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 coeur 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

Figures

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 évlution 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 efficience 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 problématique[1].
>
> 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'Etape 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.

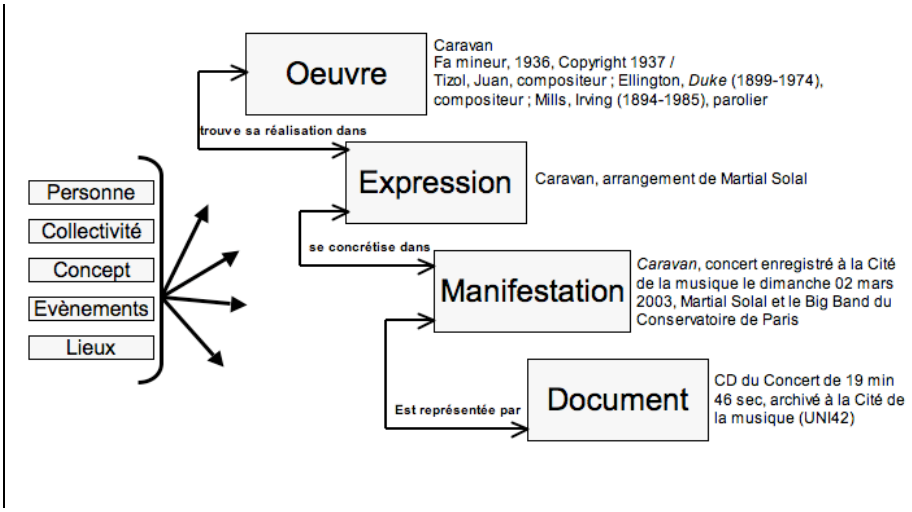
1.1 FRBR (Functional Requirements for Bibliographic Records)

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 oeuvre qui correpond 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, Evènement, Lieu.

Figure 1 - FRBR – Exemple établi à partir du catalogue de la Cité de la Musique[3]



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 Expression[4].

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 FRBR[5] 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 FRSAR (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, FRBRoo[6].

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'Ifla[7]
- 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 contrainre 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 muséogrpahie, 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 domaine[8] »(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 : desctruction, 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 relations[10].

1.3 Raprochement 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 2007[11].

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'œuvre[12]. 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'Oeuvre, le monde bibliograhique va pouvoir se rapprocher également du monde du records management et de tous les environnement 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 œuvre 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 Bœuf, 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'appuit 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 Pérénnisation sont identifiés et encapsulés d'une troisième catégorie, l'Information d'empaquetage. 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 méthologiques 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 techiquement 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 pérénnisation. Nous pouvons citer le modèle OAIS (section 1.4.), la norme 15498 ou encore la norme CEI 82045[13] 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 donné[14].
- 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 MODS [15]

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 RDA[16] – Resources Description and Access.

Développé dès le 19ième 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 : MODS[18]. Plus riche que le Dublin Core puisqu'il suite 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/rdapptjuly2005_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

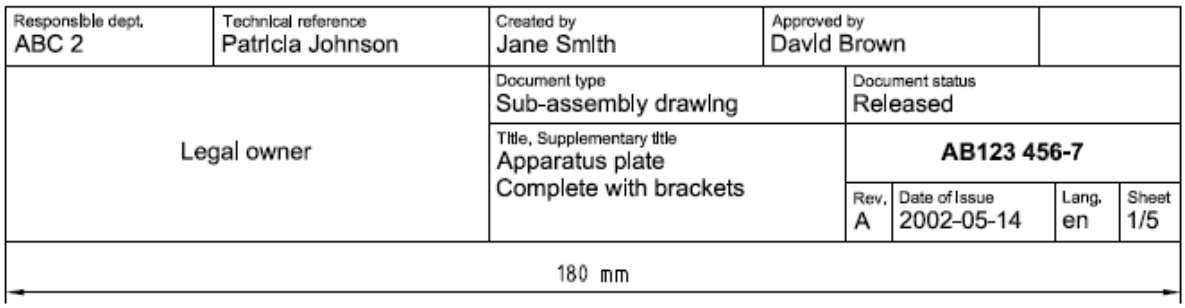
Dans le cadre de la démarche globale qui va d'un modèle métier à une application informatisée exposée dans le premier chapitre, la norme sur les références bibliographiques ISO 690 (la norme générale de 1987 et celle de 1997 sur les documents électroniques) fournit les « éléments à mentionner dans les 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 délaisser 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 fournie 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 contributeurs[21].

Figure 2 – Cartouche des documents techniques



[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 parait 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 métadonnes 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'emploi[23] pour produire (par exemple avec une application STAR) selon le format lui-même ou pour convertir à partir ou vers d'autres formats[24].

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 initative est devenue une norme ISO en 2003 (ISO 15836:2003). Cette version normative est composée de 15 éléments de données[25] optionnels et répétables, permettant la description normalisée de ressources numériques. Une version Dublin Core plus évoluée mais non normalisée autorise l'usage de qualificateurs (qualifieurs ou métadonnées de raffinement)[26]. Par exemple, l'élément Description peut être raffiné à l'aide des qualificateurs *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, Transdoc[27], 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 impots 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 numériques[28]

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 Z39.86-2005[29]). 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 2007[30] 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/