pays, de personnes ou d'organismes, etc.

Depuis le début des années 1970, ces mêmes outils produits pour les gestionnaires ont été proposés aux utilisateurs finals. Or ce modèle, qui décrit très faiblement le contenu, suppose que l'utilisateur connaisse déjà l'objet ou des éléments importants de cet objet (auteur, éditeur, collection, de façon générale les sources) qu'il souhaite obtenir. L'utilisateur acquiert en général cette connaissance grâce à des recherches préalables à la consultation du catalogue.

Le modèle qui s'exprime dans les spécifications ISBD ou dans l'Unimarc, mais également dans le modèle FRBR, repose sur un modèle d'usage implicite articulant une bibliographie qui permet d'identifier⁴⁹ des objets de lecture intéressants par rapport à un besoin, puis dans un deuxième temps à repérer ces ressources au travers du catalogue, pour *in fine* les récupérer au sein de la bibliothèque.

Dans ce modèle général, la bibliographie et ses répertoires conservent de façon non explicite un rôle d'intermédiaire fort.

Un catalogue de bibliothèque ne permet pas de répondre (aujourd'hui) à des demandes de type: « Je voudrais un roman japonais, d'un Japonais. »

Une question en lecture publique sur les dates (« Un roman qui se passe au Japon », sous-entendu le plus souvent « qui se passe aujourd'hui au Japon ») pointe sur un autre phénomène du modèle actuel centré sur l'objet physique « en main », et non sur la date de l'œuvre. Ce point a été analysé et fait partie des évolutions du modèle FRBR.

La date où se situe l'action du roman pose une troisième question: celle de la modélisation du contenu du document.

La notice présentée ci-après nous laisse l'espoir d'un roman se déroulant dans les années 1990 au Japon, alors qu'en réalité la date d'origine de cette œuvre est 1951.

Auteur:	Yokomizo, Seishi
Titre:	Le village aux huit tombes / Seishi Yokomizo; trad. du japonais par René de Ceccatty et Ryoji Nakamura
Titre original:	Yatzuhakamura
Éditeur	Paris: Denoël, 1993
Thème	genre: roman policier
Indice	895.63

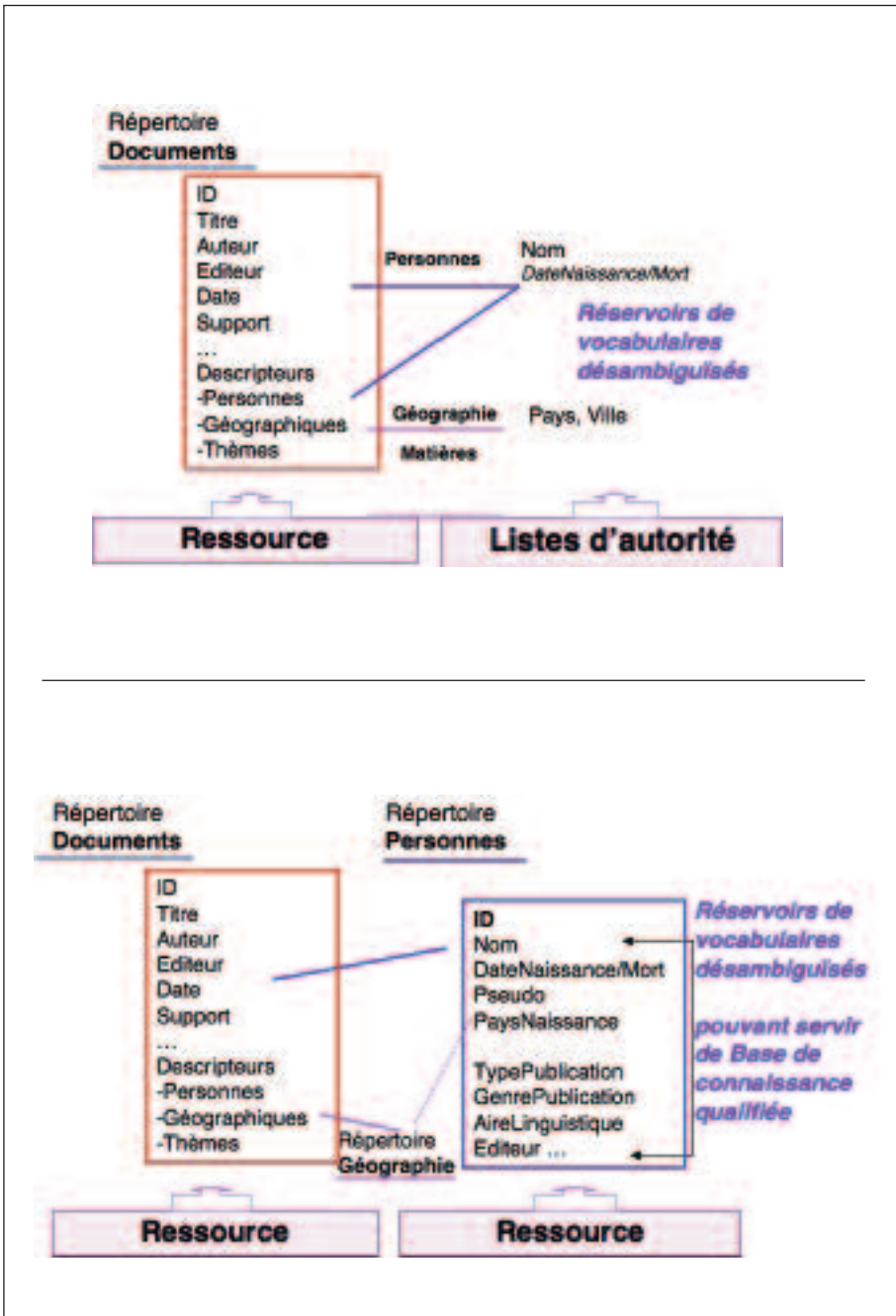
Revenons aux fichiers d'autorité. La réflexion opérée depuis plusieurs années par la profession sur ces notices se concrétise aujourd'hui avec des spécifications fonctionnelles des notices d'autorité (SFNA) associées au modèle FRBR. Ce travail étend l'information du fichier d'autorité « Personnes » à d'autres attributs (date, lieu de naissance, nationalités, langue, etc.).

Ces attributs apportés à l'entité « Personne » renforcent la fonction première de désambiguïsation des points d'accès. Mais il est également possible de s'appuyer sur l'ensemble des contenus de ces fichiers pour offrir de nouveaux accès indirects à la recherche.

Les deux graphiques de la figure 2 montrent deux approches distinctes pour modéliser la relation entre vocabulaires contrôlés et notice descriptive de la ressource. Dans les deux cas, les listes contrôlées (géographie, personnes, etc.) ont comme fonction de désambiguïser le vocabulaire fourni pour identifier les accès aux notices bibliographiques. Dans la deuxième approche, la fonction de ce réservoir est élargie: elle offre un espace sémantique exploitable à la recherche⁵⁰, et ceci quelle que soit la relation entre l'élément de donnée de la notice (auteur, éditeur ou sujet) et la base d'information constituée par ce réservoir de notices d'autorité. Les technologies actuelles permettent d'étendre les modèles d'usage de ces deux réservoirs, de façon indépendante et conjointe.

50 Un outil comme celui utilisé à la bibliothèque municipale de Lyon sur Catalog+ ou tout autre moteur d'indexation et de recherche peut exploiter ces réservoirs de connaissances pour orienter les résultats de recherche moins précise.

FIGURE 2 – MODÉLISATION DE LA RELATION ENTRE VOCABULAIRES CONTRÔLÉS ET NOTICE DESCRIPTIVE : DEUX APPROCHES



2.1.2 Représentation des langues : métadonnées composites et choix d'encodage

L'étude de l'élément de donnée « Langue » est exemplaire en ce sens qu'elle nous permet de faire le tour des problèmes à la fois de représentation et de modélisation mais aussi d'implantation informatique.

Représentation et modélisation

Le problème posé est la représentation de la langue d'une ressource, et la nécessaire distinction de ces « langues » afin qu'un utilisateur puisse sélectionner une ressource en fonction d'une langue qu'il connaît.

La logique des portails plurisources et celle des ressources multi ou plurimédias nous oblige à distinguer derrière cette unique entrée « Langue » différentes réalités en fonction des ressources :

- la langue écrite utilisée pour écrire un texte original ou celle de sa traduction (a) ;
- la langue de la mention inscrite sur une pièce ou sur une affiche ou celle utilisée sur une photo en sus de celle de la légende. Ce cas, qui rejoint le cas (a), pose deux problèmes : celui de la représentation de l'élément de donnée « Langue », mais également de la portion de texte concernée. La question de la granularité de l'unité de traitement est posée ;
- la langue parlée sur une bande-son ou une ressource audiovisuelle (b), à distinguer de celle du sous-titrage d'un film (a) ;
- la langue des annotations ou commentaires élaborés par un contributeur externe sur le contenu de la ressource.

Dans tous ces cas cités, on peut avoir des supports qui comportent plusieurs langues à la fois.

Ces questions sont bien connues des catalogueurs chevronnés, mais restent bien souvent peu ou mal identifiées.

Si l'on s'inscrit dans le monde des systèmes informatiques, du point de vue de l'utilisateur, on peut citer également :

- la langue et l'écriture utilisées pour rédiger la métadonnée. On parle alors de la méta-métadonnée (voir le schéma LOM-FR) ;
- la langue de l'interface de recherche et de l'espace intégrant les ressources ;
- la langue utilisable pour exprimer sa requête, indépendamment de l'espace documentaire exploité.

Que ce soit pour « lire » ou pour « écouter » le document, il ne suffit pas toujours de citer une langue comme le chinois, le grec, le français ou l'anglais. Dans tous les cas, le lecteur ou auditeur peut être dans l'incapacité de comprendre le contenu du document :

- dans le cas de la langue écrite, on parle de systèmes d'écriture⁵¹. Il faut faire une distinction entre la mention de la langue (grec) et celle du système d'écriture utilisé pour noter la langue (le grec ou l'alphabet latin). On « romanise » le russe, le japonais ou le chinois en le transcrivant dans notre alphabet latin. Mais quelqu'un peut connaître le russe, le japonais ou le chinois dans un système d'écriture original, mais ne pas être capable de lire la transcription (romanisation), et vice versa (cyriliser en russe);
- dans le cas de la langue parlée, un utilisateur peut parler l'anglais britannique et avoir des difficultés à entendre et comprendre un anglais américain; il en va de même pour l'ukrainien parlé en Moldavie, différent de celui parlé en Ukraine, ou encore du français parlé en Caraïbes par rapport au français québécois.

Conceptuellement, la représentation d'une « langue » est donc dépendante :

- du système d'écriture choisi (par l'auteur d'un texte, mais aussi par le gestionnaire pour présenter la référence bibliographique);
- de la langue utilisée par le locuteur telle qu'elle est parlée dans son espace culturel;
- de l'unité éditoriale et documentaire dont on parle: l'émission TV en français ou l'allocution en anglais du président de l'ONU intégrée à cette émission en français;
- de l'interface d'interrogation, de la page HTML ou de la ressource intégrée dans cette page.

Pour modéliser dans un système d'information l'attribut « Langue d'une ressource », il serait nécessaire de créer une métadonnée composite, c'est-à-dire composée de plusieurs éléments de données unitaires: la langue telle qu'elle est parlée dans un territoire et le système d'écriture utilisé dans la ressource. Les valeurs de ces éléments de données unitaires sont contrôlées par des normes. Si l'on souhaite faire de cette métadonnée composite un tout, on dira qu'il est utile de créer une métadonnée unique composée de plusieurs segments.

La norme LOM-FR comporte plusieurs métadonnées composites, dont le groupe « Contribution » composé de trois éléments de données: « Role », « Person », « Date »

Pour implanter ce modèle dans des systèmes d'information, ce micro-modèle devrait être associé à chaque unité documentaire pour lequel cette mention est nécessaire. Dans le monde du web, un document HTML constitue un tout et la mention de la langue couvre donc cette page HTML⁵².

⁵¹ Jean Meyriat, « La translittération en question », *Bulletin des bibliothèques de France*, 1993, t. 38, n° 5, p. 69-71, <http://bbf.enssib.fr>

⁵² *Les pratiques exemplaires d'internationalisation: l'indication de la langue dans les contenus XHTML et HTML*, note de groupe de travail du W3C du 12 avril 2007, www.yoyodesign.org/doc/w3c/i18n-html-tech-lang

Notation et encodage

Après avoir étudié la langue sous l'angle de sa représentation et de son modèle, nous devons aborder la question de la notation puis de l'encodage de cette métadonnée dans le schéma.

Pour la notation et l'encodage de l'élément « Langue », il existe la famille des normes ISO 639, composée de trois éléments: la norme ISO 639-1 (2002) sur deux codes et en caractères latins, bilingue français et anglais; la norme ISO 639-2 (1998) sur trois caractères incluant les éléments de la norme ISO 639-1 ainsi que les langues anciennes⁵³, la plus utilisée dans le monde documentaire⁵⁴; enfin la norme ISO 639-3 (2007) qui code sur trois caractères l'ensemble des langues connues⁵⁵. Les langues étant vivantes, les normes ISO peuvent subir quelques changements. Ces normes sont utilisées comme listes d'autorité dans les systèmes bibliographiques pour décrire la langue parlée ou écrite.

L'encodage des langues dans le monde du web⁵⁶ suit une règle qui s'est d'emblée donné un objectif – représenter et noter les langues pour tous les usages directs *via* Internet –, et une mission – celle de préserver la stabilité des codes et des noms dans le temps –, à la fois larges et précis.

La notation de l'attribut « Langue » intégré à HTML et XHTML [1] est décomposée en plusieurs segments, la notation de chacun de ces segments et l'encodage de l'ensemble constituant un composant clé de l'architecture du web. Le standard RFC 4646 (*Tags for Identifying Languages*) a succédé en 2006 au RFC 3066. Édicté par l'IETF⁵⁷ sous le contrôle de l'ICANN, ce vocabulaire est formellement enregistré à l'IANA⁵⁸.

Cette recommandation s'appuie sur trois référentiels normalisés: l'ISO 639(1) pour les langues avec une intégration prochaine de l'ISO 639-3, l'ISO 15924 pour les codes d'écritures et l'ISO 3166 pour les codes pays. Elle est complétée par la recommandation RFC 4647 qui propose des mécanismes de correspondance entre une étiquette existante et une étiquette demandée (suite à une question d'un utilisateur, pour choisir une ressource).

Cette notation est composée de deux caractères pour la sous-étiquette langue (*subtag*) désignant les langues (fr, en, be, etc.), extensibles avec d'autres codes pour représenter:

53 Précisons également qu'il existe pour certaines langues un code dit bibliographique (ISO 639-2/B) et un autre code terminologique (ISO 639-2/T) utilisé dans le monde de la linguistique. Citons le cas de l'allemand: ger (D) et deu (T).

54 www.loc.gov/standards/iso639-2/php/code_list.php

55 www.sil.org/iso639-3/codes.asp?order=639_3

56 *Les pratiques exemplaires d'internationalisation* (voir la note 52).

57 IETF RFC 4646: www.ietf.org/rfc/rfc4646.txt

58 Internet Assigned Numbers Authority (IANA) – Protocoles: www.iana.org/protocols

- la communauté culturelle ou géographique où cette langue se parle :

fr-CH (le français tel qu'il est parlé en Suisse)

en-GB (anglais britannique)

title="fr.ch - RSS-Feed" pour avoir un fil dans une langue précise

zh-TW (chinois traditionnel de Taiwan) ou zh-Hant-HK (chinois traditionnel utilisé à Hong-Kong)

es-005 (espagnol d'Amérique du Sud)

- un système d'écriture :

sr-Latn_CS: Serbe (sr) écrit dans l'alphabet latin (Latn) tel qu'il s'utilise en Serbie (CS)

Il ne suffit donc pas de posséder une métadonnée « Langue » dans deux schémas pour que l'interopérabilité et les échanges entre eux puissent avoir lieu : l'une des notations est plus riche que l'autre puisqu'elle renseigne l'ensemble des points de vue que l'on peut avoir, en anticipant les usages. Cette notation couvre donc *a priori* l'ensemble des problématiques énoncées au départ. C'est par ailleurs la recommandation pour l'élément de donnée « Langue » de la norme ISO 15836 reposant sur le Dublin Core⁵⁹ qui n'est pas celle choisie par la plupart des schémas qui sont en alignement avec le Dublin Core.

L'encodage RDF n'apporterait rien de plus dans ce cas de figure : il s'agit bien d'une question qui relève du modèle métier (étapes 2 et 3) limitant l'interopérabilité au niveau sémantique.

Derrière ces choix, qui semblent relever de critères économiques (la description très précise et complète a un coût) ou de critères techniques (la combinaison d'éléments de données est complexe à mettre en œuvre), se dessinent également des modèles d'usage différents. Dans le monde du web, tout doit être fait pour atteindre directement la ressource, alors que l'univers bibliographique ou documentaire reste encore structuré par une logique d'accès indirect à la ressource, la bibliographie puis la notice jouant chacun un rôle d'intermédiaire.

On le voit donc : la simple mention de la langue est plus complexe à analyser qu'il n'y paraît ! Et il semble bien que, pour chaque projet, il sera nécessaire de réétudier notre approche de la mention « Langue », en tenant compte à la fois de la diversité des ressources, de leur passage au numérique et des pratiques des utilisateurs finals de ces ressources.

59 DCMI Element Set (ISO Standard 15836) : <http://dublincore.org/documents/dces>

2.1.3 Représentation de la fonction « Responsabilité » et concordance entre schémas

Si les différences d'approche dans les modèles sous-jacents aux schémas sont délicates à appréhender, les variations dans les formalismes des schémas et les notations des métadonnées sont plus aisées à repérer mais tout aussi préjudiciables aux échanges.

Pour étudier ce phénomène, nous considérerons l'élément de donnée « Responsabilité » tel qu'il est représenté dans les trois schémas: Dublin Core, LOM-FR sur les ressources pédagogiques et ISO 82045-2 sur la gestion des documents techniques et commerciaux associés⁶⁰.

Le profil d'application français **LOM-FR** comme la norme ISO 82045 intègrent les communautés de contributeurs dont les auteurs, et la gestion du cycle de vie du document. Ces deux normes identifient un groupe de métadonnées « Cycle de vie de la ressource » répétable. Pour le LOM, cet ensemble d'éléments de données regroupe: le rôle des contributeurs parmi une liste d'une vingtaine de rôles liés à la production pédagogique, le nom de ces contributeurs et la date de la contribution.

Plus complexe, la norme **ISO 82045** répond aux exigences de gestion de la documentation technique: elle gère, en plus des rôles des contributeurs, des phases au cours desquelles le statut du document et les responsables peuvent varier.

Dublin Core isole, quant à lui, trois éléments de données pour noter les responsables selon trois catégories d'intervenants: « Creator », « Publisher », « Contributor », auquel il faut ajouter une date unique.

Nous voici donc avec trois niveaux de représentation et de précision différents pour trois jeux de métadonnées relatifs à une même fonction: représenter les responsabilités autour des ressources.

Le travail d'analyse de la concordance entre jeux de métadonnées doit être conduit avec attention. En particulier, il faut essayer de se détacher des dénominations pour tenir compte du périmètre réel des éléments de données, en ayant comme objectif la mise en correspondance de données avant d'envisager la création d'un nouvel élément de donnée. L'articulation entre le modèle du Dublin Core (quatre éléments de données) et celui du LOM-FR

60 ISO 82045-2:2004, décembre 2004, Gestion de documents – Partie 2: Éléments de métadonnées et modèle d'information de référence, 1^{re} édition.

La norme internationale ISO 82045, en deux parties, « spécifie des principes et des méthodes pour définir des métadonnées de gestion des documents techniques et commerciaux associés sur l'ensemble de leur cycle de vie. Ce cycle couvre généralement une plage s'étendant de l'idée conceptuelle d'un document jusqu'à sa destruction. Les principes et les méthodes mis en œuvre sont ceux de base pour tous les systèmes de gestion des documents. » Cette norme fournit un ensemble complet d'éléments de métadonnées.

(une métadonnée composite répétable) a imposé de prendre certaines dispositions particulières. La solution choisie donne la priorité au Dublin Core pour tenir compte des pratiques des indexeurs⁶¹ :

- les recommandations faites dans le profil LOM-FR préconisent, pour chaque enregistrement, d'avoir les trois rôles du Dublin Core qualifié. La concordance s'effectue entre un schéma d'encodage (la liste des rôles de contributeurs du LOM-FR de l'élément de donnée « Rôle ») et trois métadonnées dans Dublin Core ;
- pour la date, une métadonnée supplémentaire a été spécifiquement créée dans le profil français, les dates associées aux rôles dans la zone « Cycle de vie » n'étant pas exploitées pour la mise en concordance.

La notation choisie au sein du profil LOM-FR, comme pour la norme ISO 82045, prend appui sur un modèle intégrant la production de ressources et sa gestion, modèle que l'on retrouve dans les schémas de métadonnées et les référentiels liés à la gestion des documents dans une logique de traçabilité ou de réexploitation de l'information au plus proche des producteurs et des exploitants. Le format de représentation du LOM-FR offre plus de souplesse et permet de caractériser plus facilement les responsables, quels que soient le type de ressources et l'étape du cycle de vie de celles-ci.

Dublin Core, quant à lui, reprend le modèle bibliographique traditionnel, où la gestion des versions est identifiée par un élément de donnée unique, et le collectif de contributeurs resserré autour de trois rôles génériques. On note l'usage d'un même terme de « Contributeur » avec un périmètre très différent.

Si le travail de mise en correspondance donne l'impression qu'il s'agit d'une concordance entre métadonnées, cet exemple montre une fois de plus que c'est le cadre conceptuel – prise en compte ou non de l'ensemble du cycle de vie du document – qui donne au schéma de données ses caractéristiques d'hospitalité et d'ouverture, et facilite ou au contraire complexifie la mise en correspondance de schémas de métadonnées.

◆ 2.2 Du contenant au contenu

Les exemples de variations dans les modèles et les schémas, exposés rapidement dans la partie 2.1 de ce chapitre ou dans le chapitre 4 sur les normes de métadonnées, donnent à voir des micromondes centrés sur les métadonnées telles que définies à travers les notices bibliographiques. Si cette approche a du sens dans le cadre d'un repérage de premier niveau ou pour répondre à des missions de gestion et de pérennité, il paraît essentiel de

⁶¹ Profil français d'application du LOM (LOMFR), page 28 (voir la note 46).

développer une approche, si ce n'est centrée sur les contenus, au moins plus hospitalière et plus ouverte sur les contenus numériques. La convergence de ces contenus [voir la référence [13] et le chapitre 6 de cet ouvrage] et l'appel des utilisateurs [voir le chapitre 2] nous y poussent fortement.

La numérisation de l'information et des activités a permis, dès les années 1970, le rapprochement puis la fusion entre l'informatique et les télécommunications. Pour ne parler que du domaine de la documentation, l'interrogation des banques d'information professionnelles⁶² – aujourd'hui représentatives d'une part importante du web profond et invisible – a largement bénéficié de cette mutation en étendant et facilitant la mise à disposition de ces ressources et, par là même, en étendant leur audience. L'intégration des médias numériques à ce processus de convergence s'est effectuée assez rapidement, le web devenant le dispositif d'écoute/lecture pour les publics visés. « Le concept de convergence s'élargit aujourd'hui à l'ensemble des contenus même, leur dématérialisation modifiant fondamentalement leur condition d'exploitation.⁶³ »

C'est ainsi, par exemple, que nous pouvons découvrir Marie Curie en quelques clics face à notre écran [voir ci-dessous]: à partir de l'article de Wikipédia⁶⁴, de l'ouvrage sur la physicienne édité par Gallimard (qui pourrait aisément être sur support numérique), du site que lui a consacré le CNRS, associé à une exposition numérique, ou du téléfilm consacré à sa vie par Pierre Badel (Archives INA, 1965 [c'est bien 1965?]), etc.

EXEMPLE DE CONVERGENCE DE MÉDIAS NUMÉRIQUES



62 Au sujet de l'usage des banques et bases de données, voir aussi [18].

63 « Internet et audiovisuel au-delà de la convergence », *Dossiers de l'audiovisuel*, janvier 2000, n° 89.

64 http://fr.wikipedia.org/wiki/Marie_Curie

Poussés par les usagers souhaitant soit automatiser certaines de leurs tâches (recherche, rangement, diffusion, acquisition), soit se faire assister dans la réalisation d'autres activités plus centrales pour lesquelles ils préfèrent encore garder la main (production, tri des meilleures ressources, lecture et annotations, réexploitation), cette nouvelle phase dite de convergence des contenus ne peut se limiter à une harmonisation, voire à une fusion des formats des notices descriptives selon les canons bibliographiques, fussent-elles associées aux objets eux-mêmes stockés dans des entrepôts. L'impact de la numérisation atteint les contenus mêmes de ces objets, contenus sur lesquels il convient d'adopter un autre regard.

Pour dépasser le cadre de la référence bibliographique et documentaire, nous proposons d'aborder à grands traits la question de la granularité et de l'unité de traitement.

2.2.1 De l'unité documentaire à l'unité d'information

Une des premières constatations résultant de l'étude des modèles et schémas porte sur l'unité de base sur laquelle ces modèles s'appuient.

Le dictionnaire nous fournit la définition suivante de l'unité: « Élément d'un ensemble, entité (chose ou être) considéré(e) comme formant un tout indivisible.⁶⁵ »

Revenons aux modèles structurant les différentes catégories de dispositifs à caractère documentaire – édition, archives, bibliothèques, services documentaires mais aussi musées – qui nous proposent tous une terminologie spécifique: unité documentaire, unité bibliographique, unité éditoriale, unité archivistique mais aussi unité d'information (dans un hypertexte⁶⁶, dans un texte, etc.).

Tous partent d'un objet informationnel et adoptent un point de vue particulier sur cet objet. Ce point de vue se distingue sur le plan de l'unité de base et de ses composants, ainsi que sur celui des traitements et des moyens d'accès à cette unité. En creux, ces différents termes fournissent des indications sur la portée du modèle, sur ses usages et sur les audiences visées. Ils renvoient à des cadres conceptuels et des jeux de métadonnées que l'on a parfois bien du mal à aligner entre eux et à rendre, si ce n'est compatibles, au moins harmonisés pour permettre soit des recherches fédérées soit des traitements optimisés en flux continus.

65 CNRTL, www.cnrtl.fr/lexicographie

66 Bruno Bachimont, « Dossier et lecture hypertextuelle: problématique et discussion. Exemple autour du dossier patient », *Les Cahiers du numérique*, 2001, vol. 2, n° 2, p. 105-123, www.utc.fr/~bachimon/Publications_attachments/BachimontCahierNum.pdf

A. Structuration orientée par l'objet physique [figure 3, colonnes 3 et 4]

1. Modélisation éditoriale

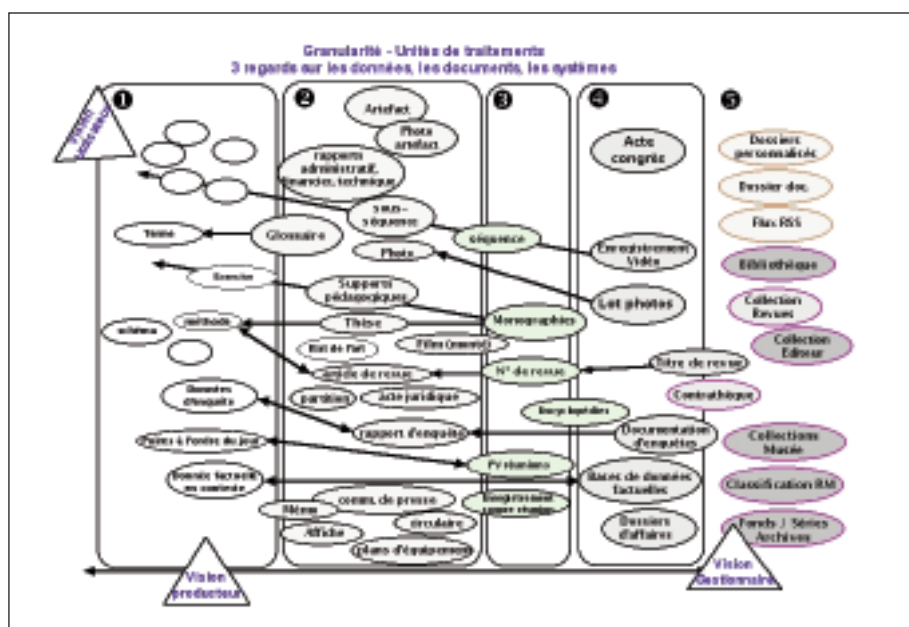
- ◆ L'unité documentaire est constituée par l'objet pris comme un tout, appartenant à une collection définie par sa filiation (Producteur, Éditeur).
- ◆ Les composants de cette structure unitaire sont la page d'accueil ou page de couverture, les parties et sous-parties nommées chapitres. Cette structure est interne à l'objet.
- ◆ Elle constitue un socle supportant la navigation intradocumentaire; elle peut être exploitée pour orienter une recherche (pondération); des outils d'accès intradocumentaires comme le sommaire et les index complètent ce schéma.
- ◆ Cette métastructure est connue par un public très large, elle est enseignée, mais elle reste d'une sémantique structurelle.
- ◆ Il existe des formats normalisés traduisant ce modèle, comme les formats de livres électroniques (*e-books*).
- ◆ Les outils de production éditoriaux ou bureautiques, ou encore les outils de lecture permettent une implantation de ce type de modèle.

2. Modélisation bibliographique

- ◆ L'unité documentaire est l'objet pris comme un tout, appartenant à une collection constituée par acquisition soit pour des raisons légales (bibliothèque de dépôt légal, bibliothèque patrimoniale), soit en vue de répondre aux intérêts de lecteurs.
- ◆ Le résultat de cette modélisation se traduit par une notice bibliographique (catalographique).
- ◆ Cette notice est distincte du document qu'elle décrit.
- ◆ La modélisation reste centrée sur le livre, les autres ressources – de la photo à la base d'information –, toujours considérées comme des « non-livres » dans les normes, sont subordonnées au schéma général élaboré avec le livre.
- ◆ Cette métastructure est connue par un public très large dont un grand nombre d'utilisateurs finals de ces objets; elle est enseignée. Mais elle reste d'une sémantique de premier niveau et n'intègre pas la sémantique structurelle mise en œuvre dans la modélisation éditoriale.
- ◆ Cette structure constitue autant de points d'accès à la ressource fondés sur la description formelle de l'objet, ce qui est pertinent pour qui connaît déjà la ressource ou une partie de celle-ci (l'auteur, par exemple) [voir ci-dessus p. 35].

- ◆ Les formats normalisés représentant ce modèle sont ceux de la famille MARC qui ont été récemment portés, avec des évolutions, dans l'environnement XML⁶⁷.
- ◆ Le cadre conceptuel récemment proposé dans le document de spécification FRBR reste centré sur le même objet, mais permet de regrouper toutes les instances d'une même création [voir le chapitre 4].

FIGURE 3 – GRANULARITÉ ET UNITÉS DE TRAITEMENT



B. Structuration orientée par les contenus [figure 3, colonnes 1 et 2]

On peut considérer que les traitements « sujets » réalisés dans le monde bibliothéconomique et de la documentation établissent un pont entre les connaissances représentées dans les ressources et leur inscription matérielle sur le support délimitant l'objet physique.

3. Modélisation Contenu de premier niveau

◆ L'unité documentaire est toujours l'objet pris comme un tout, appartenant à une collection constituée par acquisition en vue de répondre ici plus précisément aux intérêts des lecteurs (bibliothèque de lecture publique, service documentaire).

67 www.loc.gov/standards/marcxml

- ◆ Le résultat de cette modélisation se traduit par une notice bibliographique enrichie par des métadonnées facilitant la recherche en offrant une représentation synthétique du contenu de l'objet et des clés d'accès complémentaires au moyen d'indices globaux (vedette-matière globale) du sujet principal. La codification de ce sujet est réalisée par un syntagme composé de plusieurs facettes (sujet, géographique, chronologique, de forme).
- ◆ Même si les normes évoluent, la démarche bibliographique reste attachée au territoire fixé par l'objet physique.

4. Modélisation structurelle de premier niveau

- ◆ D'une unité éditoriale totalement cernée par l'objet physique manipulé (le livre, le numéro d'une revue, une émission, un disque ou une cassette, une encyclopédie, un dossier d'affaires), les traitements documentaires dans les années 1950 ont modifié la granularité de l'unité traitée en la décrochant de l'unité éditoriale, tout en conservant la logique de l'objet physique manipulé⁶⁸. Ce décrochage prenait appui sur une étude de contenu dans le sens où il s'opérait dès que l'unité documentaire distincte de l'unité éditoriale pouvait avoir un sens pour le lecteur. Ainsi les articles de revues, chapitres de livre, articles d'encyclopédie, rubriques de journal, thèmes au sein d'une émission, etc., étaient traités de façon autonome en relation inclusive avec leur source d'appartenance.
- ◆ Le résultat de cette modélisation a ajouté un niveau au modèle: collection, unité documentaire mère (numéro de revue), unité documentaire fille (article), avec des accès autonomes à chacun des trois niveaux de granularité.
- ◆ L'identification du contenu est affinée car, pour pouvoir distinguer les thématiques ou sujets d'unité documentaire pour des granularités plus fines, il est nécessaire d'effectuer une analyse de même niveau et d'attacher des descripteurs de niveau équivalents. Cette modélisation coïncide avec le développement des thésaurus.

5. Modélisation Genre

- ◆ Avec la notion de genre [11] [10] [9], les discours sur les connaissances métiers (l'information contenue dans le support document) sont « packagés » selon des formats génériques reconnus dans le domaine ou le métier qui les déploient: genres littéraires, audiovisuels et télévisuels, mais aussi techniques, juridiques, administratifs, etc.
- ◆ L'unité documentaire reste l'objet, mais elle permet des accès supplémentaires.

68 Les grandes centrales d'analyse et d'indexation (comme l'Inist) ou les producteurs de banques de données professionnelles mettaient en place des systèmes de reproduction (photocopie) à l'unité, leurs missions allant jusqu'à la fourniture du « document primaire ».

- ◆ Les genres prédéterminent des structures intradocumentaires qui offrent un autre niveau de sémantique sur l'information contenue dans le support. Nous quittons l'univers de l'édition et de la bibliothéconomie pour aller vers celui de l'édition scientifique et du producteur/lecteur intégré à un domaine ou un métier.
- ◆ Ces structures font donc sens auprès d'une communauté d'acteurs dans un environnement donné (les lettres, l'audiovisuel, la presse, l'entreprise ou tel secteur particulier).
- ◆ Ces structures sont reconnues par les utilisateurs selon des degrés variables en fonction de leur degré d'appartenance au domaine ou de leur ancienneté. L'enseignement spécialisé fait la promotion de ces formats auprès des apprenants. Certains de ces genres sont des métagenres dans le sens où ils sont génériques et exploités dans des environnements variés tels les comptes rendus de réunion; d'autres formats peuvent être imposés, comme pour les thèses dans le monde académique ou pour certains documents techniques (dossier de mise sur le marché des médicaments, appel d'offre de marché public, etc.) dans le cadre de normalisation ou réglementation dans les secteurs dits économiques.
- ◆ Dans cette approche, les objets informationnels sont appréhendés en fonction du discours qu'ils portent et qui dès lors constitue un critère précisant les contours sémantiques de l'objet. La description de l'objet, en complément de cette approche genre, est reprise du modèle bibliothéconomique.

6. Modélisation Domaine ou Métier

- ◆ L'approche par domaines structure les mémoires informationnelles en prenant appui non pas sur l'objet mais sur les activités elles-mêmes et les acteurs qui les portent, acteurs à la source de cette production informative. Les structures éditoriales ou bibliographiques traditionnelles sont ici subordonnées à cette structuration métier qui peut prendre deux formats: des modélisations du domaine en extériorité par rapport aux ressources sous la forme d'ontologies ou de taxonomies, ou bien la structure même de la ressource (corpus) représentant elle-même le modèle.
- ◆ Les éléments de nature bibliographique sont associés sous des formes variées (externes, internes à la ressource) à ces structures d'analyse et de représentation des contenus. Dans ces applications, l'interopérabilité avec les formats bibliographiques est très souvent proposée avec le formalisme du schéma Dublin Core, réduisant considérablement, lorsqu'elle n'est pas associée à la ressource, la sémantique véhiculée utilisable en recherche.
- ◆ L'indexation documentaire telle qu'elle est pratiquée dans certains lieux professionnels au plus près du domaine (bases d'information professionnelle, par exemple) s'inscrit dans cette logique; elle prend dès lors appui sur des

référentiels terminologiques métiers comme des thésaurus (très) spécialisés. Mais, pour les utilisateurs de ces informations, cette représentation reste plate car, si la sélection et la sémantique des concepts des référentiels sont intelligibles et cohérents avec le système de représentation, la sémantique des relations est soit trop pauvre soit inadéquate pour fournir une représentation efficace de ces univers. Les ontologies de représentation des connaissances ou les taxonomies métiers constituent sur ce plan des outils plus performants.

◆ Cette représentation du domaine spécialisé est connue par les acteurs du domaine qui l'ont produite ou pour qui elle a été produite; elle touche des publics variés au sein du domaine, producteurs d'information quel que soit leur niveau d'intervention, commanditaires et surtout utilisateurs finals de l'information. C'est cette re-connaissance auprès d'un public élargi qui justifie pleinement la renommée actuelle de cette catégorie de « systèmes de représentation et d'organisation des connaissances ».

◆ Il faut noter également que ces structures sémantiques, grâce à leur autonomie par rapport aux logiques descriptives des ressources, sont exploitables indépendamment du type de ressources à classer ou à représenter (données, ressources éphémères, bases de connaissances, documents traditionnels, compétences, etc.).

7. Modélisation Collection [figure 3, colonne 5]

◆ La notion de collection a comme fonction d'intégrer, de façon harmonieuse, les objets composant cette collection au sein d'un schéma d'ensemble.

◆ Ce type de modélisation – classificatoire – est pratiqué très largement dans tous les univers:

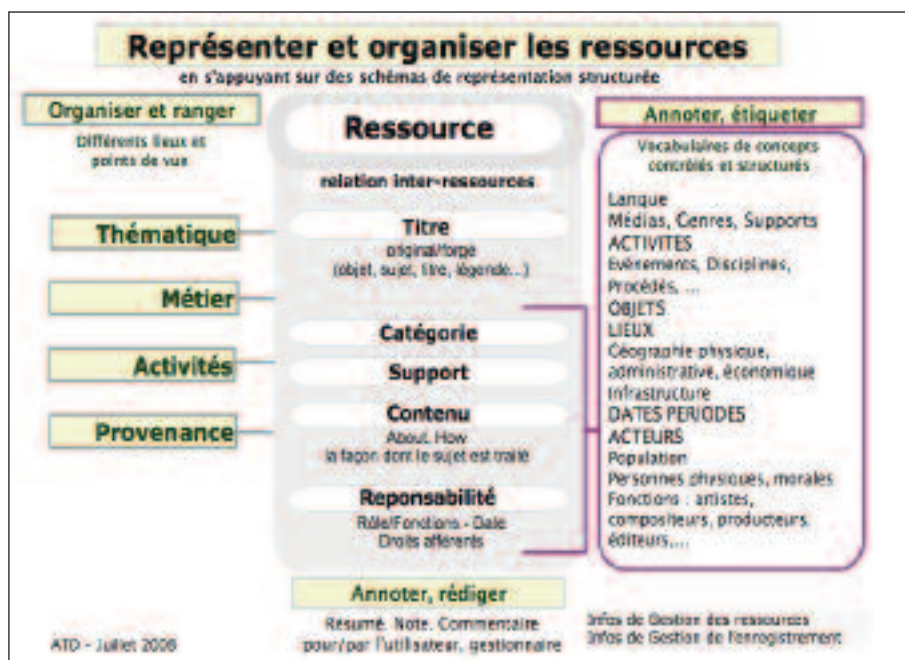
- l'unité éditoriale appartient à une collection « publication d'un éditeur »;
- des dossiers d'affaires appartiennent à un fonds, celui d'un émetteur du dossier au sein d'un service dans l'entreprise (Archives, RM);
- un objet physique est sélectionné pour être intégré à une collection de musée;
- un ouvrage est situé en rapport avec un système d'organisation, local ou central, de connaissances sectorielles ou disciplinaires.

Ces modes d'organisation renvoient à des référents de nature variée: la filiation ou le contexte de production (archives ou éditeurs ou bibliothèque de dépôt légal optent pour l'appartenance par filiation), un savoir encyclopédique (bibliothèque de lecture) ou un domaine précis (bibliothèque spécialisée, services opérationnels, etc.) auxquels il est possible d'ajouter le point de vue des utilisateurs.

Chacun de ces dispositifs adopte un point de vue différent qui conduit à proposer des services et à réaliser des traitements différents sur un objet qui, sur le plan des traitements, peut être considéré comme unique. La figure 4 montre

qu'en partant de la ressource il est possible de proposer un modèle fonctionnel intégrant les différents points de vue qui alors s'additionnent pour offrir une palette d'accès et des moyens divers de réexploitation des ressources.

FIGURE 4 – REPRÉSENTER ET ORGANISER LES RESSOURCES



2.2.2 Le cas des thèses

Pour concrétiser cette approche, nous prendrons l'exemple des thèses.

La thèse a fait l'objet d'une recommandation TEF parue en 2006⁶⁹. Cette recommandation « définit un jeu de métadonnées pour les thèses électroniques soutenues en France. [...]. Le ministère de l'Éducation nationale a voulu valoriser les thèses françaises grâce à leur diffusion électronique et engager une politique de dépôt et de conservation qui assure leur pérennité. » Le circuit pris en charge et optimisé à cette occasion démarre au « circuit administratif », c'est-à-dire après la production du contenu de la thèse, et va jusqu'au « système d'archivage pérenne ».

Dit autrement, il s'agit d'optimiser un circuit éditorial postérieur à la création de l'objet, le problème posé portant sur des métadonnées de type bibliographique ou administratif.

69 Voir la note 45.

Cette approche, qui fait écho à un réel besoin dans un cadre donné de contraintes, correspond à un modèle éditorial traditionnel postpublication. Étendre le projet de valorisation des thèses aux utilisateurs finals et à leurs besoins et pratiques fournirait d'autres pistes en termes de métadonnées.

En effet, « après le rouleau et le codex, le numérique ouvre une troisième époque de l'écrit, caractérisée par l'accès immédiat au corpus et la lecture non linéaire. [...] Il faut tenir compte en outre que le développement des lectures non linéaires a renouvelé les parcours interprétatifs propres aux usages traditionnels de l'écrit. [12] »⁷⁰ Dans ce contexte, « les projets de bibliothèques numériques actuels offrent à l'utilisateur l'accès aux thèses à partir d'une recherche qui ne permet pas d'extraire les parties pertinentes de la thèse et ne renvoie que la thèse intégrale. Ainsi l'utilisateur doit lire des chapitres entiers pour connaître les parties qui correspondent à son besoin. »⁷⁰

Les propositions évoquées dans cette recherche s'appuient sur une « insertion des connaissances propres au domaine » sous la forme d'un certain nombre de « segments sémantiques » : état de l'art, méthodologie, modèle, algorithme, architecture, prototype ou étude de cas d'une thèse scientifique.

Cette structure sémantique est utilisée ici en articulation avec la structure éditoriale générale (chapitre, sous-chapitre, etc.) et les possibilités d'annotation par l'auteur grâce à la normalisation d'étiquettes.

La structure identifiée dans ce projet est bien connue puisqu'elle fait partie de l'enseignement même de la rédaction d'une thèse. « Une bonne thèse comprend une recherche exhaustive, analysée de façon critique et rigoureuse. Elle doit inclure une description détaillée de la méthodologie utilisée. Elle doit aboutir à des résultats précis et implique une vérification systématique de toute affirmation.⁷¹ »

La thèse n'est pas évaluée directement sur la présence de ces différentes parties, mais on imagine mal une thèse qui puisse être validée sans qu'y soient clairement identifiées ces différentes parties. Il s'agit ici d'une structure sémantique liée au genre « thèse » [voir *supra* 5. *Modélisation Genre*] qui pourrait être représentée à travers un jeu de métadonnées sur le modèle intégré de la TEI (Text Encoding Initiative).

Deux avantages sont immédiatement identifiables, l'un économique, l'autre fonctionnel :

- l'indexation telle que proposée dans le projet CITHER⁷⁰ de diffusion des thèses électroniques de l'INSA de Lyon pourrait se déployer plus facilement,

⁷⁰ Projet CITHER : <http://docinsa.insa-lyon.fr/these>

⁷¹ *Qu'entend-on par thèse ou par mémoire?* Université d'Ottawa, www.grad.uottawa.ca/Default.aspx?tabid=1354

le travail amont de structuration étant mutualisé;

- l'utilisation d'un moteur d'indexation et de recherche serait optimisée par cette structure sémantique. Pour des recherches sur une méthode précise ou un type de résultat avec un renvoi précis à la partie concernée, par exemple.

En d'autres termes, en offrant des accès directs aux contenus, le numérique nous oblige à porter un autre regard sur les documents. Les technologies informatiques d'aujourd'hui nous permettent d'envisager le développement d'autres schémas à articuler avec ceux de nature bibliographique ou administrative.

Conclusion

L'univers des métadonnées, tel que nous venons de le dessiner entre données, métadonnées, registres de métadonnées et méta-métadonnées, semble bien complexe: nous sommes bien dans un monde de profusion de données qu'il nous faut apprendre à manipuler et à préserver, y compris lorsque nous ne participons pas à leur production.

Nous avons insisté également sur l'importance des visions du monde portées souvent de façon implicite par ces schémas. Toutes ces visions du monde sont justifiables et chaque modèle produit reste pertinent dans le contexte dans lequel il a été produit. Jusqu'à un certain point toutefois. Toutes ces visions du monde n'ont pas le même degré de souplesse pour prendre en compte justement les caractéristiques des documents numériques et des usages des TIC.

L'étude des différentes étapes de l'informatisation des activités d'information et de la représentation de celles-ci dans les systèmes informatiques, mises en œuvre bien sûr au sein du secteur de l'info-doc, mais plus encore celles mises en œuvre par d'autres secteurs d'activité, nous renseigne sur la situation présente et sur les orientations à prendre dans les années à venir.

Une première étape dans le développement de l'ancêtre des jeux de métadonnées a consisté à consolider, au sein d'un même environnement particulier, les modèles et les schémas existants, sans prendre en compte dans le modèle

ni les changements d'environnement ni d'autres schémas, produits il est vrai simultanément. Une deuxième étape engagée après 1990 prend appui sur une meilleure connaissance d'outils similaires dans des environnements proches. Des ponts se déploient entre communautés ayant déjà des relations anciennes: c'est ainsi que le monde des bibliothèques étudie le modèle CRM des musées, ou que des réflexions communes sont conduites autour des « notices d'autorité » entre les acteurs utilisant ces nomenclatures: archives, musées, bibliothèques.

Mais de très nombreux autres schémas et modèles exploités pour représenter et décrire des ressources documentaires, comme nous allons le constater dans le chapitre 4, se sont développés au plus près des producteurs de ces données. Et cette multiplication pose de nombreuses questions et des problèmes concrets d'interopérabilité technique et sémantique à ceux qui veulent poursuivre les deux missions structurantes du secteur de l'info-doc: celle de « mise en relation » entre ressources et utilisateurs et celle de « préservation ».

On le voit, la question des métadonnées est loin d'être close et, au-delà de l'aspect technique ou normatif, c'est certainement notre approche même de la description ou plutôt de la représentation des ressources qui doit évoluer, particulièrement en ce qui concerne la sémantique des données portées par les modèles et les éléments de données.

◆ Références

Toutes les localisations ont été contrôlées en juillet 2008.

Les localisations web de tous les jeux de métadonnées mentionnés dans ce chapitre sont récapitulées en un tableau à la fin du chapitre 4, pages 158-159.

Métadonnées et vocabulaires

- [1] Stéphane BORTZMEYER. *Les « language tags »* Exposé libre. 10 octobre 2006. www.bortzmeyer.org/files/langtags-PRINT.pdf
- [2] Jane GREENBERG, Kristina SPURGIN, Abe CRYSTAL. « Functionalities for automatic metadata generation applications: a survey of metadata experts' opinions ». *International journal of metadata, semantics, and ontologies*, 2006, vol. 1, n° 1, p. 3-20. <http://ils.unc.edu/mrc/publications/>
- [3] *Métadonnées et valorisation de l'information*. Journée d'étude organisée par l'ADBS et l'INTD-CNAM le 4 avril 2006. http://81.25.194.6/uploads/journees/4372_fr.php
- [4] Mikael NILSSON, Pete JOHNSTON, Ambjörn NAEVE, Andy POWELL. « Towards an interoperability framework for metadata standards ». In: *Proceedings of the 2006 international conference on Dublin Core and Metadata Applications: Metadata for knowledge and learning*. <http://dc2006.ucoi.mx/papers/14.00/Framework.ppt> (présentation avec schéma) ou www.scribd.com/doc/43235/Towards-an-Interoperability-Framework-for-Metadata-Standards (texte en pré-print)
- [5] Pérennisation des informations numériques (PIN), Groupe français d'études et de réflexion. http://vds.cnes.fr/pin/pin_groupe.html
- [6] John T. PHILLIPS Jr. « Metadata - Information about electronic records ». *ARMA Records Management Quarterly*, Oct. 1995. http://findarticles.com/p/articles/mi_qa3691/is_199510/ai_n8727837
- [7] Gabriella SALZANO, Abdelbasset GUEMEIDA. « Métadonnées pour les systèmes d'informations d'entreprises étendues ». In: *Journée sur les systèmes d'information élaborée, Île Rousse*, 2005. http://isdms.univ-tln.fr/PDF/isdms22/isdms22_salzano.pdf
- [8] *Understanding Metadata*. NISO Press, 2004. 20 p. www.niso.org/publications/press/UnderstandingMetadata.pdf

Méta-information Genre

[9] Jean-Paul ACHARD. *Le genre documentaire [audiovisuel]*. www.surlimage.info/ECRITS/documentaire.html

[10] François JOST. « La promesse des genres ». *Réseaux*, 1997, n° 1, 20 pages. www.enssib.fr/autres-sites/reseaux-cnet/81/01-jost.pdf

[11] François RASTIER. « L'accès aux banques textuelles - des genres à la doxa ». *Texte* [en ligne], juin 2002. www.revue-texto.net/Inedits/Rastier/Rastier_Acces.html

Information numérique

[12] Rocío ABASCAL-MENA, Béatrice RUMPLER. « Accès au contenu des thèses numériques par leur structure sémantique ». *Document numérique*, 2007, vol. 10, n° 2, p. 9-35. www.cairn.info/resume.php?ID_ARTICLE=DN_102_0009

[13] Bruno BACHIMONT. *Ingénierie des connaissances et des contenus : le numérique entre ontologies et documents*. Paris : Hermès, 2007

[14] Jean CHARLET. « Les connaissances médicales à l'épreuve de l'informatisation : entre documents non structurés et ontologies ». In: *Dossier informatisé du patient : contenu et pratiques*, réunion du groupe Dossier informatisé du patient, 27 janvier 2006. www.fing.org/jsp/fiche_actualite.jsp?CODE=1132136024576&LANGUE=0

[15] Kevin CROWSTON, Barbara KWASNIK. Wiki mis en place pour le programme de recherche « How Can Document-Genre Metadata Improve Information Access For Large Digital Collections? ». <http://genres.syr.edu/papers>

[16] *Le numérique : impact sur le cycle de vie du document*, colloque EBSI-ENS-SIB, Montréal, 2004. www.enssib.fr/bibliotheque-numerique/notice-1223 [Plusieurs problématiques sont abordées dans de nombreuses contributions : cycle de vie, imprimé/numérique, traitement et redocumentarisation, accès, etc.]

[17] Norbert PAQUEL. « Normes et standards, aspects techniques ». In: *Dossier informatisé du patient : contenu et pratiques*, réunion du 27 janvier 2006. www.fing.org/jsp/fiche_actualite.jsp?CODE=1132136024576&LANGUE=0

[18] Naoum SALAMÉ. « Les banques de données scientifiques dans l'enseignement de la biologie-géologie ». In: *L'intégration de l'informatique dans l'enseignement et la formation des enseignants*, partie sur les Banques et bases de données, coédition EPI-INRP, 1992. www.epi.asso.fr/revue/dossiers/d12p163.htm

Logiciels de gestion de métadonnées

[19] Évaluation d'éditeurs de métadonnées: www.normetic.org/Evaluation-d-editeurs-de.html

[20] Norm FRIESEN. *Quelques références d'outils*. Traduit par Karin Lundgren et Suzanne Lapointe. 2005. www.cancore.ca/editeurs.html

[21] *Étude des outils de gestion de ressources numériques pour l'enseignement ou LCMS (Learning Content Management System)*, étude réalisée pour le ministère de la Jeunesse, de l'Éducation nationale et de la Recherche par la société Business Interactif. Octobre 2003. www.educnet.education.fr/chrgt/EtudeLCMS-20030526.doc.

Bibliographie générale complémentaire

[22] Jacinthe DESCHATELETS. *Dossier sur les métadonnées*. Mis à jour le 29 mai 2008. www.bibliodoc.francophonie.org/article.php?id_article=172

[23] Annie BASSINET. Marie-Noëlle CORMENIER. Geneviève VIALA. *Indexation de ressources: métadonnées, normes et standards*. Mis à jour en mars 2007 (mise à jour variable). www.educnet.education.fr/dossier/metadatas/default.htm

[24] *La modélisation: pourquoi l'intégrer dans les systèmes d'information documentaire?* Journée d'étude ADBS, Paris, 20 mai 2003. www.adbs.fr/la-modelisation-pourquoi-l-integrer-dans-les-systemes-d-information-documentaire-journee-d-etude-adbs-paris-20-mai-2003-15141.htm?RH=ACCUEIL

[25] *Livre électronique, livre numérique*. Dossier SDTICE. Mis à jour le 1er juillet 2008. www.educnet.education.fr/dossier/livrelec/default.htm

[26] Patrick LE BŒUF. « ...“Qui se pavane et s'agite son heure sur scène et qu'ensuite on n'entend plus...”: les éléments à prendre en compte dans un futur modèle conceptuel du spectacle vivant et de l'information relative à ses archives ». Journée d'étude *De la conception à la survie: comment documenter et conserver les productions du spectacle multimédia?*, Paris, 13 janvier 2006. cidoc.ics.forth.gr/docs/2006_LeBoeuf_fr.pdf

Table des matières

Introduction

LISETTE CALDERAN	5
------------------------	---

Chapitre 1

Représentation et accès à l'information : transformation à l'œuvre

SYLVIE DALBIN	9
1. Métadonnées : processus de création et administration	10
1.1 Métadonnées : l'aboutissement d'une démarche	10
1.1.1 Métadonnée = méta + donnée	10
1.1.2 Des métadonnées : pour quoi faire?	12
1.1.3 Les étapes clés conduisant aux métadonnées	13
1.2 Du domaine normatif au domaine applicatif	21
1.2.1 Complexité de la mise en œuvre d'applications	21
1.2.2 Avec quels outils produire des métadonnées?	24
1.3 Administration des métadonnées et des registres de métadonnées	25
1.3.1 Jeu de métadonnées, registre ou référentiel de métadonnées, profil d'application	26
1.3.2 Documentation des métadonnées et des registres de métadonnées	27
1.3.3 Enregistrement du schéma de métadonnées	29
1.3.4 Quelques mots sur la qualité des données et des métadonnées	31
2. Métadonnées : de l'importance des modélisations	32
2.1 Affiner son micromonde et s'ouvrir à d'autres mondes	33
2.1.1 Listes d'autorité : simples listes contrôlées ou bases de connaissances?	35
2.1.2 Représentation des langues : métadonnées composites et choix d'encodage	38

2.1.3 Représentation de la fonction « Responsabilité » et concordance entre schémas	42
2.2 Du contenant au contenu	43
2.2.1 De l'unité documentaire à l'unité d'information	45
2.2.2 Le cas des thèses	51
Conclusion	53
<i>Références</i>	55

Chapitre 2

Moteurs de recherche : des enjeux d'aujourd'hui aux moteurs de demain

OLIVIER ERTZSCHEID	59
J'ai dix ans	59
Giant Global Graph?	60
1. Des machines sociales	60
1.1 Description, restitution, prescription	60
1.2 Algorithmes sous le moteur	62
1.2.1. Fièvre algorithmique	63
1.2.2 Complexité algorithmique objectivée ou panoptique subjectif?	63
1.3 Concentré d'économies et économie concentrée...	64
1.4 Une diversité de services... au service d'une confusion des pratiques	65
1.5 Un moteur. Des recherches	66
2. Dérive des continents documentaires et recherche universelle	67
2.1 Dérive des continents documentaires	67
2.2 Le web comme base de données	68
2.2.1 Pages statiques et web dynamique	69
2.2.2 Dans la base des moteurs	69
2.2.3 Bases de moteurs et flux de données	70
2.2.4 Read Write Web	70
2.3 La recherche universelle ou l'algorithmie ambiante?	71
2.3.1 De la recherche universelle...	71
2.3.2 ... à l'algorithmie ambiante	73

2.4 Mon nom est personne	73
2.5 Indexation marchande et indexation sociale: le thésaurus comme trésor	74
2.5.1 Indexeurs sans le savoir	75
2.5.2 L'échec des balises <meta>	75
2.5.3 Standardisation et communauté métier	75
2.5.4 Folksonomies: le retour de la communauté comme indexeur	76
2.5.5 Ontologies et web sémantique	76
3. Si loin... si proche. Rêves et réalités motorisés	77
3.1 Rêve de Web Operating System	77
3.1.1 Webtop	77
3.1.2 Collectif	78
3.1.3 Mixage	78
3.2 Rêve d'implicite	78
3.2.1 L'importance du chemin	78
3.2.2 Myware + everywhere	79
3.3 Rêve sémantique	79
3.3.1 Moteurs sémantiques: l'approche « top-down »	80
3.3.2 Moteurs sémantiques: l'approche « bottom-up »	80
3.3.3 Petite revue de troupes	81
3.4 Une question de génération...	83
3.5 Rêve de synchronicité	84
Conclusion	85
<i>Références</i>	86

Chapitre 3

Analyse des usages pour améliorer l'accès aux ressources

ANNE BOYER	89
1. Une approche de la recherche documentaire par les usages	90
2. Personnalisation de services numériques par analyse des traces d'usage	93
3. Systèmes de recommandation et filtrage collaboratif	97

3.1 Introduction au filtrage collaboratif	98
3.2 Principe général du filtrage collaboratif	99
3.3 Schéma général d'un algorithme de filtrage collaboratif fondé sur la mémoire	103
4. Défis du filtrage collaboratif	105
5. Conclusion	107
<i>Références</i>	109

Chapitre 4

Métadonnées et normalisation

SYLVIE DALBIN	113
1. Des cadres conceptuels pour représenter les données	116
1.1 FRBR pour la description bibliographique	116
1.2 CRM pour la documentation muséographique	119
1.3 Rapprochement entre FRBR et CRM : un autre travail de modélisation	120
1.4 Pérenniser les documents d'archives: la norme OAIS	121
1.5 Conclusion	123
2. Le monde de la référence des documents	123
2.1 Les formats centrés sur la description de l'objet	124
2.1.1 Description bibliographique dans le monde des bibliothèques: RDA et MODS	124
2.1.2 Formats de présentation réduite	125
2.1.3 Élargissement vers des fonctions administratives: la norme sur les thèses TEF	126
2.2 Le cas du Dublin Core	127
3. Le monde des documents numériques	129
3.1 Autour des livres numériques	129
3.1.1 DAISY	129
3.1.2 ePub Books de l'IDPF	130
3.1.3 DocBook	131
3.1.4 Conclusion	132
3.2 Informations d'enquête: DDI	132
3.3 Structure générale d'un document XML textuel: la TEI	133

3.4 Information archivistique: EAD	135
3.5 Information d'actualité	136
3.6 Conclusion	137
4. Systèmes de représentation de concepts et de dictionnaires	138
5. Fonctions de réservoir, transport et pérennisation	142
5.1 Le protocole OAI-PMH	142
5.2 Le conteneur XMP	143
5.3 Le schéma de transfert et de stockage pérenne METS	144
5.4 Et les services?	145
6. Composants transversaux	146
6.1 Numérotation et identifiants	146
6.2 Microformats	147
6.3 Droits et gestion des droits	148
7. Une famille de schémas: exemple du secteur de l'éducation	149
8. Conclusion	152
8.1 Sur le plan technique	152
8.1.1 Formalisme des normes	152
8.1.2 Schéma de métadonnées ou composants	153
8.1.3 Correspondance plutôt qu'alignement	153
8.2 Sur le plan des métiers	154
8.3 Quel terrain pour la normalisation?	156
<i>Références</i>	160

Chapitre 5

Des métadonnées à la description des ressources

Les langages du web sémantique

BERNARD VATANT	163
1. À propos de...	163
2. Des ressources, et de leur description	165
2.1 Du document à la ressource	165
2.2 RDF comme langage de métadonnées	167
2.2.1 RDF sur un exemple	168
2.2.2 Extensibilité, monde clos et monde ouvert	169

2.3 RDF comme langage de description généralisé	171
3. Questions et réponses	172
3.1 Quelle(s) syntaxe(s) pour RDF?	172
3.2 Pourquoi utiliser des URI http?	173
3.2.1 Le problème httpRange-14	173
3.2.2 « Slash », « Hash » et négociation de contenu	173
3.3 Comment construire de « bonnes » URI?	174
3.3.1 Les bonnes URI sont stables	175
3.3.2 Les URI sont des identifiants opaques... en principe	175
3.4 Qui peut dire quoi sur quoi?	176
3.5 Comment gérer la coréférence?	176
4. Des vocabulaires RDF, et de leur emploi	177
4.1 Vous avez dit ontologie?	177
4.2 RDFS: décrire simplement	179
4.3 OWL: décrire finement et raisonner	181
4.4 SKOS: classer, indexer, rechercher	183
4.5 RDFa: intégrer la description dans l'hypertexte	186
4.6 SPARQL: interroger le graphe RDF	187
5. Bonnes pratiques du web sémantique	189
5.1 Audit des vocabulaires: distinguer le terme du concept	189
5.2 Réutiliser et relier: la logique « Linked Data »	190
5.3 Réutiliser les ontologies génériques	190
5.4 Réutiliser les données publiées...	191
5.5 ... et publier ses propres données!	191
6. Le web social-sémantique est en marche	192
<i>Références</i>	193

Chapitre 6

Audiovisuel et numérique

La reconstruction éditoriale des contenus

BRUNO BACHIMONT	195
1. Le document numérique	197
1.1 Le contenu, l'inscription, le document	197

1.1.1 Contenu	197
1.1.2 Inscription	197
1.1.3 Document	197
1.2 Les dimensions documentaires	198
2. Le document audiovisuel	203
2.1 Qu'est-ce qu'un contenu audiovisuel?	203
2.2 Les spécificités de l'audiovisuel	203
2.2.1 Les contraintes de l'image	203
2.2.2 Les contraintes des séquences temporelles	207
2.2.3 Les contraintes de l'audiovisuel	208
3. L'indexation	209
3.1 Indexation documentaire ou indexation traditionnelle	209
3.2 Interpréter et manipuler	209
3.3 L'indexation fine des contenus	211
3.4 Les différents types d'indexation	214
4. L'éditorialisation	216
4.1 De l'indexation fine à l'éditorialisation	216
4.2 Différentes postures pour l'éditorialisation	216
4.3 Assister l'éditorialisation	218
4.4 Gérer les ressources	218
5. Conclusion et perspectives	219
<i>Références</i>	221

Chapitre 7

Méta-information et économie numérique

FRANÇOIS MOREAU	223
1. Les fondements économiques des industries de biens informationnels	224
1.1 Les caractéristiques économiques des biens informationnels...	224
1.2 ... et leurs conséquences	225
2. Le numérique change la donne	226
2.1 D'une vente au titre à un accès illimité: une évolution souhaitable du point de vue du bien-être collectif	226
2.2 Le déplacement de la valeur vers des biens et services rivaux	229

2.3 Un déplacement de la valeur vers la méta-information	231
3. La théorie de la longue traîne et le rôle de la méta-information	234
4. Conclusion	237
<i>Références</i>	239

Chapitre 8

Le futur du web à la lecture des recommandations du W3C

FABIEN GANDON	241
1. Web simple: toile de fond et historique	242
La naissance du web	244
2. Web structuré: la séparation du fond et de la forme	246
Du web des documents au web des données	247
Des langages pour la manipulation de XML	249
3. Web sémantique: du web qui donne à penser au web qui pense	250
Passerelles avec le web classique et le web structuré	252
4. Web sécurisé: l'assurance du surfeur	253
5. Web applicatif: vers un nouveau lieu de présence et d'action, une machine virtuelle mondiale	254
6. Web multimodal: les nouveaux visages des navigateurs	256
Les nouveaux canaux du web	258
7. Web mobile: la toile se bouge	259
8. Web accessible: des surfeurs libres et égaux en droits	260
9. Une toile perpétuellement inachevée	261
<i>Références</i>	264
Références générales	264
Recommandations du W3C	264
Notes du W3C	268
Brouillons de travail du W3C	269

Les auteurs	273
------------------------------	------------