



<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.

Chapitre IV – Métadonnées et normalisation

TABLE DES MATIERES

1. Des cadres conceptuels pour représenter les données	5
1.1 FRBR (Functional Requirements for Bibliographic Records)	5
1.2 CRM (Conceptual Reference Model) pour la documentation muséographique	6
1.3 Rapprochement entre FRBR et CRM : un autre travail de modélisation	8
1.4 Pérenniser les documents d'archives : la norme OAIS	8
2. Le monde de la référence des documents	10
2.1 Les formats centrés sur la description de l'objet	10
2.1.1 Description bibliographique dans le monde des bibliothèques : RDA et MODS	10
2.1.2 Formats de présentation réduite	11
2.1.3 Elargissement vers des fonctions administratives : la norme sur les thèses TEF	12
2.2 Le cas du Dublin Core	12
3. Le monde des documents numériques	14
3.1 Autour des livres numériques	14
3.1.1 DAISY (Digital Accessible Information System)	14
3.1.2 ePub Books de l'IDPF	15
3.1.3 DocBook V5.0 (06 Feb 2008)	15
3.2 Informations d'enquête : DDI (Data Documentation Initiative)	16
3.3 Structure générale d'un document XML textuel : la TEI (Text Encoding Initiative)	16
3.4 Information archivistique : EAD (Encoded Archival Description)	17
3.5 Information d'actualité	18
3.6 En conclusion	19
4. Systèmes de représentation de concepts et de dictionnaires	21
5. Fonctions de réservoir, transport et pérennisation	24
5.1 Le protocole OAI-PMH	24
5.2 Le conteneur XMP (eXtensible Metadata Platform)	25
5.3 Le schéma de transfert et de stockage pérenne METS	25
5.4 Et les services ?	26
6. Composants transversaux	27
6.1 Numérotation et identifiants	27
6.2 Microformats	28
6.3 Droits et gestion des droits	28
7. Une famille de schémas : exemple du secteur de l'éducation	30
8. En conclusion	32
8.1 Sur le plan « technique »	32
8.2 Sur le plan « métiers »	33
8.3 Quel terrain pour la normalisation ?	34
9. Annexes	36
9.1 Localisation Web des jeux de métadonnées abordés dans les chapitres 1 et 2	36
9.2 Références	37

Figures

Figure 1 - FRBR – Exemple établi à partir du catalogue de la Cité de la Musique	5
Figure 2 – Cartouche des documents techniques	11
Figure 3 – Modèles de représentation de concepts et termes	22
Figure 4 – XMP Right Management Schema	28
Figure 5 – Schémas et spécification dans le secteur de l'éducation	31
Figure 6 – Réservoir de métadonnées	34

Les activités dès les années 1970-1980 autour de la documentation d'entreprise, de la gestion électronique des documents ou du records management, puis celles plus récentes développées autour de la documentation audiovisuelle ont élargi considérablement les fonctions et la nature des documents pris en charge dans les dispositifs documentaires. Mais dans ces différents contextes nous sommes encore restés très attachés à une notion de la métadonnée centrée sur des pratiques descriptives visant à repérer un objet, le localiser dans un fonds, une étagère ou une boîte, le traitement s'effectuant selon une approche « boîte noire ». Le document et son contexte sont décrits ; l'objet est manipulé mais reste très peu visité : la structure interne du document qui pourrait être exploitée pour guider plus efficacement l'utilisateur n'est pas étudiée. Quant à l'identification du sujet servant à repérer un objet parmi des millions, cette représentation synthétique reste inopérante pour explorer plus avant, voire réexploiter ces contenus.

Les travaux autour de la documentation des systèmes qualité, les caractéristiques de dispositifs comme les Observatoires, le travail de groupe ou la gestion de processus (workflow), enfin la question de la conservation à long terme de l'information numérique, ont fait basculer des professionnels de l'infodoc dans le monde de la production des documents et de la gestion de cette production, avec une nouvelle exigence : celle de s'approprier plus avant les contenus des documents et leurs structures.

Plus récemment, c'est la mise à disposition des livres électroniques qui questionne d'autres catégories de professionnels de l'infodoc sur la question récurrente de l'articulation entre référence bibliographique et contenu.

Enfin le monde du Web avec son approche spécifique de la notion de « ressource » (voir Vatant 2.) rend encore plus floues certaines frontières que les professionnels de l'infodoc ont construites au fil du temps, frontières entre eux – Archives, Bibliothèque, Documentation, Musée, mais également frontières avec les objets documentaires pris en charge et leur environnement de production ou d'utilisation. Les travaux sur les modèles, les schémas de métadonnées ou les dispositifs d'accès à l'information – entrepôt, portail, intranet,..., montrent que des synergies sont possibles au sein du secteur de l'infodoc, mais qu'elles le sont également avec les environnements producteurs d'information dès lors que l'on accepte de regarder autrement les documents.

Quelques exemples de cette évolution vers le document numérique

Dans de très nombreux secteurs ou activités (archéologie, observatoires, enquêtes, géographie,...), la pratique des bases de données, seules à même de répondre aux volumes et caractéristiques des données à manipuler, se sont multipliées depuis le début de l'informatique. Ces systèmes d'information intégrant une extrême diversité de données sont une réalité quotidienne depuis plusieurs années déjà pour de nombreuses catégories d'utilisateurs qui produisent ces données et systèmes, les consultent, les exploitent. Toutefois l'utilisation de ces mêmes données sous format « livre papier » considérés jusque-là comme des publications, des rapports ou des documents de référence, perdure ou reste alors limitée à la production d'états, et ceci de façon manière très variable suivant les secteurs voire les types d'organismes,.

Dans d'autres environnements de travail qui ne relèvent pas d'une logique d'édition et de publication tel qu'évoquée, les bases d'information et les métadonnées au cœur de celles-ci correspondent dans le monde analogique à des données réparties sur différents supports. Seule la convergence sur un même support informatique permet de manipuler avec efficacité ces ensembles de données réparties par la production et les producteurs, mais rassemblées par l'usage sur un même espace. Le dossier médical (consignes médicales, analyse radiologique, ...) est un bon exemple de cette problématique¹.

Dans ces différents contextes, l'exploitation quotidienne de bases d'information diminue considérablement l'utilisation des documents sources papier traditionnellement pris en charge dans des bibliothèques ou services d'archives. Les traitements documentaires sur ces documents sources se centrent dès lors sur des questions de stockage et de conservation à long terme ou de suivi de la preuve ; les fonctions de gestion et d'administration des données elles-mêmes

¹ « Un hyperdocument particulier : le dossier » in Ingénierie des connaissances et des contenus, Bruno Bachimont, Hermès/Lavoisier, 2007, p.182

étant reportées sur la base d'information. La base d'information devenant un document à gérer et exploiter pour les utilisateurs et les gestionnaires.

Parallèlement ou concomitamment au développement des bases d'information, se sont déployées, dès le début des années 1980, des pratiques de gestion autour des documents numériques (GED ou GEIDE). La dématérialisation consistait ici à produire une image de certains documents pour la mettre à disposition sur les réseaux. Mais très rapidement le développement de la production d'information numérique a permis d'envisager une filière de production de métadonnées à la source : feuilles de style et macrocommandes permettaient dès 1990 d'intégrer dans les fichiers bureautiques les métadonnées, métadonnées récupérables directement dans des bases de données. Nous assistons aujourd'hui à une dématérialisation plus importante du contenu de ces documents faisant dire à certains que les « livres sont des bases de données »²

Nous considérerons dans ce chapitre que le basculement des données d'un support « document papier » à un support structuré de type « base d'information » ou plus globalement « document numérique structuré », s'il transforme les méthodes, les techniques et les compétences des utilisateurs et des administrateurs n'a pas retiré aux professionnels de l'information leurs missions de qualification, de mise à disposition et de préservation à long terme des données.

C'est dans ce cadre élargi que nous aimerions aborder ce présent chapitre sur les métadonnées.

Bien sûr, un panorama sur les métadonnées dans un tel contexte élargi semble utopique sans quelques limites.

Nous pourrions alors cadrer cette étude aux schémas labellisés ISO. Mais effectuer une telle sélection nous conduirait en premier lieu à exclure de fait des schémas très utilisés dans la profession mais non normalisés tels que celui de l'IPTC (section 3.5.) dans le domaine de la presse ou le schéma EAD des Archives (section 3.4.). Une recherche sur votre moteur favori vous ramènerait instantanément d'autres documents normatifs dans ce domaine : SMPTE (cinéma et télévision), livre électronique,...

Toutefois le simple label ISO est riche et élargirait cette liste de normes à étudier à des schémas en provenance d'autres environnements professionnels. Pas moins de 6 Comités techniques ont produit une ou des normes centrées sur les métadonnées dans des contextes variés : les documents techniques (TC10 - Technical product documentation), les documents liés au commerce, l'industrie (TC 154-Processes, data elements and documents in commerce, industry and administration), l'information géographique (TC211-geographic information) ou l'administration (JTC 1-Information technology) ou les ressources pédagogiques (JTC1 SC36 pour l'Education) à côté des normes issues du Comité technique Information et documentation (TC46-Information and documentation).

Nous proposons dans ce chapitre une cartographie de jeux de métadonnées ou de certains de ces composants sur la base de ce triple élargissement – le document numérique sous des formats variés, la documentation (documents techniques) vue par d'autres secteurs que celui de l'infodoc et des normes proposées par des acteurs de la normalisation autre que l'ISO. Cette liste loin d'être exhaustive, se structure autour de quelques catégories qui pourraient être exploitées pour poursuivre ce premier travail d'inventaire, et de l'étendre à d'autres schémas :

- Des cadres conceptuels pour représenter les données
- Le monde de la référence des documents
- Le monde des documents numériques
- Systèmes de représentation de concepts et de dictionnaires
- Réservoir, transport et pérennisation
- Composants transversaux

² Le livre est une base de donnée, 26 février 2008, Hubert Guillaud.
<http://lafeuille.homo-numericus.net/2008/02/le-livre-est-une-base-de-donne.html>

1. Des cadres conceptuels pour représenter les données

Cette catégorie d'outils à caractère normatif s'inscrit dans l'Étape de formalisation du modèle conceptuel métier tel qu'exposé dans le premier chapitre (A.A.3).

Les « principes directeurs » proposés donnent un cadre de travail permettant de préciser le périmètre de dispositifs ou d'applications informatiques à développer, et d'en définir des caractéristiques fonctionnelles et techniques.

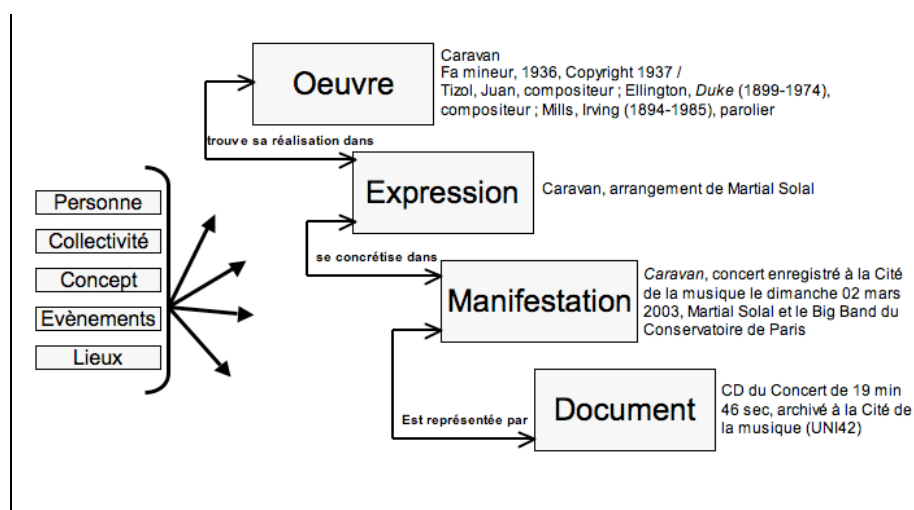
1.1 FRBR (Functional Requirements for Bibliographic Records)

Développé au sein de l'IFLA dès 1997, le modèle FRBR (Spécifications fonctionnelles des notices bibliographiques ou SFNB) est un langage d'analyse et de conception non normalisé, permettant de modéliser une description bibliographique d'un objet documentaire (ouvrage, photographie, ...).

Ce modèle s'appuie sur un ensemble d'entités réparties en 3 groupes :

- Groupe 1 centré sur les objets représentés : l'Œuvre qui ne fait pas référence à un objet matériel particulier ; l'Expression d'une œuvre qui correspond à une réalisation individuelle de l'œuvre sous une forme sonore, visuelle, textuelle... ou une combinaison de ces formes ; la Manifestation de l'expression d'une œuvre qui correspond au niveau de matérialisation sur un support d'enregistrement de l'expression d'une œuvre ; et enfin le document ou Item, c'est-à-dire l'exemplaire isolé. Ces 4 niveaux s'emboîtent les uns aux autres.
- Le Groupe 2 d'entités est composé des Personnes et collectivités. Il s'articule avec la mention de Responsabilité d'un objet du Groupe 1 ou à l'entité « Sujet » du Groupe 3 ;
- Le Groupe 3 regroupe les entités Concept, Objet, Évènement, Lieu.

Figure 1 - FRBR – Exemple établi à partir du catalogue de la Cité de la Musique³



Les Entités (et catégories d'entités) sont caractérisées par des attributs (titre, identifiant, audience d'une œuvre, état matériel pour les objets ; date et lieu de naissance pour des Personnes,...). La détermination de ces attributs provient « d'une analyse logique des données normalement exprimées dans les notices bibliographiques » (2.1.1.). Entités et attributs sont reliés entre eux par différents types de relations.

Cet ensemble organisé et structuré – entité, relation, attributs – constitue un modèle encore incomplet selon les dires des développeurs ; en particulier tous les besoins concernant les entités des Groupes 2 ou 3 ne sont pas encore totalement couverts. Mais cet outil méthodologique permet déjà de conduire des réflexions préalablement au développement d'applications. Par exemple en fonction des exigences

³ Catalogue de la médiathèque de la Cité de la Musique - <http://mediatheque.cite-musique.fr>

en termes de suivi des droits sur les œuvres, il est possible de déterminer si l'application doit couvrir trois ou quatre de ces niveaux, le niveau Œuvre pouvant être dissocié formellement du niveau Expression⁴.

Une fois le modèle conceptuel établi, il faut le faire connaître auprès des praticiens, le déployer en l'articulant avec les outils opérationnels fonctionnant sous d'autres logiques ou approches, établir des passerelles, l'affiner et le consolider. Par exemple :

- Une mise en correspondance avec les éléments des ISBD (descriptions bibliographiques internationales normalisées) et les entités du modèle FRBR⁵ a été établie par un groupe adhoc au sein de l'Ifla
- Le modèle FRBR est complété par les modèles conceptuels plus récents pour les entités du Groupe 2 par les FRANAR (Functional Requirements of Authority Numbering and Records) et pour ceux du Groupe 3 par les FR SAR (Functional Requirements for Subject Authority Records).
- Le modèle conceptuel lui-même fait l'objet d'un travail d'explicitation de même nature que celui proposé avec le modèle muséographique CRM en prenant appui sur une démarche objet. Ce travail a abouti à une nouvelle version du FRBR, FRBRoo⁶.

Munis de ces outils conceptuels, que peut-on faire ?

Si la modélisation aboutit à un modèle que tout le monde s'accorde à louer pour sa clairvoyance, sa pertinence et ses potentialités applicatives, la transformation des fonds existants et des applications et logiciels, et leur alignement à ce modèle reste une question épineuse. Projets de recherche et réalisations se multiplient :

- L'OCLC en raison de l'impact du FRBR sur les principes même de la description bibliographique et du catalogage, a engagé des actions de recherche en partenariat avec l'Ifla⁷
- La médiathèque de la Cité de la Musique (Paris) a appliqué pour partie ce modèle, offrant ainsi à l'utilisateur du catalogue, des liens croisés entre œuvres et expressions, manifestations ou documents qui lui sont rattachés, accès qui auraient nécessité de nombreuses requêtes dans un système plus conventionnel.

Plusieurs années sont utiles pour développer un modèle, le proposer à la collectivité, l'améliorer et le consolider, permettre aux praticiens et à tous les acteurs du secteur de s'approprier ces nouvelles approches, de dessiner des pistes et des orientations... On le voit, le chemin est long et le déploiement d'applications démarre ... au moment où un autre formalisme, intégré à la version FRBRoo, semble plus efficace pour construire des ponts avec un autre modèle dans le monde muséographique.

Nous retiendrons que dans le monde très changeant dans lequel les technologies de l'information nous conduisent, il est illusoire d'attendre une hypothétique stabilité d'un format ou d'une norme avant de l'exploiter. Il s'agit au contraire de participer à ces travaux pour être à la source de ces changements et ainsi les intégrer plus facilement dans ces dispositifs.

1.2 CRM (Conceptual Reference Model) pour la documentation muséographique

Le CRM (Conceptual Reference Model) a été élaboré par le Documentation Standards Working Group du CIDOC (Comité international pour la documentation), émanation du Conseil international des Musées (ICOM). Il est devenu une norme ISO en 2006 (ISO21127:2006).

⁴ Il est bien question ici de la gestion des droits, et non de l'information déclarative sur les droits effectifs des documents possédés, qui elle est toujours présente dans les systèmes d'information.

⁵ Mapping ISBD Elements to FRBR Entity Attributes and Relationships, 28/07/2004
<http://www.ifla.org/VII/s13/isbdrg/index.htm#frbr> ; <http://www.ifla.org/VII/s13/pubs/ISBD-FRBR-mappingFinal.pdf>

⁶ FRBR object-oriented definition and mapping to the FRBRER. Version 0.6.7 », International Working Group on FRBR and CIDOC CRM Harmonisation, Martin Doerr, Patrick Le Bœuf (Eds), August 2006.

⁷ <http://www.oclc.org/research/projects/frbr/default.htm>

Ce modèle conceptuel de référence vise à expliciter le sens des informations relatives aux objets patrimoniaux que l'on trouve dans des musées ainsi que des objets de nature patrimoniale tels les sites ou les monuments qui peuvent apparaître comme des objets moins classiques. La fonction de ce modèle est identique à celle des FRBR : expliciter la logique qui sous-tend les objets pris en charge dans les musées ou ceux des bibliothèques. Mais ce modèle va ici plus loin, en cherchant à faire « émerger la signification réelle de tout ce qui est considéré comme « implicite » et « évident » dans une structuration d'informations.

Nous retiendrons ici deux spécificités de ce modèle.

En premier lieu au cœur du modèle se trouve non pas la notion d'œuvre mais celle d'événements et de phénomènes temporels. La finalité du travail de documentation muséographique vise à contextualiser un objet avant même de le décrire ou de le localiser ; c'est cette finalité qui est considérée ici comme centrale dans le modèle. Celui-ci distingue ainsi les phénomènes temporels (Temporel entity) des entités persistentes (Persistent item) comme les Choses, les Acteurs, les Appellations.

Cette démarche qui part du fondement même de l'activité et des missions du domaine, ici la muséographie, montre bien que la modélisation ne peut se réduire à l'identification et l'organisation d'un jeu de métadonnées décrivant des objets, mais qu'elle vise bien à donner une vue particulière sur un événement, un fait, une tâche, un objet, celle-ci servant de point de départ pour structurer un réseau de connaissances ce qui explique l'appellation d'ontologie au sens de « spécification rendant compte d'une conceptualisation d'un domaine⁸ »(Gruber, 1990).

Un Evènement est une entité temporelle. Elle se décline en :

- Actions (Activity) : « La Seconde Guerre mondiale, la bataille de Stalingrad, le tremblement de terre de Lisbonne, la naissance de Cléopâtre, la fête donnée pour mon anniversaire le 28 juin 1995, la conférence de Yalta, « une tuile est tombée de mon toit », la conférence CIDOC de 2005 »
- Début d'existence : naissance, création, formation,..
- Fin d'existence : destruction, dissolution, mort,..

En second lieu plus que de savoir si la démarche emprunte au modèle entité/relation (FRBR) ou au modèle objet (CRM)⁹, nous retiendrons la notation choisie pour représenter le modèle conceptuel sous une forme explicite pour l'humain (i.e. en langage naturel). Ce formalisme, malgré la complexité intrinsèque du modèle, facilite la compréhension de celui-ci par les praticiens.

une instance d'une classe *un élément physique fabriquée par l'homme* (E24 Physical Man-Made Thing), par héritage des propriétés de tout objet physique, *a une localisation ou une position* (P53: has former or current location = is former or current location), dans *un lieu précis* (E53 Place)

une instance d'une classe *Lieu* (E53) *est identifiée* (P87 – is identified by) par une *Appellation* (E44 Appellation)

L'explicitation du modèle et le formalisme adopté facilitent sa prise en main pour définir et construire des applications personnalisées (certaines propriétés ou classes peuvent ne pas être déployées), tout en garantissant l'interopérabilité sur les classes et propriétés. Les éditeurs de logiciels commencent à

⁸ Nous n'entrerons pas dans les débats sur la définition et le périmètre même de l'ontologie. Voir par exemple De l'utilité des ontologies en génie logiciel, Ivan Maffezzini, Génie Logiciel, , no 78, p. 47-57, <http://www.archipel.uqam.ca/358/>

⁹ On parle ici de classes (sous-classes), d'instances de classes, de propriétés et sous-propriétés, de hiérarchie de classes. Les avantages de ce modèle sont rapidement présentés par Patrick Le Bœuf dans « Le modèle CRM pour la documentation muséographique [...] ». Journée d'Etude de l'ADBS « La modélisation : pourquoi l'intégrer dans les systèmes d'information documentaire ? », Paris-La Défense, 20 mai 2003. http://cidoc.ics.forth.gr/docs/adbs_crm.pdf

Notons que l'ancien modèle utilisé dans l'environnement muséographique s'appuyait sur un formalisme entité/relation. De l'avis même de ces concepteurs, ce modèle était devenu ingérable.

travailler dans ce sens et proposer des logiciels dans lesquels il est possible d'effectuer un paramétrage des objets et de typer des relations¹⁰.

1.3 Rapprochement entre FRBR et CRM : un autre travail de modélisation

Un rapprochement entre les modèles des FRBR et celui du CRM s'est mis en route en 2004 ; un nouveau modèle bibliographique, le FRBRoo, est paru en 2006 puis mis à jour en 2007¹¹.

Il s'agissait au niveau du CRM d'intégrer un niveau bibliographique dans son modèle, et au niveau des FRBR à la fois d'appliquer la démarche *objet* mise en oeuvre avec le modèle CRM mais aussi d'harmoniser le modèle FRBR avec le modèle et le formalisme choisis par le CRM. Par exemple les attributs des entités ou des relations du FRBR s'expriment dans FRBRoo en « propriétés » de nature relationnelle entre les classes.

Dans ce rapprochement, un travail approfondi a été réalisé en cherchant à bien analyser les modèles sous-jacents. Après avoir fait le constat de la nature statique de l'information bibliographique dans les FRBR, l'Entité temporelle, les événements et les processus temporels propres au modèle du CRM furent pris en compte dans ce remodelage des FRBR. Parmi les re-modélisations opérées, l'entité « Œuvre », boîte noire dans le modèle FRBR, se voit complétée par une sous-classe appelée « conteneur » correspondant à une enveloppe globale permettant de modéliser plus finement les processus de création et production de l'œuvre¹². En prenant ainsi en compte l'activité de conception et de création préalable à celle de la publication, la publication n'étant alors qu'un événement particulier dans la vie de l'Œuvre, le monde bibliographique va pouvoir se rapprocher également du monde du records management et de tous les environnements pour lesquels « le cycle de vie » est un événement primordial (voir Chapitre 1- 2.1.3.).

Ce travail de mise en concordance entre modèles est un travail complexe mais nécessaire. Les résultats montrent aussi que ce travail ne se réduit pas à une mise en équivalence de métadonnées mais qu'elle consiste à reprendre et à expliciter chacun des modèles pour les rapprocher à un niveau « meta ».

1.4 Pérenniser les documents d'archives : la norme OAIS

Dans le monde du numérique, la préservation des fichiers informatiques devient un facteur incontournable de l'accès à l'information dans le temps. En effet nous avons tous fait le constat un jour, souvent au hasard de la recherche d'un document précis, de fichiers devenus illisibles. Si l'on pouvait il y a une décennie encore, supposer qu'une version papier puisse exister quelque part, il est impossible aujourd'hui de se contenter de cette réponse : la question de l'archivage à long terme des documents numériques est donc posée.

Les solutions techniques aux problèmes posés (l'obsolescence technologique des supports et formats de fichier, le document a disparu du serveur où il se trouvait, ...) restent encore très partielles et largement insatisfaisantes. Mais il existe un « Modèle de référence pour un Système ouvert d'archivage d'information » normalisé ISO (14721 :2003) qui fournit un cadre conceptuel. L'application de ce modèle à son environnement permet de comprendre les enjeux de la conservation à long terme et de relever les questions clés spécifiques à son contexte. La norme précise l'architecture logique et les fonctionnalités d'un système d'archivage et ceci quels que soient le type et la nature des données à archiver. Il identifie les acteurs, et décrit les fonctions et les flux d'information ; il propose un modèle d'information adapté à la problématique de l'archivage numérique.

Ce cadre indépendant de produits commerciaux définit 2 modèles complémentaires : un modèle fonctionnel et un modèle d'information.

¹⁰ Par exemple, la solution de MuseumPlus autorise la mise en oeuvre de relations typées à associer à des entités que l'on peut définir à partir du modèle CRM.

¹¹ FRBR, object-oriented definition and mapping to FRBRER, (version 0.9 draft), Editors: Chryssoula Bekiari, Martin Doerr, Patrick Le Boeuf, www.ifla.org/VII/s13/wgfrbr/FRBRoo_V9.1_PR.pdf

¹² Container Work, which provides a framework for conceptualising works that consist in gathering sets of signs, or fragments of sets of signs, of various origins ("aggregates"), p.12, FRBRoo cité.

Le *modèle fonctionnel* de l'OAIS emprunte au modèle d'un SGED (système de gestion électronique de document) avec un découpage en 6 grandes fonctions [Auffret, 2005] : Entrée / Stockage / Gestion de données / Administration / Planification de la pérennisation / Accès.

Cet ensemble de fonctions est représenté selon une structure par couches où chaque couche représente un traitement qui rend un service à la couche immédiatement supérieure, et sur lequel le sous-traitement suivant de la chaîne opère. A chaque couche, sont associés des sous-traitements et des catégories de métadonnées.

Le *modèle d'information* s'appuie sur des « paquets d'information », des conteneurs (voir Section 5.).

Trois catégories de conteneurs en fonction de la nature de l'information qui s'y trouve ont été définies:

- l'Objet-contenu, objet physique ou numérique, celui dont l'intelligibilité doit être préservée, associé à son Information de Représentation (information de structure, information sémantique, format,..) L'Information de Représentation permettra la compréhension de l'Objet-contenu par la Communauté d'utilisateurs cible. Un Objet-information spécifique nommé Information de description, contient les données d'accès aux documents ou aux applications d'accès.
- l'Information de Pérennisation constituée par des métadonnées précisant l'Identification, le Contexte, la Provenance ainsi que l'information de Préservation fournissant des moyens de contrôler l'intégrité des données.

Ces deux paquets d'information – Contenu et information de Pérennisation sont identifiés et encapsulés d'une troisième catégorie, l'Information d'emballage. Il existe plusieurs types de Paquet d'informations utilisés dans le processus d'archivage. Ces différents Paquets d'informations peuvent être utilisés pour structurer et stocker les fonds de l'OAIS, pour transporter l'information requise du Producteur vers l'OAIS, ou pour transporter l'information demandée par les Utilisateurs à l'OAIS.

Ce schéma avec son vocabulaire particulier est complexe, mais la question de la pérennisation des ressources numériques, de façon globale et dans le temps, est également complexe, d'autant plus si l'on prend en compte toutes les contraintes de volumes, flux, diversités des types d'objets documentaires, acteurs,...Le cadre conceptuel de l'OAIS s'étudie en parallèle de la norme ISO 15498* et du document MoReq [Ref2.] résultant de travaux récents menés par des professionnels du records management et de l'archivistique.

Mais la lecture de ces outils méthodologiques nous montre aussi qu'il est important d'agir pour que les métadonnées, que celles-ci soient descriptives, administratives ou structurelles, produites et éventuellement en cours de normalisation, dans d'autres contextes bibliographiques ou métiers soient interopérables techniquement et sémantiquement avec celles des différents paquets d'information d'un OAIS.

Conclusion de la section

Ces modèles conceptuels - FRBR, CRM et OAIS – constituent tous les trois des cadres de réflexion et de travail. Nous le disions dans le premier chapitre : ils offrent des canevas qui peuvent guider et conduire à la construction de systèmes. Mais il est nécessaire aussi, à partir de ces modèles, de définir entre ces cadres conceptuels et les applications concrètes, des schémas de référence préservant la cohérence et l'harmonisation entre cadre conceptuel et cadre applicatif. Les sections suivantes présentent ce type de schémas dans des environnements ou pour des objets différents.

2. Le monde de la référence des documents

Les périmètres des modèles ou schémas portant sur les références des documents sont variables suivant le poids accordé :

- Aux étapes du « cycle de vie » des ressources, qui démarre à la conception ou création et conduise aux questions d'archivage et de pérennisation. Nous pouvons citer le modèle OAIS (section 1.4.), la norme 15498 ou encore la norme CEI 82045¹³ sur la Gestion des documents comme outils prenant en compte les caractéristiques de cette fonction « cycle de vie », alors qu'un schéma comme le Dublin Core se focalise sur une information figée à un moment donné, pour une version ou un état donné¹⁴.
- l'orientation « référence » ou « contenu » du périmètre du schéma

2.1 Les formats centrés sur la description de l'objet

2.1.1 Description bibliographique dans le monde des bibliothèques : RDA et MODS¹⁵

Parallèlement aux évolutions du modèle bibliographique que nous venons de présenter (FRBR, FRBRoo) et en s'appuyant sur ceux-ci, le secteur des bibliothèques du monde anglo-saxon a dressé en 2002 un plan stratégique pour l'évolution des règles de description bibliographique : les RDA¹⁶ – Resources Description and Access.

Développé dès le 19^{ème} siècle et plusieurs fois remaniées, ce projet de règles de description bibliographique vise aujourd'hui à réviser en profondeur les règles utilisées jusque-là (AACR) en les adaptant au contexte actuel. Concrètement il s'agit de développer une nouvelle norme de description bibliographique et d'accès à ces descriptions de ressources, norme utilisable dans un environnement numérique et prenant en compte tous les médias. Ce travail initié en 2003 devrait se clore en 2009. Il touche à tous les éléments qui organisent et structurent les règles de description et d'accès : les types et formes de contenus, la notion de genre, les types et formes de support, la prise en compte de la structure proposée dans les modèles FRBR pour la description et FRAD (Functional Requirements for Authority Data) pour les autorités, une ouverture sur d'autres formats que le livre, ... Ce plan prend acte des principes-clés de séparation entre forme et fond et offre ainsi des règles formalisant le modèle métier (cf. Chapitre 1 Etape 2), tout en restant indépendant de tout format de présentation ou d'affichage (ISBD) ou de tout système d'encodage informatique (MARC)¹⁷.

A un niveau applicatif plus bas, il existe dans le monde bibliographique un schéma XML correspondant à un format simplifié de MARC21 : MODS¹⁸. Plus riche que le Dublin Core puisqu'il suit le format MARC, il est compatible avec les formats bibliothéconomiques et le format éditorial, ONIX (ONline Information Exchange)¹⁹. Ce dernier standard de métadonnées a été proposé en 1999 par le groupe EDItEUR pour favoriser le commerce électronique du livre et des séries à l'attention des éditeurs, libraires et autres intermédiaires. Il complète le modèle de la référence bibliographique par des données administratives comme la licence de publication, ou des données d'accès comme des listes contrôlées adaptées aux catalogues d'éditeurs.

¹³ CEI 82045 :2004 - Gestion de documents – Partie 2: Eléments de métadonnées et modèle d'information de référence

¹⁴ Dublin Core propose des « relations » qui mettent donc en relation deux versions d'un même objet. Ces relations devraient être typées, et les fonctions des auteurs, éditeurs et contributeurs agencés à ces typages pour assurer une gestion des versions.

¹⁵ Pour un panorama historique des travaux menés dans le monde bibliographique, voir EGiulianiNouveauxOutils

¹⁶ Le Site : <http://www.collectionscanada.gc.ca/jsc/rda.html>
RDA: Description des ressources et accès, Préparé par le Joint Steering Committee for Revision of AACR, 2005, Traduit par Bibliothèque et Archives Canada. http://www.collectionscanada.gc.ca/jsc/docs/rdaptjuly2005_fre.pdf

¹⁷ Vers un code international de catalogage, journée ABES des 20 et 21 mai 2008, <http://www.abes.fr/abes/page,395,journees-abes.html> [Atelier6_Fleresche.pdf]

¹⁸ <http://www.loc.gov/standards/mods/mods-overview.html>

¹⁹ Format ONIX - <http://www.editeur.org/onix.html> ; http://www.bisg.org/onix/onix_faq.html

En conclusion de cette section

Trois schémas pour trois périmètres distincts : un modèle de description de ressource indépendant (DRA), un schéma de description étendu uà des fonctions d'usages (licence) et un troisième schéma, technique d'encodage pour compléter la palette d'outils.

2.1.2 Formats de présentation réduite

Références bibliographiques ou citation : un format de présentation réduite

Dans le cadre de la démarche globale qui va d'un modèle métier à une application informatisée exposée dans le premier chapitre, la norme sur les références bibliographiques ISO 690 (la norme générale de 1987 et celle de 1997 sur les documents électroniques) fournit les « éléments à mentionner dans les références bibliographiques [...]. Elle détermine un ordre obligatoire pour les éléments de la référence et établit des règles pour la transcription et la présentation de l'information provenant de la publication source [...] pour l'établissement de listes de références bibliographiques à inclure dans une bibliographie et pour la formulation des citations dans le texte, correspondant aux entrées de la bibliographie »²⁰. Cette norme constitue bien un schéma formel au sens où nous l'avons précisé au Chapitre 1 : « Les schémas identifient les éléments constitutifs des références bibliographiques de documents et précise une séquence normalisée pour la présentation de ces éléments » ; un schéma applicatif qui n'est pas loin d'un format informatique. L'objectif des normes à l'époque était de permettre le développement des applications sans qu'il y ait trop de variations entre elles.

Les principes du document numérique structuré, à l'inverse nous convie à délaissier ce type de schéma qui pourrait rester au niveau applicatif, pour formaliser et normaliser des outils plus conceptuels structurant les applications sans les enfermer dans des formats de représentation contraignants.

Documents techniques

Dans la même catégorie de format de présentation réduite avec des contraintes supplémentaires ici de format de présentation et de dimension, citons la norme ISO 7200:2004 du TC10 pour les documents techniques de produits. Cette norme réglementaire dans un grand nombre de secteurs présente un ensemble d'éléments de données bibliographiques à inscrire dans les cartouches d'inscription et les têtes de documents techniques de produits. Cette norme s'applique à tout type de documents, pour tout type de produits, dans tous les domaines techniques et à tous les stades de leur durée de vie. Concernant les stades de vie, cette norme n'assure pas la gestion du cycle de vie du document, mais fournit toutefois en dehors des mentions de titres, d'identifiant, de propriétaire ou de type de documents, des indications sur la version de l'exemplaire ainsi que le statut et les différents contributeurs²¹.

Figure 2 – Cartouche des documents techniques

Responsible dept. ABC 2	Technical reference Patricia Johnson	Created by Jane Smith	Approved by David Brown		
Legal owner	Document type Sub-assembly drawing		Document status Released		
	Title, Supplementary title Apparatus plate Complete with brackets		AB123 456-7		
	Rev. A	Date of Issue 2002-05-14	Lang. en	Sheet 1/5	
180 mm					

²⁰ Extrait de la norme ISO 690 « 1. Objet et domaine d'application ». <http://www.collectionscanada.gc.ca/iso/tc46sc9/standard/690-1f.htm>

²¹ Future directions for IFC (Industry-Foundation class)-based interoperability, Väino, July 2003 at <http://www.itcon.org/2003/17>, FIG.7 : Schematic of dimensions in Unified Project Management

RFC 1807 - A Format for Bibliographic Records

Nous pouvons également citer dans cette catégorie, la recommandation RFC 1807 datant de 1994 et qui porte sur un format bibliographique. Il est difficile d'évaluer la portée de ce format, mais il paraît plus intéressant d'étudier les différences avec le Dublin Core par exemple. Ce format intègre des métadonnées de gestion de la production (commanditaire, version) et des éléments de contact sur l'auteur, la finalité étant de faciliter la mise en relation des auteurs et des lecteurs. Un axe très certainement à développer dans la logique des réseaux sociaux.

2.1.3 Elargissement vers des fonctions administratives : la norme sur les thèses TEF

Autre schéma de métadonnées orientée « référence », la recommandation française sur les thèses publiée par le Groupe AFNOR CG46/CN357/GE5 (voir chapitre 1, 2.2.2. Cas des thèses)²².

Ce schéma est à la fois :

- Un modèle pour le genre Thèse tenant compte des trois « dimensions qui caractérisent toute thèse [...] : un travail universitaire validé par des pairs, une oeuvre de l'esprit soumise au droit de la propriété intellectuelle et un document administratif qui conditionne la délivrance d'un diplôme national ;
- Un format informatique d'organisation des données selon le vocabulaire METS (voir 5.), qui permet d'articuler différentes catégories de métadonnées, ici des métadonnées descriptives et des métadonnées de gestion, ces dernières regroupant des métadonnées administratives relatives au suivi de la thèse, de droits et de conservation relatives à la pérennité de l'archivage ;
- Un format informatique de structuration et d'encodage XML selon le vocabulaire Schematron (voir note 18 du chapitre 1)

TEF est un schéma informatique prêt à l'emploi²³ pour produire (par exemple avec une application STAR) selon le format lui-même ou pour convertir à partir ou vers d'autres formats²⁴.

Racine METS (**mets:mets**)

- En-tête METS (**mets:metsHdr**)
- Blocs de métadonnées descriptives (**mets:dmdSec**).
- Blocs de métadonnées de gestion (**mets:techMD** ou **mets:rightsMD**)
- Inventaire des fichiers (**mets:fileSec**)
- Arbre des entités TEF (**mets:structMap**)

Le bloc de métadonnées descriptives intègre des précisions sur la version avec une métadonnée encodée avec la balise <tef-manque>.

```
...
<tef:version>
  <tef:manque>
    <tef:ressourceID>tiers1</tef:ressourceID>
    <tef:noteVersion>Manquent toutes les images
    de cette thèse</tef:noteVersion>
  </tef:manque>
</tef:version>
```

Dans cet exemple, ce nouveau schéma ouvre la description bibliographique à des informations administratives liées à la production de la thèse.

2.2 Le cas du Dublin Core

Impossible d'écrire un paragraphe sur les métadonnées sans au moins citer le schéma Dublin Core.

²² Giloux, Marianne ; Mauger Perez, Isabelle, « Le dispositif national d'archivage et de signalement des thèses électroniques », BBF, 2007, n° 6, p. 46-49 [en ligne] <<http://bbf.enssib.fr>> Consulté le 13 juillet 2008

²³ Ou presque, mais les outils s'appuyant sur ce schéma formel se développent en parallèle d'autres formalismes.

TEF en RDF : premiers essais, Yann Nicolas (ABES), Diffusé le 4 juillet 2007, <http://www.scribd.com/doc/156199/TEF-en-RDF-premier-essai>

²⁴ Voir le Site officiel : <http://www.abes.fr/abes/documents/tef/index.html>
Le Blog collaboratif sur les métadonnées des thèses numériques françaises - <http://tefsav.canalblog.com/archives/pratiques/index.html>

Le Dublin Core est un jeu de métadonnées défini en 1995 par le NSCA (National Center for Supercomputing Applications) et l'OCLC (Online Computer Library Center). Les objectifs assignés au début du Web étaient de trouver un format minimal consensuel. Ce format se devait d'être simple en création et gestion, d'une sémantique comprise largement, d'envergure internationale, utilisable avec HTML et XML et applicable au plus grand nombre de formats, tout en respectant la contrainte d'être sous un format interprétable par des moteurs de recherche et par des humains. Le Dublin Core devait être facilement extensible.

Cette initiative est devenue une norme ISO en 2003 (ISO 15836:2003). Cette version normative est composée de 15 éléments de données²⁵ optionnels et répétables, permettant la description normalisée de ressources numériques. Une version Dublin Core plus évoluée mais non normalisée autorise l'usage de qualificatifs (qualifieurs ou métadonnées de raffinement)²⁶. Par exemple, l'élément Description peut être raffiné à l'aide des qualificatifs *tableOfContents* et *abstract*. Le schéma est complété par des vocabulaires d'encodage comme le type de ressources.

Ce travail était conçu au départ pour traiter la communication entre ressources web, en établissant un jeu minimal utilisable par tous. La forte utilisation de cette norme vient de son caractère générique, de sa simplicité de mise en œuvre et de son statut normatif pour le jeu de 15 métadonnées.

Face à cet engouement pour le format et ses extensions, le DCMI a pris en mai 2008 la décision de relancer un groupe de travail sur les registres de schémas de métadonnées pour référencer les profils d'application basés sur le Dublin Core et les vocabulaires associés qui se multiplient.

This work is primarily designed to support knowledge sharing between initiatives with an interest in metadata schema registry work, and to address communication between established registries. This work focuses on metadata schema, vocabulary, and terminology registries. http://wiki.metadataregistry.org/DCMI_Registry_Community#DCMI_Registry_Community_Projects

Malgré la liste de contraintes, le résultat constitue une base (core) qui se veut commune à tout projet de métadonnées.

Avec le Dublin Core nous fermons la parenthèse des schémas de références pour entrer dans le monde des objets eux-mêmes.

²⁵ Pour rappel les 15 éléments de données du Dublin Core de base : type (catégorie de ressources), title, creator, publisher, contributor, date, source ; identifier ; relation, language, format, coverage, description, subject, rights,

²⁶ DCMI - <http://dublincore.org/documents/dcmi-terms/>

3. Le monde des documents numériques

25 ans après un des premiers projets de grande envergure en France de gestion électronique de documents, Transdoc²⁷, l'appropriation de l'information numérique s'effectue progressivement mais de façon assez disparate suivant les milieux professionnels. La vie en société a pris le relais de l'environnement du travail pour assurer un certain niveau de pratique autour de l'ordinateur et de l'information numérique : le changement de carte grise, de domiciliation ou le paiement des impôts promus par l'Administration électronique, constituent des activités fréquentes qui initient les citoyens au monde de l'information numérique. Ne parlons pas des plus jeunes (les « digital natives ») qui arrivent en masse dans les organisations avec beaucoup moins de craintes que leurs aînés.

Il semble donc important d'investir rapidement le champ du document et de l'information numérique, et de traduire, concrètement dans les modèles métiers, ses spécificités fonctionnelles et techniques.

En effet, que répondre aux Utilisateurs qui exploitent déjà des outils d'annotation ou de coproduction numérique associés à une lecture sur écran, et qui souhaitent pouvoir faire de même avec les documents qu'ils obtiennent par notre entremise ? Cette question n'est pas qu'un problème de droit et les formats ou les fonctions proposés au cœur de nos dispositifs limitent quand ils n'interdisent pas une exploitation personnalisée des contenus voir une ré-exploitation des contenus faute d'une réelle prise en compte des caractéristiques des objets et des besoins (voir Bruno ? Anne ? Olivier ?).

Nous proposons ici quelques schémas pour certains normalisés, orientés « documents » : les livres numériques eux-mêmes, la documentation d'enquête (DDI), la norme TEI de balisage de corpus textuel et le format d'encodage de l'information archivistique (EAD), l'information de presse avec le format NewML ou encore l'information géographique.

3.1 Autour des livres numériques²⁸

3.1.1 DAISY (Digital Accessible Information System)

Pour de nombreuses personnes les livres en braille ou en audio constituent l'unique moyen d'accès à l'information. Depuis plusieurs décennies, des documents sont transcrits en braille ou enregistrés vocalement sur cassettes, sur des cédéroms numériques puis plus récemment enregistrés selon des formats audio structurés. Cependant, les livres enregistrés sont laborieux à utiliser ; la recherche d'une information précise dans le contenu du livre exige d'interminables défilements sur les supports. Le livre en braille présente également des limitations pour la recherche rapide d'information.

La structuration des données en XML pour les livres au format audionumérique constitue la solution la plus efficace. Le format XML du consortium DAISY, DTBook, offre des fonctionnalités facilitant la navigation dans la structure du livre et l'affichage de texte synchronisé avec la bande audio. Destiné à être lu par une synthèse vocale, il est basé sur les standards XHTML et SMIL (Synchronized Multimedia Integration Language) du W3C pour la synchronisation. Ce format proposé et maintenu par le consortium DAISY a été adopté par BrailleNet en 2002 et normalisé en 2002 puis révisé en 2005 par l'institution de normalisation américaine (ANSI/NISO Z39.86-2005²⁹). Aux USA, les évolutions de la norme sont prises en charge par un réseau national de bibliothèques coopérant pour l'élaboration de matériels audio et en braille, la National Library Service for the Blind and Physically Handicapped (NLS), réseau piloté à la Bibliothèque du Congrès. Microsoft a adopté le format DAISY dans sa suite Office 2007³⁰ assurant une pérennité aux investissements effectués par les éditeurs.

²⁷ TRANSDOC : projet de transmission électronique de documents, Groupement TRANSDOC, Documentaliste Sciences de l'information, Volume 21, N° 3, paru le 1 mai 1984, page(s) 119-121
Sakoun, Caroline, « Transdoc : archivage et fourniture électroniques de documents », BBF, 1985, n° 6, p. 482-495 - [en ligne] <<http://bbf.enssib.fr>> Consulté le 13 juillet 2008

²⁸ Charge cognitive du livre électronique - <http://www-apa.lip6.fr/GIS.COGNITION/livr3.html>

²⁹ DAISY, est depuis juin 2008, dans une procédure de révision sur les aspects liés à la distribution et à la définition de l'autorat - <http://www.niso.org/workrooms/daisy/Z39-86-2005.html>

³⁰ DAISY - <http://www.openxmlcommunity.org/daisy/>

Le paquet d'éléments de données descriptifs du document audio numérique, appelée métadonnées de publication³¹, contient deux parties normalisées : l'une conforme à ce qui est demandé par l'OpenEbook et donc pour partie au Dublin Core, l'autre spécifique au document audio numérique et introduite par la balise *x-metadata* qui permettent entre autres des manipulations des contenus. Les éléments de données spécifiques parce que manquantes dans le Dublin Core concernent le statut du document par rapport à des étapes de production et des informations structurelles et techniques, données essentielles aux Utilisateurs.

3.1.2 ePub Books de l'IDPF

Les formats propriétaires actuels pour les livres électroniques sont dépendants de leur plate-forme de lecture. Conscient du frein que représente cette multiplication des formats, le Forum international de l'édition électronique (IDFL) regroupant de grands acteurs industriels comme Adobe, Amazon, des éditeurs comme Harper Collins ou Hachette ou des institutions du secteur de l'infodoc comme l'OCLC ou des Universités, a développé un format ouvert connu aujourd'hui sous le sigle de ePub. Ce format se base sur le format Open eBook développé depuis 1999.

ePub est un assemblage de 3 spécifications : Open Publication Structure (OPS), Open Packaging Format (OPF) et Open Container Format (OCF). Ce vocabulaire s'appuie sur la DTBook développée dans le cadre du consortium DAISY dans sa version normalisée par le NISO, et sur XHTML1.1.³² Les métadonnées de référence du livre sont donc les mêmes.

Sur le plan de la mise en application, la solution d'Adobe, Digital Editions d'Adobe, prend en charge ce format en mode natif en sus du format PDF/A. Le Groupe Hachette USA a adopté également ce format et BooksOnBoard généralise les ebooks au format EPUB sur son catalogue en ligne. ePub semble se présenter comme le format enfin harmonisé pour les livres électroniques.

3.1.3 DocBook V5.0 (06 Feb 2008)

DocBook est un schéma pour la documentation, initié en 1992. Il est maintenu par le comité technique DocBook du consortium OASIS.

Ce schéma est composé d'un noyau standardisé avec des possibilités d'extension et de personnalisation, qui permettent certaines adaptations aux contextes d'utilisation. Des filtres ou formats de conversion se développent. Ainsi le filtre Ooo2DBK³³ pour OpenOffice permet de créer des documents DocBook au format "article" et "book". Outre les formatages, l'export gère les métadonnées descriptives (champs utilisateurs), la bibliographie, le glossaire, la préface, les annexes, la table des matières, les images insérées dans le document, les bordures de tableaux etc ... OpenOffice devient ainsi un éditeur graphique simple pour produire de la documentation au format DocBook XML, ce qui a toujours constitué la difficulté majeure avec le format d'édition Tex/LaTeX, utilisé depuis de nombreuses années pour ce type de documents. Écrit initialement en Relax NG, il est aujourd'hui proposé en W3C Schéma XML, Schematron et DTD.

Sont proposés dans ce schémas un certain nombre d'attributs communs à tout type de documentation (audience, rôle, statut de révision, niveau de sécurité, langue), d'autres spécifiques à un type. Par exemple, pour les livres : remerciement (acknowledgements), appendice (appendix), article, bibliographie (bibliography), chapitre (chapter), colophon (colophon), dédicaces (dedication), glossaire (glossary) et index, , partie (part), préface (preface), référence (reference), sous-titre (subtitle), titre (title), titre abrégé (titleabbrev), table des matières (toc), ainsi que des métabalises regroupant plusieurs éléments de données : info (db.info), info (db.titleforbidden.info)

Rien d'original à une structure éditoriale classique, mais des éléments utiles pour naviguer dans la ressource. Une référence au Dublin Core pour trois éléments de donnée simplement : *Coverage*, *Relation* et *Source*.

En conclusion à cette section

³¹ DAISY - <http://www.niso.org/workrooms/daisy/Z39-86-2005.html#PubMed>

³² Sur quelques balises. « Formats numériques : un exemple », Hadrien Gardeur, mars 2008, <http://blog.feedbooks.com/fr/?p=50>

³³ Ooo2DBK sur le Site d'Indesko - <http://www.indesko.com/telechargements/>,

Après une décennie d'hésitation, il semble que nous nous trouvions aujourd'hui avec un format normalisé pour les livres numériques ou la documentation numérique, l'un plutôt orienté vers l'édition commerciale (DTBook, ePub) nécessitant en particulier un composant pour la gestion des droits, l'autre schéma plus orienté pour la production documentaire intégrée aux outils bureautiques (DocBook). Les deux formats libres et décrits dans l'environnement XML, s'appuient certes sur les métadonnées Dublin Core. Mais celles-ci ne constituent qu'une petite partie de l'ensemble des métadonnées disponibles sur les contenus et la structuration de ces contenus, les plus importantes en nombre et en possibilités fonctionnelles pour les lecteurs.

La récupération de ces métadonnées directement à la source lors de l'acquisition des documents ou leur articulation dans le cadre d'un portail multisources, deux fonctions qui semblaient techniquement délicates deviennent nettement plus aisées. Et lorsque des règles d'encodage des éléments de métadonnées et d'indexation sont connues voire harmonisées, ces fonctions deviennent transparentes.

Les praticiens auront à se déterminer suivant deux axes :

- s'ils veulent avoir la maîtrise de l'utilisation des ressources par leurs lecteurs, ils doivent prendre en charge toutes les métadonnées utiles aux lecteurs et élargir la palette des outils mis à disposition avec ceux de lecture et d'exploitation de ces contenus
- s'ils préfèrent orienter leurs activités sur le repérage des sources et ressources et la mise à disposition d'outils d'orientation et de sélection, ils devront prêter attention à coordonner leurs propres dispositifs avec les autres dispositifs que les lecteurs utiliseront in fine pour lire et exploiter.

3.2 Informations d'enquête : DDI (Data Documentation Initiative)

Le schéma DDI est une initiative internationale portant sur la documentation d'enquêtes en sciences humaines et sociales. Initiée en 1995 et maintenue par le Data Documentation Initiative, une version beta du schéma est sortie en 1999. La dernière version 3.0 date d'avril 2008.

Ce format est organisé en 5 parties principales :

Partie 1 : Description du document DDI (environ 58 items).

Partie 2 : Description de l'enquête (environ 100 items).

Partie 3 : Description des fichiers de données (environ 26 items).

Partie 4 : Description des variables (environ 50 items).

Partie 5 : Description des autres documents relatifs à l'enquête (environ 6 items).

Ce schéma est proposé au format DTD et W3C XML Schema ainsi qu'une version allégée.

Une traduction française du dictionnaire de balises de la version 1.2-2 datée du 28 août 2002 a été réalisée au sein du Réseau Quetelet³⁴. Ce réseau associe des centres français ayant des missions nationales en matière de diffusion et d'archivage d'ensemble de données statistiques pour les sciences humaines et sociales³⁵.

Une mise en correspondance avec le Dublin Core normalisé a été réalisée par les promoteurs de ce schéma. Cette mise en correspondance s'appuie sur le Dublin Core qualifié. En particulier l'élément Coverage et Date sont affinés pour pouvoir prendre en compte les caractéristiques de l'information d'enquête. Le schéma organisant un ensemble de données de nature différente, l'élément *relation* est également exploité pour rassembler ces différents composants.

Encore une fois les métadonnées Dublin Core qui décrivent la référence de ce corpus ne constituent qu'une petite partie du schéma. Les métadonnées qui en précisent les contenus offrent aux utilisateurs dans leur recherche au sein de corpus importants, des points d'accès métiers précis.

3.3 Structure générale d'un document XML textuel : la TEI (Text Encoding Initiative)

Issue des travaux de chercheurs de Vassar College (Etats-Unis) en 1987, la TEI s'est élaborée à partir des besoins de structuration, de conceptualisation et de mise en réseau de textes dans le milieu de la

³⁴ Centre Quetelet - http://www.centre.quetelet.cnrs.fr/ddi/DDI_versionFR.htm

³⁵ Depuis cette période, il semble que la France ne soit plus représentée dans cette Alliance dans laquelle pourtant de nombreux pays européens collaborent (<http://www.ddialliance.org/org/structure.html>)

recherche. Ce système de balisage fournit un cadre général, un répertoire d'éléments de données normalisés (ISO 12083 - P4:2001), accompagné de recommandations pour conduire des travaux permettant l'échange de corpus de documents numériques. Ce modèle intègre deux niveaux d'usage par le biais de sous-jeux de métadonnées spécifiques à des genres ou des types de documents (métadonnées liées aux usages) complémentaires à un sous-jeu sur l'objet (métadonnées descriptives de l'objet). L'outillage proposé permet d'encoder le texte lui-même ainsi que les annotations faites par des contributeurs non auteurs de la ressource. Cette norme, fondée sur le langage SGML, s'appuie aujourd'hui sur XML.

C'est donc à la fois un schéma descriptif d'un objet et de son contenu et des usages sur son contenu. Il rend les activités de rédaction, publication et de mise à jour plus efficace grâce au partage des outils et d'une culture commune au sein des communautés qui développent ces outils. Il offre une visibilité puisque tous ces corpus sont plus facilement réutilisables, mis en réseau.³⁶

Deux catégories d'éléments sont identifiables au sein de la TEI :

- Le noyau correspond à des balises et des éléments communs à toutes disciplines ou domaines : la structure en division et paragraphes du texte, la description documentaire du contenu, etc... que nous nommerons "structure éditoriale et bibliographique ».
- Les balises et des éléments propres à des disciplines ou domaines - le théâtre, la poésie, la transcription d'interviews, les dictionnaires, l'histoire,...-, que nous nommons "structure métier" ou de contenu.

Autour d'un module-noyau (« core tagset ») rassemblant les éléments communs à tous les types d'information s'organisent un jeu de métadonnées de base (« base tagset ») et des modules additionnels (« additional tagset »).

La structure générale du document XMLTEI, à travers son balisage, est décomposée en deux jeux de métadonnées complémentaires :

- l'en-tête (header) qui intègre à la fois des données sur la source et des méta-métadonnées sur la transcription TEI (responsables, version, mode de disponibilité,...).
- le balisage proprement dit du document

Ce balisage peut être assez léger, mais toujours formel. Il permet alors essentiellement l'échange de références ou de corpus. Un balisage plus riche peut être exploité pour décrire de façon fine les textes, mais aussi pour encoder les annotations conceptuelles elles-mêmes faites sur ces textes.

"Avant la TEI, cette transmission [entre chercheurs] ne pouvait se faire que par la lecture et la digestion des articles et ouvrages, suivie d'une reprise des éléments du corpus selon les résultats transmis par ces articles et ouvrages. La TEI ne dispense pas de lire nos collègues, bien au contraire, mais elle nous permet, comme en sciences exactes de disposer directement et de façon normalisée des textes travaillés selon les hypothèses d'autrui"³⁷.

3.4 Information archivistique : EAD (Encoded Archival Description)

Les difficultés qui se posent dans le développement des systèmes informatisés des archives viennent pour partie de l'importance des masses, souvent très anciennes de documents d'archives. La pratique archivistique s'est donc appuyée fortement sur la description de fonds, complétée par la production d'inventaires précisant sous une forme rédactionnelle le contenu des fonds.

La première étape d'informatisation s'est attaquée naturellement à la gestion des fonds eux-mêmes. Pour les fonds d'archives plus récents une activité d'indexation à la pièce s'est mis en place en particulier pour les archives courantes ou intermédiaires, s'est développée en suivant les pratiques documentaires traditionnelles³⁸. Ces bases documentaires constituent des outils d'orientation vers un

³⁶ La Foire aux DTDs ou la TEI pour tous ?, Etienne Petitjean et Susanne Alt, Atelier ATILF Septembre 2005. Un exemple dans le monde des lexiques.

³⁷ Laurent Romary (CNRS-Loria) et Henri Hudrisier (Université de Paris8). TEI - Text encoding initiative, [sd] <http://www.culture.gouv.fr/culture/dglf/riofil/tei.htm>

³⁸ PRIAM - <http://www.culture.gouv.fr/documentation/priam3/pres.htm>

inventaire ou directement vers les archives. Mais ce cas d'indexation à la pièce n'est pas si fréquent laissant d'importants fonds difficiles à explorer. L'étape suivante des projets d'informatisation, s'est donc portée sur l'informatisation des inventaires d'archives et de manuscrits, offrant une vue plus précise des fonds eux-mêmes. C'est dans ce contexte que la DTD EAD (Encoded Archival Description Document Type Definition) s'est développée. Initiée dès 1993, une première version de ce format a été publiée en août 1998. La Bibliothèque du Congrès assure la maintenance et la distribution du schéma et de la documentation, le standard lui-même restant sous la responsabilité de l'Association des Archivistes Américains (SAA).

Ce format propose un ensemble de balises (environ 146) pour décrire contenu et structure des inventaires. Il permet la création d'inventaires numériques à structure hiérarchisée, optimisant la recherche, les affichages et les échanges, et ceci indépendamment de tout logiciel. Des liens avec des images numérisées des documents décrits peuvent être ajoutés à l'inventaire numérique. Pour faciliter le travail de rétroconversion d'inventaires existants, un élément <odd> pour Other Descriptive Data, permet l'encodage "en bloc" d'information. Les éléments de l'en-tête EAD <eadheader> peuvent correspondre aux éléments du Dublin Core en s'appuyant sur l'attribut RELATEDENCODING. Le schéma est proposé en version Relax NG et W3C XML Schema.

Exemples de quelques balises³⁹ de l'EAD:

Élément obligatoire Mention de titre et de responsabilité

```
<filedesc><titlestmt><titleproper>Répertoire méthodique du versement de la Chambre de commerce et d'industrie de Saint-Dizier</titleproper></titlestmt></filedesc>
```

```
<author>établi par [--], rédacteur territorial<lb/>sous la direction de [--], Conservateur en chef, Directeur des Archives départementales de l'A </author>
```

Description physique de l'ensemble, avec physdesc

```
<did>
<unittitle>Fonds du haras national de Pompadour </unittitle>
<physdesc> 469 cartons et registres, 96 mètres linéaires</physdesc>
</did>
```

Informations sur des documents séparés du fonds, avec <separatedmaterial>

```
<archdesc level="fonds">
```

```
<did>
```

```
<unittitle>Fonds Paul Reynaud</unittitle>
```

```
<unitid>74AP/1 à 74AP/106</unitid>
```

```
<repository>Centre historique des Archives nationales</repository>
```

```
</did>
```

```
<separatedmaterial><p>D'autres papiers provenant de Paul Reynaud sont conservés au ministère des Affaires étrangères : dossier sur l'affaire norvégienne ; notes pour le ministre et télégrammes (Balkans, Finlande, Extrême-Orient), dossiers sur l'Italie et l'intervention italienne, le blocus, les armements, et l'armistice, 1939-1942.</p></separatedmaterial>
</archdesc>
```

3.5 Information d'actualité

L'IPTC (International Press and Telecommunications Council) est un consortium regroupant plus de 50 sociétés, agences et éditeurs de presse. Comme le SMPTE dans le monde du cinéma et de la télévision avec le format MXF⁴⁰, l'IPTC a en charge le développement de normes techniques visant à optimiser les échanges d'information dans un contexte numérique. Les standards proposés fournissent des jeux de métadonnées sémantiques pour décrire les contenus, les identifier et les communiquer, initialement pour des médias textuels aujourd'hui étendus à tous les médias. Ces standards sont très utilisés dans ce secteur professionnel⁴¹.

³⁹ Extrait de : Faire un répertoire ou un inventaire simple avec l'EAD, Groupe de travail AFNOR/CG46/CN357/GE3, juin 2005 - <http://www.archivesdefrance.culture.gouv.fr/static/1068>

⁴⁰ SMPTE – voir aussi en fin de chapitre

⁴¹ Pour plus d'information sur l'utilisation de ce format : site *Controlled vocabulary* animé par David Riecks : http://www.controlledvocabulary.com/imagedatabases/iptc_naa.html

Dès 1979, l'IPTC diffuse un premier format d'échange, l'IPTC 7901.

En 1991, un format pour les métadonnées de ressources numériques, l'IPTC IIM (Information Interchange Model) apparaît. La version 4 date de 1999.

En 2001, un sous-ensemble du format IIM, l'IPTC core Schema est défini pour être intégré au format XMP développé sous l'auspice d'Adobe (section 5.2.)

Un format comme l'IPTC intègre de façon normative, différents vocabulaires contrôlés⁴²

Poussé par des contraintes de rapidité et de fiabilité de la transmission de l'information, le secteur de l'information d'actualité est très actif dans le domaine de schémas standardisés de documents intégrant leurs métadonnées.

NewsML de l'IPTC est un format de type conteneur qui permet d'intégrer les métadonnées et les contenus eux-même, y compris photo, texte, vidéo ou graphique. IPTC Photo est dédié aux métadonnées pour la photo.

SportsML est un schéma adapté aux résultats, programmes et statistiques sportifs.

D'autres langages spécialisés sont en cours de définition également sous l'égide de l'IPTC comme ProgramGuideML (<http://www.programguideml.org/pages/index.php>) pour l'échange des programmes de télévision et radio.

Le site constitue un portail de référence regroupant toutes les informations sur ces référentiels.

3.6 En conclusion

Nous pourrions poursuivre ce panorama des formats de documents numériques à l'infini, en fonction du secteur ou plutôt du type d'information concernée : l'information géographique, l'information biomédicale (Bruno ?), l'information touristique, l'information audiovisuelle (Bruno) ...

Par exemple la norme pour l'information géographique (ISO 19115 : 2003) permet d'encoder tout à la fois des métadonnées descriptives globales sur la ressource et des éléments plus précis sur l'identification, la superficie, la qualité, le schéma spatial et temporel, la référence spatiale et la distribution, de données géographiques numériques. Ce format et les schémas techniques associés ont permis un formidable développement des applications de géolocalisation. Les données de nature géographique (mesures, résultats d'expertise, cartes, bases de données, photos aériennes...) se comptent en millions. Cet ensemble de normes deviennent ainsi immédiatement accessibles, lisibles sur une carte et exploitables. Par exemple, une commune peut rapidement diffuser aux citoyens des informations géolocalisées (qualité de l'eau, site de déchetteries...).

En France, l'IGN a opté pour un service de géocatalogue proposé aux collectivités.

D'autres schémas que nous n'avons pas évoqués jusque-là concernent les données techniques. Ce sont des schémas intégrés aux outils de production de documents ou qui correspondent aux informations de production et de capture numérique.

Le format Exif (EXchangeable Image File), qui n'est plus maintenu aujourd'hui, est une spécification de format de fichier pour les images utilisées avec les appareils photographiques numériques. Il fournit des métadonnées sur l'auteur dans le même esprit que la zone Propriété des logiciels bureautiques. Mais plus important : des métadonnées comme la date, l'heure ou la position sont ainsi capturées à la source. Bien sûr, comme toute métadonnée, il est nécessaire de bien contrôler à quoi réfèrent les données relevées : ainsi il ne faut pas confondre la position d'où est prise la photo et la position de l'objet sur la photo⁴³. Mais les deux métadonnées sont utiles et peuvent être articulées aux métadonnées de la norme géographique.

Pour tous les types de médias - le son, le texte, la vidéo ou le cinéma numériques, il existe des formats qui produisent des métadonnées techniques.

⁴² Vocabulaires contrôlés du schéma de l'IPTC - <http://www.iptc.org/NewsCodes/>

⁴³ Les limites de la prise en compte automatique de ces données : « Géolocalisation des images numériques fixes », Patrick Peccatte, 19 mars 2008, <http://www.scribd.com/doc/2320095/Geolocalisation-des-images-numeriques> <http://www.scribd.com/doc/4014582/Geolocalisation>

Le développement du numérique en faisant circuler les médias numériques entraînent avec eux des lots de métadonnées en amont liées à la gestion et au contenu et en aval à son usage (annotation), qu'il devient nécessaire de prendre en compte et de gérer, particulièrement mais non exclusivement, à l'intention de deux catégories de publics : des utilisateurs spécialisés souhaitant agir sur ces données, et les administrateurs pour la conservation à long terme.

Avec ce rapide panorama de schémas, référentiels ou modèles centrés sur les contenus et non sur la référence bibliographique, nous avons simplement voulu montrer l'intérêt et la nécessité de redonner au contenu, à l'essence, bref au document primaire, une place de choix dans les systèmes d'information documentaire.

4. Systèmes de représentation de concepts et de dictionnaires

L'utilisation de langages documentaires, nomenclatures ou vocabulaires contrôlés s'étend très largement au delà des frontières du secteur de l'infodoc. Un langage documentaire est un « *code sémantique de représentation des sujets permettant à un système documentaire de repérer les documents par une formulation rigoureuse de leur contenu et aux utilisateurs d'ajuster leurs interrogations à ces formulations* »⁴⁴. Dans ce contexte où l'accès aux ressources elles-mêmes semble plus immédiat et grandement facilité par les moteurs, que deviennent ces outils de normalisation des vocabulaires comme les thésaurus ou les classifications ? Comment eux-mêmes sont-ils représentés ? quels sont les modèles sous-jacents ?

Parler de représentation de concepts et de termes nous conduit à étudier plus particulièrement dans le cadre de ce chapitre, trois schémas à caractère normatif.

La *norme TMF* (ISO 16642 :2003 Terminological Markup Framework) propose un cadre général pour la représentation des bases de données terminologiques multilingues en XML. Le modèle général proposé part du concept pour aller au(x) terme(s). Ce schéma est découpé en plusieurs niveaux : données sur les concepts communes à toutes les langues, données propres à une langue, données propres à un terme.

Le *projet de norme sur les thésaurus* (ISO NP25964) en cours d'examen s'appuie sur un modèle qui établit des relations entre concepts. Les concepts étant représentés par des termes contrôlés, un seul terme dit préférentiel représente un concept dans une langue donnée, les variantes de ces termes étant subordonnées aux termes préférentiels. Ce schéma distingue des données sur les concepts et leurs relations, et des données sur les termes en relation avec les concepts.

SKOS est un vocabulaire RDF développé au sein du W3C (voir Vatant) centré sur les concepts reliés entre eux selon les mêmes principes que le thésaurus. Mais dans ce schéma, toutes les formes de représentation des concepts rattachées directement aux concepts, sont autorisées⁴⁵. Ce schéma distingue des données sur les concepts et leurs relations.

Pourquoi avoir trois modèles de représentation de choses (ici des concepts et des termes) qui semblent au premier abord, identiques ?

Les pratiques et l'existant, aussi bien en termes de données que de logiciels ou d'applications peuvent expliquer certaines différences. Mais ce sont surtout les finalités et les contextes d'usage qui fournissent les meilleurs arguments : en terminologie, pour la traduction par exemple, il est indispensable de déployer toutes les dénominations utilisées pour désigner un même concept, alors que dans le monde de l'accès à l'information avec un thésaurus, ces dénominations doivent converger vers un terme considéré comme préférentiel à l'indexation, à l'exclusion de tous les autres formes lexicales considérés comme des équivalents pour la recherche documentaire, et à l'inverse il doit s'ouvrir à toutes les formes équivalentes à la recherche. Quant au monde du Web sémantique avec SKOS, il cherche à développer une approche résolument sémantique, ce qui caractérise le modèle SKOS au moins sur deux plans. Tout d'abord la façon de désigner un concept peut ne pas reposer nécessairement sur un signe linguistique. Même si poussé par les contraintes de reprise des données existantes le modèle SKOS a varié sur cette question pour revenir dans sa toute dernière version de Juin 2008 à des choses plus connues⁴⁶, il propose en sus du signe linguistique (lexical label), une autre catégorie de signe désignée sous le label de « notation » pour des codes ou des symboles. Ensuite l'idée est que le graphe de relations et de noeuds permet de revenir à un concept quel que soit le signe (ou la variante d'un signe linguistique) utilisé. Ainsi toutes les désignations du concept sont « raccrochées » au concept lui-même et non à un label lexical (terme). Ce point tout comme la

⁴⁴ Actualité des langages documentaires, Jacques Maniez, ADBS Éditions, 2002.

⁴⁵ Pour faciliter les échanges, l'attribut « terme préférentiel » a toutefois été conservé.

⁴⁶ Voir les différences avec la version de 2005 : <http://www.w3.org/TR/2005/WD-swbp-skos-core-spec-20051102/#symbol>

question d'un « monde ouvert »⁴⁷ versus un « monde fermé » (un répertoire de concepts qui n'a de sens qu'utilisés ensemble).

Diversité des points de vue⁴⁸ :

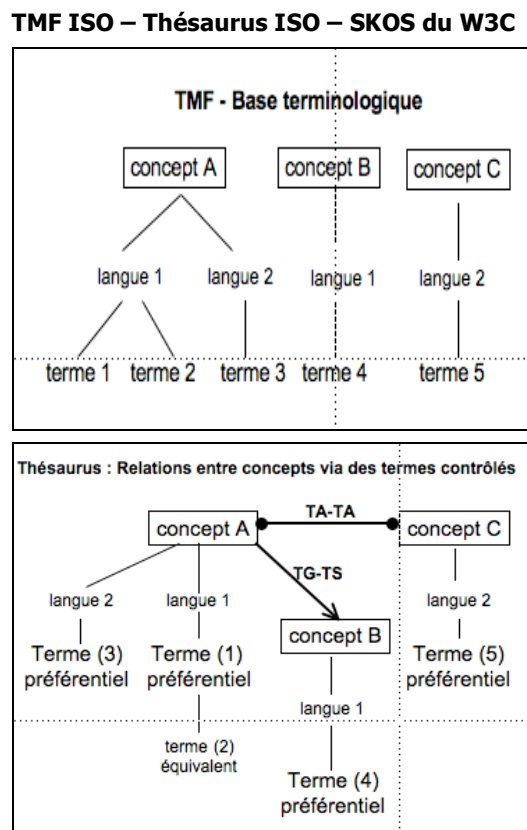
Base terminologique: termes, relations entre notion et dénomination, équivalences entre dénominations de 2 langues, procédés et dénominations linguistiques

Thésaurus: relations entre notions représentées par des termes contrôlés

Ontologie et web sémantique: notions et relations sémantiques entre les notions

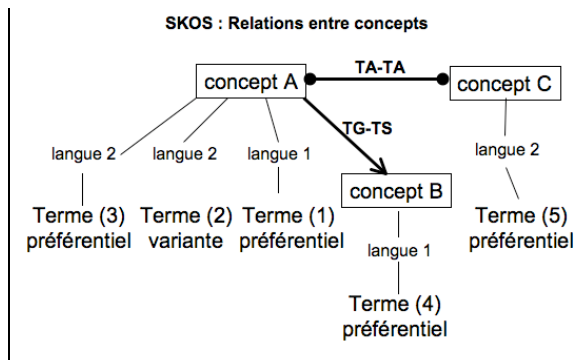
Une bonne connaissance des contextes d'usage et des modèles de ces trois schémas est nécessaire avant de constituer éventuellement un répertoire terminologique ou d'effectuer des travaux de mise en correspondance et ceci afin de préserver les spécificités des uns et des autres.

Figure 3 – Modèles de représentation de concepts et termes



⁴⁷ Notons l'approche choisie récemment pour le développement de Rameau qui, sans encore donner de réponse complète quant à la question du contexte d'utilisation d'un concept, s'oriente toutefois vers le principe de grappes ou bouquets terminologiques en rendant plus autonomes des micros-ensemble de concepts/termes. In Bilan et perspectives, Michel Mingam, Journée nationale RAMEAU, 30 mai 2008, http://rameau.bnf.fr/informations/pdf/journee2008/rameau_bilan.pdf

⁴⁸ Constitution de bases de données terminologiques sur le web, Samuel Jolibois, mars 2005 http://samuel.jolibois.free.fr/Formations/fr/ctb_web_term_db_2005-03-19_fr.ppt



A côté de ces schémas ou modèles normalisés de représentation de réservoirs de concepts se développent des applications concrètes qui ne visent pas (seulement) à montrer ces réservoirs (dictionnaire, corpus terminologiques, thésaurus, ontologie de recherche, classification), mais à les intégrer chacun dans une application de traduction automatique ou assistée par ordinateur, de recherche documentaire,.... Si les principes normatifs de chacun de ces schémas de concepts tendent à dégager des « cadres » ou des « principes directeurs », en atteignant le niveau applicatif, il est nécessaire d'encoder, de normaliser, de formater ces langages pour tenir compte d'environnement particulier.

Ceci a conduit au développement de certaines spécifications techniques comme:

- ClaML : dans le domaine très réglementé des systèmes médicaux, la spécification normalisée ClaML (EN 14463:2007 - Informatique de la santé) est une syntaxe permettant de représenter le contenu des systèmes de classification médicaux ;
- VDEX⁴⁹ est une spécification du secteur de l'éducation pour l'échange des listes de valeurs contrôlées, quelles que soient leurs formes (thesaurus, taxonomie, liste à plat). Cette spécification normalise la syntaxe et éventuellement l'ordre des entrées de différents types de vocabulaires (glossaire, thésaurus, liste à plat et une liste hiérarchisée) relève d'autres normes, les normes ISO et NISO sur les thésaurus monolingues par exemple.

Des cadres conceptuels normalisés génèrent des schémas applicatifs, eux aussi normalisés. Toute la difficulté est de pérenniser la mise en correspondance de ces différents schémas.

⁴⁹ VDEX - <http://www.imsglobal.org/vdex/>

5. Fonctions de réservoir, transport et pérennisation

La boîte à outils s'étoffe avec d'autres catégories de schémas utilitaires. Avec cette famille de schémas standardisés, il s'agit de constituer des enveloppes ou des réservoirs et de transporter efficacement références et documents. Nous revenons aux fonctions traditionnelles de gestion, stockage, transfert et conservation.

5.1 Le protocole OAI-PMH

Le Protocole de collecte de métadonnées de l'initiative Archives Ouvertes (Open Archives Initiative - Protocol for Metadata Harvesting ou OAI-PMH) est à la fois un protocole de récupération de métadonnées sur des sites fournisseurs de données et une spécification de présentation des ressources moissonnées, ou plutôt de leurs références. Cette spécification permet de constituer et d'alimenter automatiquement des réservoirs de métadonnées unifiées interrogeables par tout moteur de recherche.

Le principe a été initié en 1991 dans le secteur de la physique pour mettre à disposition de la collectivité de spécialistes, la production scientifique disciplinaire. Cette approche fonctionnelle et technique a abouti à une norme technique en 1999. Ce travail s'est accompagné de prises de position à caractère politique quant à la disponibilité publique de la production scientifique. Cet autre axe de développement du phénomène « open » faisait suite à la hausse des prix des revues mais aussi à la nécessité de diffuser plus rapidement les résultats de recherche et aux nouveaux usages du numériques. Il a conduit en 2002 à la définition d'une politique de publication, l'initiative de Budapest⁵⁰ ou BOAI. En quelques années, le paysage de la publication scientifique en a été bouleversé. En France, citons le serveur HAL⁵¹ à la fois outils de dépôt et outil d'interrogation de documents scientifiques de toutes disciplines, mis en place par le CCSD, Centre pour la communication scientifique directe au CNRS.

L'architecture de ce protocole distingue 3 entités, en dehors de la ressource (le document) rendu accessible par un lien à partir de l'entrepôt mais qui n'entre pas dans le champ de la norme OAI :

- L'item ou enregistrement côté entrepôt source ; il peut être associé à des formats normalisés de métadonnées multiples et accessibles de façon sélective aux moissonneurs
- L'ensemble des métadonnées dans un format spécifique XML transmises par le biais de moissonneurs. Ce paquet d'information est structuré en trois zones : l'en-tête avec les identifiants, les dates de gestion..., les métadonnées proprement dites de la ressource. Au minimum le Dublin Core non qualifié est exigé, mais d'autres formats sont supportés : EAD, MARC ou RSLP⁵² (un format de description des collections), eprint ou UDDI (protocole pour les annuaires de services web).
- Une zone de méta-métadonnées « about »

La force de l'environnement XML et des modèles normalisés de métadonnées s'exprime pleinement avec ce format OAI qui permet la constitution automatique de bases de références de documents préexistantes. Ce mécanisme de récupération automatique de métadonnées permet une optimisation de l'architecture du système puisque les documents primaires ne sont pas manipulés ni doublonner dans le réservoir d'interrogation, avec bien sûr la limite liée à la disponibilité de cette ressource lors de sa consultation. Les actions actuelles visent surtout à capitaliser les ressources et à mettre un réservoir de documents, les questions de recherche dans ces réservoirs se posant dans les mêmes termes que pour les formats bibliographiques.

⁵⁰ Budapest Open Access Initiative/Initiative de Budapest pour l'Accès Ouvert, <http://www.soros.org/openaccess/fr/read.shtml>

⁵¹ Serveur HAL du CNRS - <http://hal.archives-ouvertes.fr/>

⁵² Site officiel de RSLP - <http://www.ukoln.ac.uk/metadata/rslp/>

5.2 Le conteneur XMP (eXtensible Metadata Platform)

Développé par Adobe puis devenu en 2001 une norme internationale, XMP peut être vu comme un réservoir structuré de métadonnées particulièrement adapté à une ressource composite c'est-à-dire composée d'éléments pouvant avoir chacun un jeu de métadonnées spécifique. Ce format permet ainsi d'intégrer des familles différentes de métadonnées⁵³ en sus du schéma de base XMP (méta-métadonnées) : version simplifiée du Dublin Core, schéma de Gestion des droits et schéma de gestion des médias, schémas spécifiques en fonction des médias intégrés comme Exif et Exif IPTC pour des images fixes, gestion des vidéos (ou Dynamic Media) ou encore le schéma propre aux données issues de Photoshop.

Ce réservoir de métadonnées est encapsulé dans les ressources numériques PDF qui ainsi transporte une ressource composite et l'ensemble des métadonnées de ces différents composants. PDF se transforme en conteneur rendant autonome la ressource. Mais il est également possible d'exporter le fichier XMP XML contenant l'ensemble structuré des métadonnées pour être exploité dans d'autres applications. De nombreux outils en production ou en conversion sont déjà compatibles XMP⁵⁴.

5.3 Le schéma de transfert et de stockage pérenne METS

METS (Metadata Encoding and Transmission Standard)⁵⁵ est une spécification au départ développée sur la base du modèle de l'OAIS (ISO 14721:2002).

Cette spécification XML a une finalité d'encodage de métadonnées descriptive, administrative et structurelle pour le transfert (et le stockage pérenne) de ressources. C'est un format non propriétaire, ouvert, modulaire, extensible et indépendant du support (papier/numérique) ou du type de médium. Ce format prend en compte des granularités variables reliées elles. Il décrit la structure hiérarchique des objets numériques constituant la ressource (logsheets d'un enregistrement vidéo, chapitre d'une monographie, article d'un numéro d'une publication, schéma d'un rapport, ...) et permet de reconstituer la ressource dans son contexte.

Le format se structure autour de deux sections obligatoires :

- la liste des fichiers (File Section)
- la carte de structure (Structural Map) de l'objet METS

Cinq autres sections, facultatives et répétables en fonction des ressources intégrées structurent un document METS : l'en-tête, les métadonnées descriptives et administratives, les liens structurels entre les différents éléments de la carte ainsi que des exécutables. Le format repose également sur la notion de pointeurs qui permet de mettre en relation des éléments de métadonnées et des fichiers entre eux. Pour le moment, les schémas de métadonnées pris en compte par METS sont peu nombreux : Dublin Core, MARCXML/ MODS, EAD.

Maintenu par la Bibliothèque du Congrès, le format est totalement documenté sur le site qui lui est dédié : présentation, schéma XML, extensions, profils d'application enregistrés. Ce format est utilisé depuis l'origine par la NASA, le CNES ou l'ESA. Plus récemment, il a été mis en application à la Library of Congress en 2004 pour le dépôt légal audiovisuel. Il est exploité pour le transfert d'importants volumes dans le cadre par exemple de projets de numérisation entre prestataires et commaditaires.

Dans le même registre de schéma type « conteneur », on peut citer dans le secteur audiovisuel, le format MXF, un ensemble de normes établies récemment par le SMPTE (Society of Motion Picture and Television Engineers). MXF intègre dans une enveloppe unique, tous les contenus (essences), leurs métadonnées ainsi que des informations administratives sur la ressource ainsi constituée.

Les trois formats présentés rapidement dans cette section, se positionnent chacun à la fois sur des fonctions générales (récupération et stockage, enveloppe, enveloppe pour transfert) et sur un

⁵³ XMP Specifications, Adobe, september 2005

⁵⁴ Des solutions existent en dehors des produits d'Adobe. Sur le site de l'IPTC (<http://www.iptc.org/photometadata/software/supportlist1.php>) ou de Soft Experience (<http://peccatte.karefil.com/Software/Logiciels.html>)

⁵⁵ Métadonnées pour les nuls, épisode 3 : METS, BlogOKat, 20 juin 2005, <http://blogokat.canalblog.com/archives/2005/06/20/589285.html>

territoire précis qui ne se recouvre pas. Enfin notons que l'utilisation d'outils liés à l'environnement XML facilite la mise en œuvre des applications⁵⁶

5.4 Et les services ?

Lorsque nous arrivons à la question des services à offrir – recherche, achat, prêt, lecture, exploration, ... - il semble qu'il n'y ait plus rien à dire !

En définitive tous les travaux menés autour des schémas de métadonnées visent à simplifier et optimiser l'accès, la diffusion et la réexploitation des ressources, grâce à une structuration et une formalisation standardisées ou normalisées de celles-ci.

Tout protocole, tout service web se structure autour de métadonnées, qu'ils utilisent pour mener à bien leur tâche. Nous pensons à des protocoles comme SRU ou OpenURL par exemple ou encore à des outils de tris ou de catégorisation. Certaines métadonnées qui leur sont utiles leur sont fournies par un humain lorsque celui-ci fait sa demande (une requête par exemple), mais elles peuvent également provenir de traitements automatiques entre services, entre applications. Ces agents sont d'une très grande souplesse et les moteurs et leurs algorithmes (voir Olivier) n'attendent qu'une chose : un peu plus de métadonnées de qualité pour effectuer correctement leurs tâches.

⁵⁶ Un logiciel comme Greenstone propose depuis 2004, un serveur OAI et peu de temps après le format METS.
www.greenstone.org

6. Composants transversaux

Une des dernières catégories brièvement traitées ici, transversale à toutes les autres, mérite notre attention même si le périmètre et la dimension de ces schémas peuvent paraître dérisoires par rapport aux autres catégories.

Nous avons déjà évoqué la question des langues et des pays dans le premier chapitre. Nous citerons ici trois sous-catégories de composants transversaux : les identifiants, les microformats du Web et les composants de gestion de droits.

6.1 Numérotation et identifiants

ISBN (livre), ISSN (revue), ISMN (musique), ISAN (audiovisuel), ISRN (rapports d'étude), ISBN (livres et publications en série), DOI (Digital Object Identifier), ISSN (périodiques), PII (identificateur d'articles)⁵⁷ ou encore ISIL (identifiants normalisés pour les bibliothèques et les organismes apparentés). La liste est longue et s'enrichit encore dans ce domaine en plein bouillonnement [Réf3.Identifiants]

Ces systèmes d'identification nous permettent à la fois d'associer à l'identique (un code et l'objet représenté auquel il réfère) et de distinguer entre deux objets ayant des caractéristiques proches (deux traductions différentes d'un même ouvrage). Cette fonction d'identification à base d'une numérotation est tout à fait adaptée au monde numérique qui s'appuie grandement sur des codes numériques. Et le monde du Web avec les URIs (Fabien, Vatant) renforce cette pratique, même si leurs fonctions dépassent celles des identifiants internationaux dont nous parlons.

Revenons aux identifiants numériques internationaux pour la plupart pilotés au niveau du Comité technique de l'ISO, le TC46/SC9.

Des différences importantes existent entre toutes ces numérotations⁵⁸ : sur le sens de la numérotation, le fonctionnement et sa place dans les dispositifs. Mais dès la création de l'ISBN, un certain nombre d'éléments toujours en vigueur sont mis en avant :

- L'unicité et la permanence du lien entre le numéro et l'objet qu'il identifie
- La nécessité d'une organisation centralisée de la gestion des numéros attribués
- Une structure automatisable de l'identifiant, grâce à une séquence contrôlée
- Des directives sur l'inscription physique de l'identifiant sur le support identifié.

Les nouveaux systèmes de numérotation s'appuient sur ces mêmes éléments de base. Mais ils ont fait évoluer d'autres éléments comme les classes d'objets et parmi ceux-ci les objets documentaires qui se sont précisés et élargis, ainsi que les objectifs (et bien sûr les acteurs porteurs de ces projets) qui se sont largement diversifiés : circulation marchande au même titre que tout autre produit, information bibliographique, gestion des droits.

De façon très synthétique, nous voudrions insister sur quelques éléments dans l'évolution des objectifs et des numérotations elles-mêmes qui impactent les applications :

- L'élargissement des types d'objets identifiés et complémentairement la précision apportée à l'élément identifié : entre une œuvre et une instance ou une version, une matérialisation de cette œuvre. L'un est pérenne, l'autre peut subir des transformations ou des adaptations ; les droits peuvent être différents ou attachés l'un à l'autre. Les espaces de nommage dans le monde du Web vont plus loin, puisqu'il s'agit d'identifier formellement toute ressource y compris une instance d'un schéma d'encodage par exemple (un terme d'un thésaurus). Ces principes d'identification et de numérotation imposent que la formalisation et la structuration du niveau de granularité (granularité et relations) soient préalablement établis (Voir Bruno).
- Concrètement, un même objet documentaire peut avoir plusieurs identifiants en fonction de sa composition (chaque partie d'un document composite peut avoir son propre identifiant, le tout

⁵⁷ Autres numérotations internationales. BNF. 13/9/2005, <http://www.bnf.fr/pages/infopro/numeros/is-autres.htm>

⁵⁸ Cette partie emprunte à l'excellent article de Guilliani Elizabeth. L'essor de la numérotation internationale, BnF/AFNOR 20 novembre 2003

pouvant en avoir également un) ou de la nature de l'identifiant. Ainsi un ouvrage ou une vidéo peuvent avoir simultanément un numéro ISAN (contenu) et un EAN (contenant).

Ceci peut nous amener à définir dans les schémas de référence eux-mêmes, les profils d'application ou les applications informatiques elles-mêmes, une métadonnée composite (système de numérotation – numéro et éventuellement source ou contact), répétable pour tenir compte de ces différentes éventualités.

6.2 Microformats

Conçus d'abord pour les humains puis pour les machines, les microformats sont des jeux de formats d'encodage de données, ouverts et simples, conçus pour fonctionner avec d'autres standards. Les microformats tendent à résoudre des problèmes simples, de périmètre relativement autonome. Ces microformats proposent pour les objets qu'ils représentent (événement, contact...) des schémas standardisés facilitant l'interopérabilité entre dispositifs suivant ces formats.

Les microformats les plus usités sont :

- hCard implantation du format préexistant vCard (RFC 2426 - <http://www.ietf.org/rfc/rfc2426.txt>) intégré dans les logiciels bureautiques.
- hCalendar met en œuvre le format préexistant iCalendar (RFC 2445 - <http://www.ietf.org/rfc/rfc2445.txt>) pour les événements et les calendriers, utilisé largement dans HTML ou XHTML, ainsi qu'au sein des formats Atom ou RSS.
- hReview⁵⁹ pour des critiques ou des notes. Ce format intègre hCard pour l'auteur et hCalendar pour l'événement ou le fait, qui peut être une Personne, une Activité, un Evènement...
- rel="license" est un microformat déjà embarqué dans HTML, XHTML, Atom et RSS
- rel="tag" utilisé par exemple par Technorati

Exemples de microformat

```
<a href="http://creativecommons.org/licenses/by/2.0/" rel="license">cc by 2.0</a>
```

```
<a href="http://www.apache.org/licenses/LICENSE-2.0" rel="license">Apache 2.0</a>
```

```
<a href="http://technorati.com/tag/tech" rel="tag">tech</a>
```

Ces microformats constituent ainsi des composants autonomes que l'on peut facilement articuler au sein d'autres schémas d'envergure plus large. Ainsi le profil d'application LOM-Fr s'appuie-t-il sur vCard pour les éléments de données de cette nature (personne, collectivité).

6.3 Droits et gestion des droits

Le microformat « licence » évoqué précédemment est dans un mode déclaratif simple : il correspond à la mention d'un type de licence. D'autres jeux de métadonnées orientés droits et gestion de droits de structure plus complexe se sont développés ces dernières années. Citons deux formats :

- XrML (eXtensible Rights Markup Language) est un schéma XML développé par ContentGuard, une joint venture entre Xerox et Microsoft. C'est le format adopté pour MPEG comme infrastructure de référence pour son propre système DRM et par OpenBook. Il intègre le format de métadonnées Dublin Core et la numérotation DOI.
- ODRL (Open Data Rights Language) diffusé sous licence libre, a été adopté par plusieurs fournisseurs de solutions serveur ; elle est recommandée par le W3C.

Figure 4 – XMP Right Management Schema

⁵⁹ Un exemple de critique : <http://www.xfront.com/microformats/examples/hReview/example01/review-Introduction-to-mathematics.html>

Property	Value Type	Category	Description
xmpRights:Certificate	URL	External	Online rights management certificate.
xmpRights:Marked	Boolean	External	Indicates that this is a rights-managed resource.
xmpRights:Owner	bag ProperName	External	An unordered array specifying the legal owner(s) of a resource.
xmpRights:UsageTerms	Lang Alt	External	Text instructions on how a resource can be legally used.
xmpRights:WebStatement	URL	External	The location of a web page describing the owner and/or rights statement for this resource.

Les systèmes documentaires construits sur une logique de description bibliographique n'intègrent bien souvent qu'une mention déclarative du droit d'auteur. La mention des propriétaires et des droits associés n'est intégrée que dans les schémas de métadonnées les plus récents et surtout dans les secteurs professionnels où cette question est centrale (commerce, industrie du livre, audiovisuel...). Cette catégorie de métadonnées des droits peut donc faire l'objet d'une adaptation à étudier avec les spécialistes qui pourront déterminer le niveau de précision qu'il revient de traiter. en fonction de vos fonds et des publics.

Contrairement aux schémas de métadonnées qui se comprennent comme un tout représentant un objet, nous avons évoqué ici des métadonnées de périmètre plus réduit et qui offrent une grande souplesse dans le cas où les travaux applicatifs conduisent à l'intégration de plusieurs schémas.

7. Une famille de schémas : exemple du secteur de l'éducation

Nous venons de balayer un panorama d'outils méthodologiques et techniques, standardisés ou normalisés, liés aux ressources numériques. Cette présentation s'est appuyée sur des catégories de nature documentaire, c'est-à-dire des catégories structurées par des fonctions liées aux documents : gestion, production, conservation et transfert documentaires. Ce panorama qui visait à fournir un échantillon représentatif des formats existants et de leurs spécificités peut donner une impression hétéroclite contraire aux objectifs visés d'interopérabilité et d'efficacité.

L'étude d'un secteur particulier et des différents modèles, formats conceptuels et formats techniques qui s'y déploient nous apporte une autre vision, celle d'une cohérence métier. Cette cohérence métier est perceptible à partir des activités et des processus métiers qui s'y déroulent.

Nous avons choisi le secteur de l'éducation et de la formation pour montrer cette cohérence entre différents outils de représentation de différentes entités métiers.

Le secteur de l'éducation et de la formation est largement porté par une structure internationale, le consortium IMS - Global Learning Consortium, à but non lucratif et porté par plus de 120 organisations ou entreprises.

Le consortium IMS GLC participe à l'élaboration et à la promotion de spécifications ouvertes pour favoriser les activités d'éducation, de formation et d'apprentissage en ligne. La Figure 5 extraite d'un document-cadre de l'IMS, auquel nous avons ajouté l'ensemble des modèles et formats développés en 10 ans y compris ceux développés au sein d'autres structures que l'IMS, donne une bonne vision de cet ensemble de travaux.

La démarche au sein de cette communauté est progressive et itérative : les expérimentations autour d'une spécification, enrichissent la suite des travaux ou la maintenance des schémas eux-mêmes. Ce travail de standardisation a pour objectifs d'assurer tout à la fois : le dialogue entre les systèmes en particulier les plateformes d'e-learning, l'interopérabilité des contenus et des données dont celles liées aux apprenants et à leurs parcours ou aux tuteurs, la ré-utilisation de contenu avec d'autres outils ou dans d'autres cours, le management pas tant du système technique que des contenus, des acteurs et de leurs activités, l'accessibilité et la pérennité des systèmes. Ces spécifications sont interopérables entre elles, y compris avec celles qui n'émanent pas directement de l'IMS comme la spécification SCORM développée par le consortium ADL (Advanced Distributed Learning). Elles sont fournies accompagnées à la fois d'un modèle ou a minima d'un contexte explicatif, de schémas techniques et de la documentation associée.

Quelles spécifications pour quoi faire ?

Les programmes sont consultables en ligne (Course Description Metadata⁶⁰).

Les cours produits sont empaquetés et enrichis de métadonnées pour pouvoir être embarqués et utilisés sur des plateformes de formation en ligne (SCORM et une partie du standard LOM de métadonnées pour les ressources pédagogiques) ; ces ressources sont personnalisables par les apprenants (IMS Learner Information Package, IMS AccessForAll Meta-data) ; des exercices en lignes sont proposés (IMS Question & Test interoperability⁶¹ model).

Le suivi des activités et des progrès des apprenants est relevé et le dispositif offre des possibilités de dialogues entre systèmes et en particulier ouvert sur d'autres systèmes (IMS LIP, IMS Entreprise).

La scénarisation des cours (implantée dans SCORM dès 2004, IMS Learning Design, IMS Content packaging, IMS simple sequencing) améliore la qualité de l'offre pédagogique. Les ressources pédagogiques décrites précisément sont mises à disposition des enseignants comme ressources

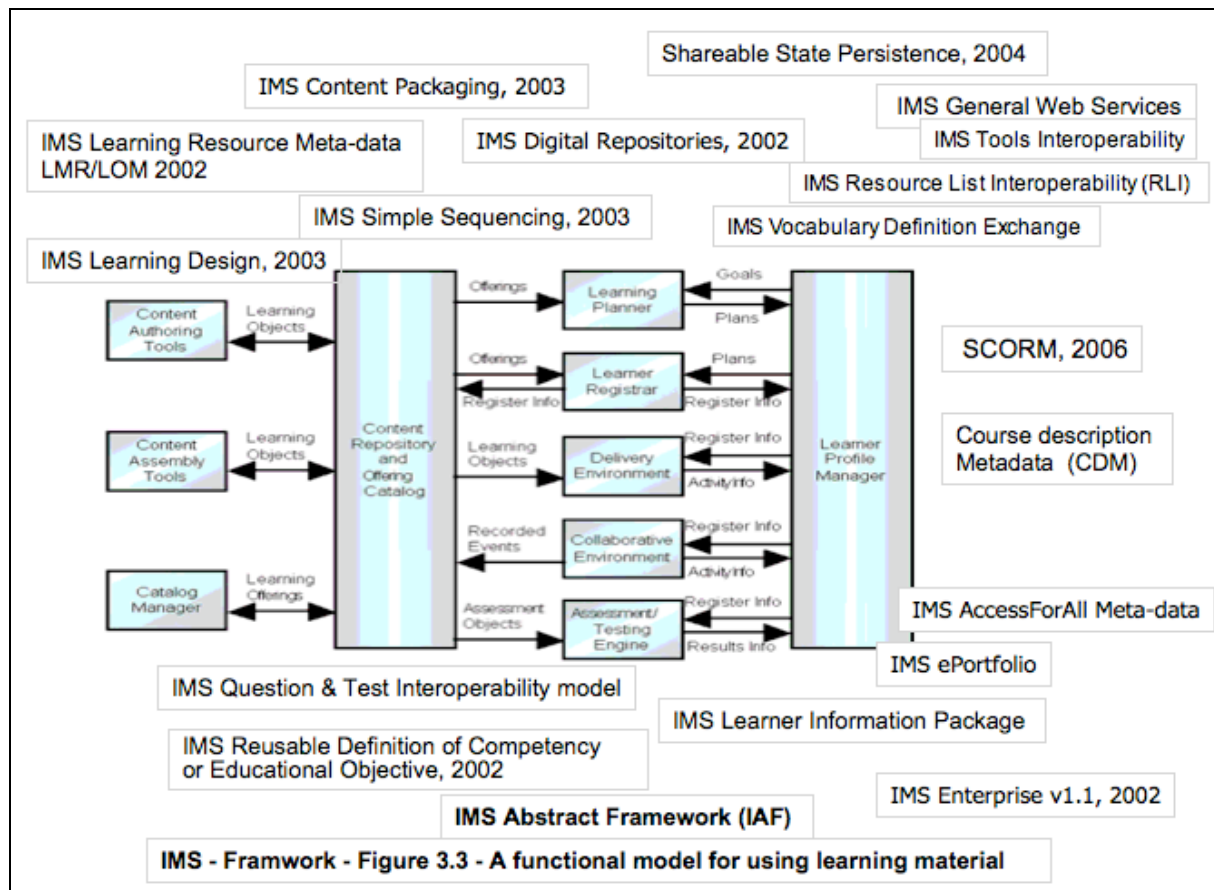
⁶⁰ Développé en 2001 par l'University d'Oslo en Norvège (<http://cdm.utdanning.no/cdm>), ce format est largement étendu aujourd'hui à d'autres pays dont la France (<http://cdm-fr.fr/>).

⁶¹ Cette spécification propose un modèle pour la représentation de questions et réponses possibles et un modèle pour représenter les réponses des utilisateurs

réexploitables pour la conception de cours ou aux apprenants (LOM, LOMFR⁶² ; IMS Digital Repositories, Interoperability et extension SCORM).

La définition des compétences (IMS Reusable Definition of Competency) et le e-portfolio (IMS ePortfolio) participent de la qualité des actions de formation et répondent aux objectifs de formation tout au long de la vie. Ils sont utilisables par les apprenants et les concepteurs.

Figure 5 – Schémas et spécification dans le secteur de l'éducation



Bien sûr nous ne réduisons pas la pédagogie ou le tutorat à distance à un ensemble de spécifications. Mais face aux contraintes et aux coûts de mise en place de programmes de formation à distance, ces spécifications interopérables entre elles garantissent que les investissements réalisés, y compris ceux des apprenants soient accompagnés et optimisés. De plus l'articulation entre ces différentes spécifications permet de mettre en place, de façon progressive ou partielle, un projet d'e-learning et de mettre en pratique plus aisément des approches individualisées.

Dans cet ensemble guidé par un cadre conceptuel (IMS Abstract Framework), l'interopérabilité des différents composants est optimisée, mais les composants restent toutefois utilisables individuellement en fonction de ses besoins. Dans ce contexte, l'intégration de ressources indexées selon le schéma du LOMFR dans un système documentaire doit préserver la cohérence d'ensemble et garantir la richesse de cette description afin d'être reexploitable en particulier dans un contexte de e-learning.

⁶² Voir par exemple le portail Pr@tic ENS, pour la production des ressources pédagogiques à l'ENS Lyon. Ces ressources sont indexées au format LOM / LOMFR et sont partagées sur le réseau internet. <http://pratic.ens-lyon.fr/>

8. En conclusion

Cette exploration du monde des schémas de métadonnées pour les ressources d'information, exploration structurée par grandes catégories fonctionnelles, nous a conduit sur un terrain où se mélaient aussi bien des aspects techniques que des aspects que nous qualifierions de « métiers ». Enfin la question de la normalisation ou plus largement de la « référence » se pose en toile de fond.

8.1 Sur le plan « technique »

Formalisme des normes

Les normes ou les standards que nous avons étudiés intègrent des formalismes assez variés complexifiant la lecture de ces documents. Or le développement d'application suppose une appropriation par de nombreux professionnels des contenus de ces documents qui devraient être plus largement diffusés. Bien sûr, une documentation pédagogique pourrait être produite en complément des documents normatifs mais en raison du nombre de ces documents normatifs et surtout en raison du rythme des changements que ceux-ci sont amenés à subir, il nous semble plus judicieux que les praticiens s'habituent à repartir des documents source.

Nous avons identifié des difficultés qui conjointes peuvent compromettre le déploiement de ces outils auprès des praticiens. Hormis le choix de la méthode qui a servi à la modélisation et qui ne semble pas négociable, d'autres règles de production des normes pourraient être améliorées :

- Une représentation graphique des modèles ou des schémas. Celle-ci peut perturber certains lecteurs mais une description sous une forme exclusivement rédactionnelle ou sous forme de liste non organisées de métadonnées, en perturbe d'autres. Une présentation des normes mixant ces deux formes conviendrait au plus grand nombre
- Parmi ces types de représentations visuelles, certaines sont porteuses de sens. Ainsi les éléments graphiques et leurs relations dans la notation UML a une signification qu'il faut apprendre à connaître. Sur ce point et en parallèle de l'apprentissage des rudiments des outils de modélisation (entité/relation, modèle objet), il serait nécessaire d'étoffer la culture générale de certaines catégories de professionnels de l'infodoc sur ces méthodes de modélisation.
- Une normalisation du nommage des éléments de données ainsi que des règles de description moins variées, permettrait de s'approprier plus rapidement schémas ou modèles. Les règles imposées pour les métadonnées pourraient être appliquées à la documentation associée, et des normes en la matière seraient la bienvenue.

Schéma de métadonnées ou composants

Un jeu de métadonnées peut être étudié dans sa globalité comme un schéma offrant une vue d'ensemble, ou comme une composition d'éléments de données - identifiant, droit, langue, pays, personnes et géographie. Cela correspond à deux hypothèses et deux approches de l'architecture des systèmes qui se dessinent dans le monde du Web⁶³ :

- L'hypothèse - qui est plutôt celle du Web sémantique - vise une intégration de schémas via des standards architecturaux qui en assurent l'interopérabilité. Cette intégration est difficile, mais elle produit une « fusion » fonctionnelle et opérationnelle même imparfaite des métadonnées.
- L'hypothèse dite de « métadonnées pidgin » est basée sur des vocabulaires de granularité faible, généraux et utilisés transversalement aux secteurs (cas des microformats).

La solution n'est peut-être pas si binaire, et l'utilisation de microformats (droit, personne, ...) à l'intérieur de schéma pourrait être valorisée.

Correspondance plutôt qu'alignement

L'activité de mise en correspondance de jeux de métadonnées doit aujourd'hui s'effectuer avec l'idée de réservoirs contenant de très nombreuses métadonnées destinées à des fins multiples (voir Figure

⁶³ Introduction to metadata, Thomas Baker, Delos Summer School, 2002 - [www.is.informatik.uni-
duisburg.de/courses/dl_ss04/folien/05-metadata-baker.pdf](http://www.is.informatik.uni-duisburg.de/courses/dl_ss04/folien/05-metadata-baker.pdf)

6). Il ne s'agit pas d'aligner strictement une métadonnée d'un jeu à une autre métadonnée d'un autre jeu, mais de rechercher la meilleure correspondance en usant éventuellement de procéder indirects comme la concaténation ou des scripts conditionnels, avec comme objectif de ne pas perdre d'informations⁶⁴ qui pourrait être utile ultérieurement. Ce problème est particulièrement frappant dans l'approche de mise en concordance avec le Dublin Core.

8.2 Sur le plan « métiers »

Nous avons vu au cours du premier chapitre que tout jeu de métadonnées prend appui sur un modèle métier sous-jacent, explicite ou implicite. L'étude d'un certain nombre de ces schémas ou cadres conceptuels nous permet d'identifier des catégories de standards ou normes en fonction du poids donné à telle ou telle phase du cycle de vie d'une ressource ou telle fonction documentaire.

La profession de l'infodoc peut encore rester centrée sur le traitement du document une fois produit et sur les métadonnées descriptives associées, en se limitant aux métadonnées bibliographiques ou en s'ouvrant aux autres métadonnées descriptives représentant les documents (le contenu technique, sa structure, des informations de gestion ou d'emballage, ...).

Mais il est possible d'intégrer des éléments liées en amont aux fonctions de production et de gestion de la production. Les avantages de cette solution sont importants puisqu'ils limitent les transactions complexes entre systèmes de production et de gestion de référence de documents, à un moment où les flux de production de l'information numérique mais aussi ceux de réexploitation et de transformation de l'information source, se démultiplient. De plus, cette configuration permettrait d'intégrer plus aisément des utilisateurs « auteurs » et des communautés utilisatrices en proximité plus ou moins étendues des auteurs.

il ne s'agit pas ici d'adopter une vision centrée sur la création de métadonnées après avoir pendant longtemps opter pour une vision centrée sur la gestion du document produit. L'enjeu est ici « autant le recyclage infini que [leur] la création initiale »⁶⁵ de métadonnées. Mais pour qu'il puisse y avoir recyclage et réexploitation, la donnée doit être présente et administrée.

Pour s'orienter vers l'une ou l'autre de ces solutions il faut tenir compte des logiques portées par XML visant à séparer contenu et présentation. Ces logiques s'appliquent bien évidemment aux systèmes prenant en charge l'ensemble des métadonnées y compris d'ailleurs celles produites par des utilisateurs. Avec ces réservoirs, seules les métadonnées pertinentes pour certaines audiences ou pour certaines transactions, sont exploitées et affichées, l'ensemble des métadonnées restant disponibles à tout moment pour de nouvelles exploitation (Cf. Figure 6).

Dans la profession, les archivistes ou les records managers sont déjà dans cette configuration, même si la prise en compte des documents numériques reste souvent calée sur des fonctions de préservation et donc d'administration, de stockage ou de sécurité. Dans ce cas, il faut rentrer dans le vif des documents numériques, de leurs formes et formats, des circuits d'édition, des caractéristiques des supports numériques.

Pour ceux plus concernés par la médiation, l'impact du document structuré numérique et des métadonnées associées porte sur différents points. Tout d'abord il faut accepter de ne pas chercher à s'appropriier les documents pour les référencer dans un système local ; il faut apprendre à appréhender les contenus de ces documents et les possibilités fonctionnelles susceptibles d'intéresser les utilisateurs. Ensuite dans cette orientation d'intermédiation et de services, les professionnels doivent identifier rapidement pour chacune des sources d'information et des filières d'acquisition, les catégories de métadonnées utiles à la construction des dispositifs et interfaces d'accès aux documents. L'enjeu est de se trouver en capacité d'articuler un réservoir de métadonnées riches pour mettre en œuvre outils (moteur, classification et visualisation graphique, ...) (Olivier) et interfaces

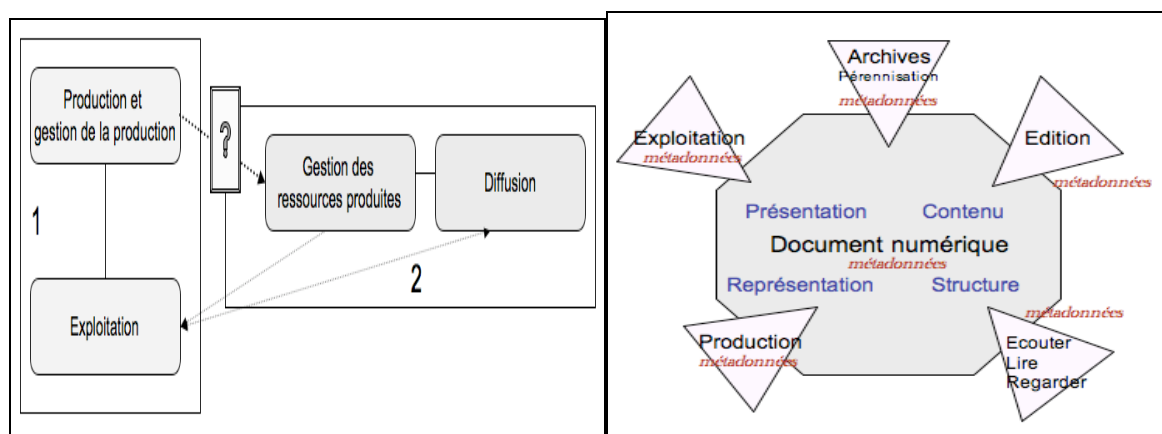
⁶⁴ Voir cet exemple de tableau de correspondance entre le Dublin Core qualifié (<http://www.pleade.org/fr/documentation/descriptions/qdc.html>) et l'EAD (<http://www.pleade.org/fr/documentation/descriptions/implicit-extract.html>)

⁶⁵ Métadonnées de droit : l'impact de la DADVSI, Yann Nicolas. 25 Février 2007. <http://tefsav.canalblog.com/archives/2007/02/25/4131295.html>

d'accès. Dans ce domaine de l'architecture d'information, il convient de ne pas se laisser aller à la facilité du Dublin Core au risque de réduire à ce jeu minimal, les clés d'accès aux ressources mais plutôt d'adopter une approche que nous qualifierons de « libre » par rapport à la diversité de métadonnées.

Bien sûr, le terrain professionnel ne se réduit pas à ces deux situations extrêmes, et les cas où la bibliothèque ou le centre documentaire assurent à la fois des missions de préservation du patrimoine et de médiation de sources externes auprès de différentes audiences, sont très nombreux. Les solutions sont fonction bien sûr des contextes locaux, mais une des orientations pourrait consister à adopter deux filières pour répondre précisément aux exigences de ces deux contextes plutôt que de trouver un juste milieu réducteur, bien souvent au détriment des utilisateurs.

Figure 6 – Réservoir de métadonnées



Dans cette deuxième configuration, le travail d'interopérabilité et de mise en correspondance de jeux de métadonnées prend une tout autre ampleur : le capital de métadonnées devient s'enrichit au fur et à mesure et réutilisable dans de nombreux contextes et applications.

8.3 Quel terrain pour la normalisation ?

En partant du périmètre informationnel élargi à la fois sur les objets, les utilisateurs et les fonctions tel que nous venons de l'évoquer, le terrain de la normalisation prend ainsi une tout autre dimension.

Les structures normatives, dont la production s'impose aux professionnels sont multiples :

- En tout premier lieu, le Comité technique TC46 – Information et documentation⁶⁶ et le miroir français la Commission Générale CG 46 : sont produites ici toutes les normes ISO des Musées, des Archives ou du Records management et des Bibliothèques, les identifiants numériques internationaux, la norme sur les thésaurus.
- Le monde de l'Internet et du Web avec l'IETF (Internet Engineering Task Force)⁶⁷ pour les normes techniques de l'Internet ou le W3C (Consortium du WWW).
- Enfin par métiers ou collectifs de métiers, de nombreux regroupement d'institutions ou d'entreprises de composition variable en fonction des domaines, sont porteurs d'une activité normative qui n'aboutit pas toujours à une norme ISO mais dont la production joue le rôle de norme pour la profession. Citons l'IFLA dans le secteur bibliographique, l'IPTC consortium d'agences de presse, l'Open Geospatial Consortium, l'IMS Global Learning Consortium. Dans ce contexte d'autres structures peuvent se trouver dans une position d'acteur stratégique fort, un rouage essentiel. C'est le cas pour notre secteur du rôle particulier détenu par la Bibliothèque du

⁶⁶ « La galère du travail normatif en infodoc », Sylvie Dalbin, Descripteurs, 22 juin 2008. <http://dossierdoc.typepad.com/descripteurs/2008/06/la-galre-du-tra.html>

⁶⁷ Cet organisme produit les RFC - request for comment - qui ont valeur de standards pour l'Internet. Les RFC qui ont été normalisées ISO, continuent à se dénommer RFC.

Congrès dont la mission est de soutenir ces développements technologiques et qui se trouve à la tête de 17 référentiels, du MARC21 aux codes des langues ISO 639-2⁶⁸.

La liste est longue de ces structures à vocation normative, et c'est bien un des problèmes posés aux professionnels. Ce déluge de modèles et de schémas et la myriade de structures à caractère normatif qui les déploient et en font la promotion, rendent très difficile une simple veille de ces dispositifs et de leur production⁶⁹ et complexifient notre éventuelle participation à tous ces travaux. L'ISO, en particulier le TC46 qui reçoit de nombreuses demandes pour améliorer et transformer les standards en normes, a bien du mal avec ces moyens à répondre à toutes ces sollicitations. Les enjeux culturels, économiques⁷⁰, techniques et sociaux du capital constitué par les ressources et leurs métadonnées, restent certainement à clarifier et à valoriser auprès des praticiens afin de les convier à une participation plus active de leurs développements.

⁶⁸ Portail sur les normes supportées par la Bibliothèque du Congrès - <http://www.loc.gov/standards/>

⁶⁹ Sans compter le fait que la majorité de ces ressources sont en anglais

⁷⁰ Les enjeux économiques de l'interopérabilité. Le cas de la gestion des droits numériques, Fabrice Rochelandet, Michèle Francine Mbo'o Ida, ADIS-Aneis, Université Paris-Sud, aneis.adislab.net (version préliminaire, version corrigée publiée dans *Revue Lamy – Droit de l'Immatériel*, 01/2007)

9. Annexes

9.1 Localisation Web des jeux de métadonnées abordés dans les chapitres 1 et 4

Nom ou numéro de norme	Sources
CDM	Course Description Metadata Représentation de programmes d'enseignement http://cdm.nou.no/ Site collaboratif pour le profil d'application français CDM-Fr. http://accés.inrp.fr/cdm/ Exemples - http://formations.univ-lille1.fr/cdm/
CRM ISO 21127:2006	Conceptual Reference Model (CRM) http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=34424
DAISY ANSI/NISO Z39.86-2002	http://www.daisy.org/ http://www.loc.gov/nls/z3986/v100/
DDI	Le site : http://www.ddialliance.org/ddi3/index.html Traduction française du dictionnaire de balises de la version 1.2-2 datée du 28 août 2002 : http://www.centre.quetelet.cnrs.fr/ddi/DDI_versionFR.htm
DocBook v5.0.	http://www.oasis-open.org/docbook/ http://www.docbook.org/tdg5/en/html/docbook.html
Dublin Core ISO 15836:2003	Norme ISO: http://www.niso.org/international/SC4/sc4docs.html (site du sous-comité 4) Dublin Core Metadata Initiative (DCMI). http://dublincore.org http://www.rfc-editor.org/rfc/rfc5013.txt Modèle conceptuel DC - http://dublincore.org/documents/domain-range/index.shtml
EAD	Site officiel : http://www.loc.gov/ead/ Site de SAA EAD Roundtable's Web site - http://www.archivists.org/saagroups/ead/ Encoded Archival Description : Tag Library. Version 1.0 / Description Archivistique Encodée : Dictionnaire des balises. Society of American Archivists (traduit de l'anglais par le groupe AFNOR CG46/CN357/GE3), octobre 2002 Site des Archives de France : http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html http://www.archivesdefrance.culture.gouv.fr/gerer/classement/normes-outils/ead/
ePUB	Site officiel de l'IDPF - http://www.idpf.org/specs.htm (extension .epub)
Exif	Image fixe - http://exif.org/
FRBR	http://www.ifla.org/VII/s13/frbr/frbr.htm Functional Requirements for Bibliographic Records / Spécifications fonctionnelles des notices bibliographiques. Final Report. IFLA Study Group on the Functional Requirements for Bibliographic Records, Approved by the Standing Committee of the IFLA Section on Cataloguing, September 1997. As amended and corrected through February 2008 Site de la BNF : http://www.bnf.fr/pages/infopro/normes/no-acFRBR.htm (version française de 2001)
IEC 82045:2:2005	Document management -- Part 2: Metadata elements and information reference model ...
ISO 3166	Noms et codes de pays en Français http://www.iso.org/iso/fr/french_country_names_and_code_elements
ISO 690 :1987 ISO 690-2:1997	Références bibliographiques - Contenu, forme et structure - http://www.lac-bac.gc.ca/iso/tc46sc9/standard/690-2f.htm Références bibliographiques -- Partie 2: Documents électroniques, documents complets ou parties de documents - http://www.lac-bac.gc.ca/iso/tc46sc9/standard/690-2fx.htm
ISO 7200:2004	Document technique http://www.iso.org/iso/fr/CatalogueDetailPage.CatalogueDetail?CSNUMBER=35446
Microformat	http://microformats.org/wiki/Main_Page
NewsML	http://www.newsml.org/
OAI-PMH	Open Archives Initiative's Protocol for Metadata Harvesting (Protocole du transfert de métadonnées d'une archive ouverte) http://www.openarchives.org/OAI/openarchivesprotocol.html Version 2.0. http://www.openarchives.org/OAI/2.0/guidelines-repository.htm Un exemple : « Archive Ouverte en Sciences de l'Information et de la Communication » http://archivesic.ccsd.cnrs.fr/
OAIS ISO 14721:2003	Site de l'ISO, http://www.iso.org/iso/fr/ Modèle de référence pour un Système ouvert d'archivage d'information (OAIS) : Projet de standard CCSDS, CCSDS 650.0-B-1 (F). (version française proposée à l'ISO) http://vds.cnes.fr/pin/documents/projet_norme_oais_version_francaise.pdf

	Site du Groupe PIN. http://vds.cnes.fr/pin/pin_normes.html
OASIS	http://www.oasis-open.org/
ODRL	http://www.odrl.net/
RSS 2.0. Atom...	RSS 2.0 version originale. http://blogs.law.harvard.edu/tech/rss version française, JY Stervinou. http://www.stervinou.com/projets/rss/ The Atom Syndication Formation. December 2005. http://www.ietf.org/rfc/rfc4287.txt
SCORM	Sharable Content Object Reference Model - http://www.adlnet.gov/about/index.aspx (Modèle de référence pour les objets de contenus à partager)
ISO 19115 et 19139	ISO/TS 19139:2007 - Geographic information -- Metadata -- XML schema implementation http://www.iso.org/iso/fr/CatalogueDetailPage.CatalogueDetail?CSNUMBER=26020 IGN - http://eden.ign.fr/xsd/isotc211/index_html?set_language=fr&cl=fr
SKOS	http://www.w3.org/2004/02/skos/ La spécification. http://www.w3.org/2004/02/skos/core/spec/2005-10-06/
TEF	Les métadonnées des thèses électroniques françaises. TEF, Groupe AFNOR CG46/CN357/GE5, seconde Édition, Mars 2006 http://www.abes.fr/abes/documents/tef/recommandation/index.html
TEI	http://www.tei-c.org
TMF	http://www.loria.fr/projets/TMF/ Sur le site Termsciences. [sd]. http://www.termscienc.es.fr/article18.html
XMP	http://www.adobe.com/products/xmp/main.html

9.2 Références

Ref1. Métadonnées

Indexation de ressources : métadonnées, normes et standards, Dossier numérique, Annie Bassinet. Marie-Noëlle Cormenier. Geneviève Viala du SDTICE, mars 2007, <http://160.92.130.159/dossier/metadata/default.htm>

Portail du groupe de travail Métadonnées et interopérabilité, Catherine Morel-Pair (animatrice), http://artist.inist.fr/rubrique.php3?id_rubrique=41

Des métadonnées pour bien utiliser les ressources électroniques, Journée d'information AFNOR/CG46, Mardi 7 juin 2005, Bibliothèque nationale de France, <http://www.bnf.fr/pages/infopro/journeespro/no-Afnor2005.htm>

Le document doit gérer de nouvelles métadonnées, Jean-Pierre Blanger, Ricoh, Forum de la GEIDE. Paris, 2005. http://www.ricoh.fr/Binary/14_pdf.pdf

Ref2. Archivage et pérennisation

MoReq2 - Modèle d'exigences pour l'organisation de l'archivage électronique, EUROPA, http://ec.europa.eu/transparency/archival_policy/moreq/moreq2_fr.htm

L'archivage pérenne des documents numériques, Michel Auffret, JRES 2005, <http://2005.jres.org/paper/47.pdf>,

Réf3. Identifiants

L'essor de la numérotation internationale, Guillian Elizabeth, BnF/AFNOR 20 novembre 2003
Identifier les identifiants, Douglas Campbe, Trad. Catherine Gunet (INIST), In DCMI Conference Proceedings à Singapour en 2007, Ametist, n°2, <http://ametist.inist.fr/document.php?id=503>

Ref4. TEI – Text encoding initiative

[Poupeau] - Un dossier : A la découverte de la TEI. G. Poupeau, <http://www.lespetitescases.net/a-la-decouverte-de-la-tei>

[Poupeau 2004] - Un exemple : « Réflexions sur l'utilisation de la TEI pour coder les sources diplomatiques à partir de l'exemple du Cartulaire blanc de l'abbaye de Saint-Denis », Le Médiéviste et l'ordinateur, 43, 2004 [En ligne] <http://lemo.irht.cnrs.fr/43/43-12.htm>

Réf5. Thèses

Les métadonnées des thèses électroniques françaises. TEF. Groupe AFNOR CG46/CN357/GE5. seconde Édition. Mars 2006

A structured approach to presenting PhD theses : notes for candidates and their supervisors, Chad Perry, <http://credo.quebec.com/redaction.html>

Réf6. OAI

Une économie renouvelée de la publication scientifique, Ghislaine Chartron, Perspective documentaire en éducation, n°62, 2006. http://archivesic.ccsd.cnrs.fr/sic_00117798/en/

Dossier Les archives ouvertes, Service Commun de la Documentation - Université Paris Descartes, mise à jour Décembre 2007, http://www.bu.univ-paris5.fr/spip.php?page=imprimer&id_article=301

Réf7. Gestion des droits

Digital Rights Management (DRM) Architectures. Renato Iannella. June 2001 [en].

<http://www.dlib.org/dlib/june01/iannella/06iannella.html>

Gestion des droits numériques. http://www.supinfo-projects.com/en/2005/gestion_droits_numeriques/1/

ODRL, langage ouvert des droits numériques. R. Parent. 2002.

http://www.services.gouv.qc.ca/fr/publications/enligne/normes/securite_attestation/droits_numeriques.pdf

DRM : Digital Rights Management systems. Quand l'idéologie rencontre la réalité, Hervé Le Crosnier, [2005], <http://ecole-ouverte.ens-lsh.fr/hlc/>

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 + everyware	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>	<i>193</i>

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
------------------------------	------------