Section 9 - Métadonnées

FRANÇOISE BANAT-BERGER
CLAUDE HUC

version 1

22 novembre 2011

Table des matières

Chapitre 1. Introduction

A. 1.1 - Objet de la section et premiers éléments

Il s'agit ici de donner une définition des métadonnées, de positionner les métadonnées au sein du modèle OAIS et enfin de donner une typologie des métadonnées à travers quatre grandes familles que nous reprendrons en détail plus loin : **les métadonnées descriptives, les métadonnées techniques, les métadonnées de structure, et enfin les métadonnées administratives.**

B. 1.2 - Premiers éléments sur les métadonnées

Les métadonnées sont un vaste sujet à la mesure du rôle qu'elles jouent dans le processus de pérennisation. Essayons d'en donner une définition. Étymologiquement, « méta » provient du grec signifiant « après, au-delà de, avec » :« méta » données signifie « au-delà des données », « qui dépasse les données », « qui englobe les données ».

Il s'agit donc de données sur les données, à propos des données, qui définissent, décrivent des données. Le terme est récent. Néanmoins, il y a toujours eu des métadonnées. Selon l'activité, cela s'appelle cataloguer, indexer, classifier, décrire, élaborer un instrument de recherche, que l'on soit bibliothécaire, documentaliste, archiviste, scientifique. Il s'agit, à partir des archives collectées en provenance d'un producteur, de pouvoir communiquer ces archives bien plus largement, à de nouvelles communautés d'utilisateurs pour lesquels les métadonnées « métier » élaborées par les producteurs doivent être enrichies et explicitées pour des communautés plus larges.

Les métadonnées sont ainsi tout d'abord un outil d'aide à la recherche. Avec les données numériques est apparu en outre, le besoin d'avoir des données qui fournissent une information factuelle sur les façons d'employer et de manipuler les données (métadonnées techniques). De même les métadonnées relatives aux droits d’accès deviennent plus complexes à gérer, dans un contexte de diffusion sur les réseaux internet.

Les métadonnées sont des données, soumises aux mêmes défis de production, gestion, conservation que les données elles-mêmes...

Chapitre 2 - Le modèle OAIS et les métadonnées

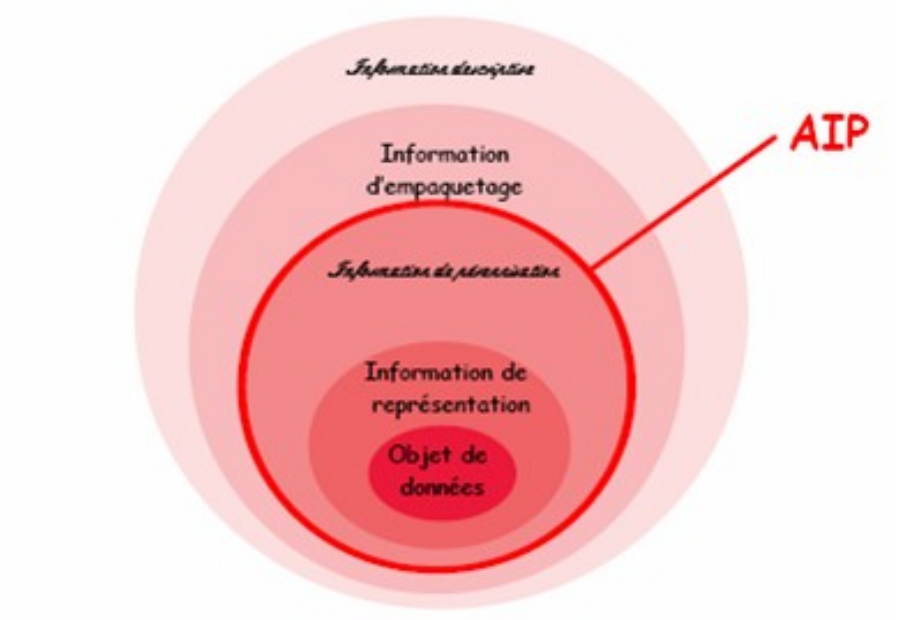
A. Le modèle OAIS et les métadonnées

Le modèle OAIS nous montre clairement que nous avons besoin de conserver une somme importante d'informations pour espérer pourvoir préserver au mieux l'information enregistrée dans les objets de données. Ces informations sont de différentes natures et ont des fonctions différentes.

En effet, un des concepts basiques de l'OAIS est le Paquet d'information archivé (AIP). Nous avons déjà vu dans la partie IV relative aux normes, que ce paquet contient :

• L'Information de pérennisation (provenance, contexte, identification, intégrité, droits),

• L'Objet information proprement dit contenant l'information à pérenniser proprement dite, sous forme de données et l'information de représentation correspondante, permettant d'interpréter les données sous la forme d'une information compréhensible



Les métadonnées dans le modèle OAIS

Légende :

Cette image représente de façon symbolique la somme d'informations complémentaires qui doivent s'ajouter à l'objet numérique proprement dit, l'objet « données », afin d'en assurer la préservation et la réutilisation future.

Chapitre 3 - Une typologie des métadonnées

Les qualificatifs pour caractériser les métadonnées sont très nombreux : descriptives, administratives, contextuelles, structurelles, techniques, de pérennisations, statique, évolutives, externes de contenus...

Certaines de ces données sur les données servent à organiser la connaissance et à utiliser et exploiter le document, d'autres servent à les maintenir accessible dans le temps et à garantir et controler leurs accès.

Nous allons tenter ici de classifier les métadonnées nécessaire à la pérennisation en nous appuyant autant que possible sur le modèle OAIS.

A. 3.1 – Le classement des métadonnées

On peut classer les différents types de métadonnées en fonction :

- de ce qu'elles décrivent (le contenu)
- de la façon dont elles sont créées (leur provenance)
- du moment où on les crée (leur historique)
- de l'endroit où on les trouve (leur localisation)
- de l'aspect qu'elles ont (leur forme)
- de l'usage qu'on en fait (leur objectif)

Naturellement, ces catégories ne sont pas disjointes.

La gestion d'un objet numérique complexe implique :

- une gestion complexe du descriptif (une ou plusieurs notices, descriptions à niveaux)
- une gestion complexe des aspects techniques (s'il y a différents formats de fichier par exemple)
- une gestion complexe des aspects administratifs (droits, suivi des évolutions du document, suivi des évolutions des métadonnées)
- une gestion complexe de sa structure physique (organisation des fichiers entre eux)
- une gestion complexe de sa structure logique (organisation des parties du contenu entre elles)

C'est le rôle des métadonnées.

Complément : ce qui permet de définir un format de métadonnées

En pratique, la définition d'un format de métadonnées inclut généralement les éléments suivants :

• un dictionnaire, qui rend compte de la sémantique du schéma de métadonnées : voir sur le site de la direction des Archives de France, le dictionnaire des balises de la DTD EAD (voir paragraphe suivant), traduit en français : http://www.archivesdefrance.culture.gouv.fr/static/1066

• un thesaurus, instrument de maîtrise du vocabulaire : quel sens précis donner à chaque terme ?,

• un schéma qui traduit la structuration des concepts applicables à l'objet numérique. Ce schéma est spécifié le plus souvent sous la forme d'une DTD ou d'un schéma XML ou encore d'une modélisation UML.

B. 3.2 – Plusieurs exemples de formats de métadonnées

Ces dernières années, nous avons assisté à la naissance d'un nombre important de définitions de métadonnées à l'initiative d'organisations internationales, d'universités, de bibliothèques nationales, chacune cherchant à répondre à une problématique générale ou particulière à leur communauté.

Dans le domaine bibliographique, nous pouvons énumérer :

Dublin Core (http://dublincore.org/) qui est un format de métadonnées très général, adapté aux ressources mises en ligne sur Internet. Ce format peut être enrichi et spécialisé en fonction des métiers et des domaines. Un exemple sera donné plus loin pour la description de données orales.

Marc21 (http://www.loc.gov/marc/marcdocz.html) est un format de métadonnées pour le catalogage, très utilisé par les bibliothèques. Il est basé sur le format MARC (MAchine-Readable Cataloging),

EAD (Encoded Archival Description) (http://www.loc.gov/ead/) est utilisé la description des documents d'archives,

TEF. Version 2.0 (Thèses électroniques françaises) (http://www.abes.fr/abes/documents/tef/) La recommandation TEF qui est produite par un groupe d'experts de l'AFNOR, définit un jeu de métadonnées pour les thèses électroniques soutenues en France. Son objectif est d'organiser de manière cohérente des métadonnées de thèse riches et normalisées, pour faciliter leur échange et leur diffusion, au niveau national ou international. Ces métadonnées sont hétérogènes. Elles recouvrent des métadonnées descriptives (bibliographiques) et des métadonnées de gestion (administration, droits, conservation).

Voir un exemple de notice TEF en HTML à l'adresse suivante : http://www.abes.fr/abes/documents/tef/recommandation/ex1_theseSimplePDF.html

Les normes ISO de la série 19000 couvrent le domaine géographique (ISO 19115 – métadonnées, ISO 19119 – services, 1SO 19101 – modèle de référence…). Elles prennent une importance grandissante dans le cadre de la directive européenne INSPIRE.

D'autres normes et standards de métadonnées couvrent d'autres domaines.

C. 3.3 – Comment choisir un format de métadonnées ?

Deux constatations :

Premièrement, la réflexion sur le sujet devient mature. Ainsi les formats sont normalisés ou bien certains formats deviennent des standards de fait parce que largement utilisés par une communauté. Dans la suite de cette section, nous essayerons de retenir les plus importants ou prometteurs.

Deuxièmement, le langage XML est devenu la norme pour formaliser les métadonnées et faire en sorte que les métadonnées soient aussi bien lisibles et compréhensibles par les personnes que par les ordinateurs.

Rien d'obscur ni de magique dans tout cela !

Dans tout projet d'archivage numérique, il conviendra donc d'examiner le contexte, les différentes contraintes, les pratiques et les besoins de la communauté d'utilisateurs avant de choisir un format de métadonnées et un profil d'application pour ce format.

Complément : le plan de métro des métadonnées

Le professeur James Turner de l'École de bibliothéconomie et des sciences de l'information, Université de Montréal, et son équipe effectuent un intéressant travail de catégorisation et de représentation de cette myriade de métadonnées et des organisations qui les élaborent. Il a conçu la carte du métro de métadonnées «MetroMeta » que l'on peut consulter à l'adresse : http://www.mapageweb.umontreal.ca/turner/meta/francais/

Chapitre 4 - Les métadonnées descriptives

A. 4.1 – Définition

Les métadonnées descriptives sont les métadonnées qui servent à organiser la connaissance. Ce sont les métadonnées qui vont permettre d'identifier, classifier, hiérarchiser l'information contenue dans l'objet numérique. Il s'agit typiquement d'un titre ou d'un nom, d'auteurs, de dates, de termes permettant la classification.

Organiser le savoir ou du moins un domaine de connaissance est une activité potentiellement génératrice d'une quantité importante de métadonnées. Nous parlons alors de taxonomie.

Les métadonnées descriptives sont appelées Information descriptive dans le modèle OAIS (voir les définitions données dans la partie 4 sur les normes et standards au terme Information descriptive). Il s'agit d'une information qui est construite à partir des métadonnées présentes dans l'AIP (voir les définitions données dans la partie 4 sur les normes et standards.au terme AIP) à destination d'une communauté d'utilisateurs ciblée. Les communautés et leurs domaines d'activités peuvent être extrêmement variés. La nature de l'Information descriptive est donc extrêmement liée aux spécificités du public auquel elle s'adresse.

Pour la description des documents d'archives, la DTD EAD est un format de métadonnées descriptives. Ce format est étudié précisément dans le module « Traitement des archives définitives », dans les parties traitant de l'EAD).

Exemple

Lien vers *le fichier FRAD024_91J.xml*[1]. Intitulé du fichier : instrument de recherche concernant le fond du commandant RIZZA concervé aux archives départementale de la Dordogne sous la côte 91J.

B. 4.2 - Un exemple de métadonnées descriptives : le format Dublin Core

Le Dublin Core est un schéma de métadonnées générique et simple. Initialement, il

1 - http://pleade.cg24.fr/sdx/pl/doc-tdm.xsp?id=FRAD024_0000000EF_de-163&fmt=arkheia&base=fa

s'appuyait sur 15 éléments de description formels (titre, auteur, éditeur), intellectuels (sujet, description, langue...) et relatifs à la propriété intellectuelle. Cet ensemble historique, créé en 1995, constitue le « Dublin Core Metadata Element Set ».

Il s'est enrichi d'éléments complémentaires : des relations (isPartOf, isVersionOf, isFormatOf, etc.), des référentiels complémentaires avec la liste des types permettant de caractériser plus finement la syntaxe de tel ou tel élément (DC: type). L'ensemble constituant avec les éléments d'origine, les « DCMI Metadata Terms ». Ce format est maintenu par le « Dublin Core Metadata Initiative » qui est une organisation constituée de membres émanant principalement de bibliothèques nationales et universitaires. Le Dublin Core est devenu la norme ISO 15836 en 2003.

Exemple : Un exemple de métadonnées Dublin Core pour la description d'un corpus oral

Nous présentions ci-après le code XML commenté du fichier de métadonnées puis sa représentation à des fins de consultation par les utilisateurs :

La langue par défaut étant l'anglais, si le contenu d'une balise est exprimée dans une autre langue, cela est indiqué par l'ajout d'un attribut :lang (cf. :title)

Les deux espaces de noms et correspondent respectivement aux 15 étiquettes de la norme ISO 15836 (Dublin Core) et pour aux « refinements et encoding » définis par le DCMI.

L'espace de nom OLAC est celui de l'organisation de même nom (Open Language Archives Community).

Les commentaires sont en italiques.

<!-- titres -->

<dc:title xml:lang="fr">L'arrivée à Marseille</dc:title>

<dcterms:alternative xml:lang="lad">La vinida a Marseya</dcterms:alternative>

<!-- durée du corpus en heures minutes secondes … -->

<dcterms:extent xsi:type="crdo:Duration">PT01M30S</dcterms:extent

<!—L'objet d'étude est le Judéo-espagnol (lad en ISO-639-3) -->

<dc:subject olac:code="lad" xsi:type="olac:language">Judéo-espagnol</dc:subject>

<!-- La ressource contient du Judéo-espagnol (lad en ISO-639-3) -->

<dc:language olac:code="lad" xsi:type="olac:language"/>

<!-- portée du contenu objet d'étude de la ressource --> <dc:coverage xml:lang="fr">Langue traditionnellement parlée dans l'ex-Empire ottoman. Variante de Salonique (Grèce).</dc:coverage>

<!-- classement suivant des typologies et des vocabulaires contrôlés proposés par OLAC -->

<dc:subject olac:code="text_and_corpus_linguistics" xsi:type="olac:linguistic-field"/>

<dc:type olac:code="dialogue" xsi:type="olac:discourse-type"/>

<dc:type olac:code="primary_text" xsi:type="olac:linguistic-type"/>

<!-- Identification du lieu de l'enquête : verbeuse pour l'humain, identifiant du Thesaurus of Geographic Names + longitude/lattitude dans une syntaxe proposé par DCMI -->

<dcterms:spatial xml :lang="fr">France, Bouches-du-Rhône, Aubagne.</dcterms:spatial>

<dcterms:spatial xsi:type="dcterms:TGN">1031883</dcterms:spatial>

<dcterms:spatial xsi:type="dcterms:Point">east=5.5833; north=43.2833</dcterms:spatial>

<!-- date d'enregistrement -->

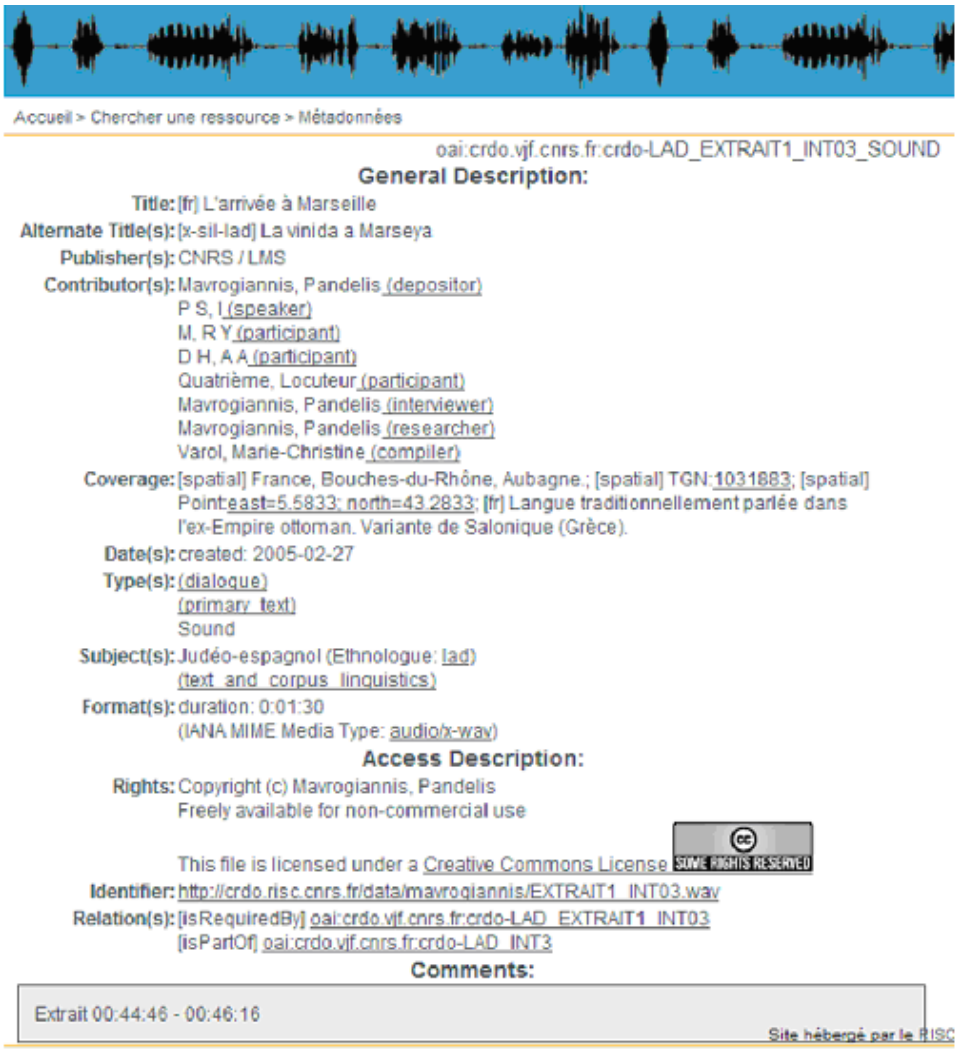
<dcterms:created xsi:type="dcterms:W3CDTF">2005-02-27</dcterms:created>

<!-- éditeur, laboratoire de rattachement du responsable de la ressource -->

<dc:publisher>CNRS / LMS</dc:publisher>

<!-- participants à la création de la ressource avec leur rôle (vocabulaire contrôlé OLAC). Certains participants ont été anonymisés -->
<dc:contributor olac:code="depositor" xsi:type="olac:role">Mavrogiannis, Pandelis</dc:contributor>
<dc:contributor olac:code="speaker" xsi:type="olac:role">P S, I</dc:contributor>
<dc:contributor olac:code="participant" xsi:type="olac:role">M, R Y</dc:contributor>
<dc:contributor olac:code="participant" xsi:type="olac:role">D H, A A</dc:contributor>
<dc:contributor olac:code="participant" xsi:type="olac:role">Quatrième, Locuteur</dc:contributor>
<dc:contributor olac:code="interviewer" xsi:type="olac:role">Mavrogiannis, Pandelis</dc:contributor>
<dc:contributor olac:code="researcher" xsi:type="olac:role">Mavrogiannis, Pandelis</dc:contributor>
<dc:contributor olac:code="compiler" xsi:type="olac:role">Varol, Marie-Christine</dc:contributor>
<!-- Type mime de la ressource et type DCMIType (vocabulaire contrôlé) -->
<dc:format xsi:type="dcterms:IMT">audio/x-wav</dc:format>
<dc:type xsi:type="dcterms:DCMIType">Sound</dc:type>
<!-- réafirmation du copyright et cession de droits par la licence CC + indication des moyens d'accès à la ressource -->
<dc:rights>Copyright (c) Mavrogiannis, Pandelis</dc:rights>
<dcterms:accessRights>Freely available for non-commercial use</dcterms:accessRights>
<dcterms:license xsi:type="dcterms:URI">http://creativecommons.org/licenses/by-nc-nd/2.5/</dcterms:license>
<!-- La ressource est un extrait d'une ressource plus vaste -->
<dcterms:isPartOf xsi:type="dcterms:URI">oai:crdo.vjf.cnrs.fr:crdo-LAD_INT3</dcterms:isPartOf>
<!-- La ressource est requise pour la consultation d'une autre ressource qui est en fait sa transcription -->
<dcterms:isRequiredBy xsi:type="dcterms:URI">oai:crdo.vjf.cnrs.fr:crdo-LAD_EXTRAIT1_INT03</dcterms:isRequiredBy>
<!-- URL de la ressource -->
<dc:identifier xsi:type="dcterms:URI">http://crdo.risc.cnrs.fr/data/mavrogiannis/EXTRAIT1_INT03.wav</dc:identifier>

Les mêmes métadonnées présentées aux utilisateurs



Accueil » Chercher une ressource » Métadonnées

oai:crdo.vjf.cnrs.fr:crdo-LAD_EXTRAIT1_INT03_SOUND

General Description:

Title: [fr] L'arrivée à Marseille
Alternate Title(s): [x-sil-lad] La vinida a Marseya
Publisher(s): CNRS / LMS
Contributor(s): Mavrogiannis, Pandelis (depositor)
P S, I (speaker)
M, R Y (participant)
D H, A A (participant)
Quatrième, Locuteur (participant)
Mavrogiannis, Pandelis (interviewer)
Mavrogiannis, Pandelis (researcher)
Varol, Marie-Christine (compiler)
Coverage: [spatial] France, Bouches-du-Rhône, Aubagne.; [spatial] TGN:1031883; [spatial] Point:east=5.5833; north=43.2833; [fr] Langue traditionnellement parlée dans l'ex-Empire ottoman. Variante de Salonique (Grèce).
Date(s): created: 2005-02-27
Type(s): (dialogue)
(primary_text)
Sound
Subject(s): Judéo-espagnol (Ethnologue: lad)
(text_and_corpus_linguistics)
Format(s): duration: 0:01:30
(IANA MIME Media Type: audio/x-wav)

Access Description:

Rights: Copyright (c) Mavrogiannis, Pandelis
Freely available for non-commercial use

This file is licensed under a Creative Commons License SOME RIGHTS RESERVED
Identifier: http://crdo.risc.cnrs.fr/data/mavrogiannis/EXTRAIT1_INT03.wav
Relation(s): [isRequiredBy] oai:crdo.vjf.cnrs.fr:crdo-LAD_EXTRAIT1_INT03
[isPartOf] oai:crdo.vjf.cnrs.fr:crdo-LAD_INT3

Comments:

Extrait 00:44:46 - 00:46:16

Site hébergé par le RISC

Vue utilisateurs

Chapitre 5 - Les métadonnées techniques

Elles constituent une des plus grandes nouveautés par rapport aux métadonnées nécessaires pour les archives sur support papier. Elles sont absolument nécessaires pour permettre la conservation sur le long terme des archives mais également pour la restitution, afin de savoir comment visualiser ce que l'on a conservé.

A. 5.1 – Définitions

Les métadonnées techniques sont les métadonnées qui servent à identifier, caractériser, définir l'environnement technique des objets numériques. Dans le modèle OAIS, les métadonnées techniques correspondent à l'Information de représentation OAIS (voir les définitions données dans la partie 4 sur les normes et standards au terme Information de représentation) qui sert à définir comment transformer un train de bits en une information intelligible.

Afin de réduire les risques, il est absolument nécessaire de conserver des données techniques périphériques aux informations à pérenniser. Ces données décrivent ce que peut contenir un format de représentation et comment l'exploiter.

L'identification peut se limiter à la reconnaissance du type de fichier soit à partir de son extension, de son « magic number » ou de son type « Mime » :

• le « Magic number » est une technique qui consiste à référencer les entêtes de fichier pour chaque format permettant ainsi de rapidement déterminer le type de fichier. Introduite dans les systèmes UNIX, elle correspond à la commande «file»,

• Les types MIME (Multipurpose Internet Mail Extension) sont des formats standards enregistrés par l'IANA (Internet Assigned Number Authority), http://www.iana.org/assignments/media-types/

Dans ces deux cas, le niveau d'information obtenue est assez pauvre : par exemple, nous allons savoir que nous détenons un fichier image de type TIFF mais sans connaissance de la version concernée. Pour la pérennisation, ce simple niveau d'information n'est pas suffisant. Il convient d'identifier le type de format de la manière la plus précise possible et en particulier obtenir la version exacte du format de l'objet numérique.

Pour aller plus loin, il peut s'avérer nécessaire de caractériser de manière complète un objet numérique. Il s'agit alors non seulement d'identifier précisément le type de format mais également de définir les choix techniques qui ont été retenus pour l'application de ce format : type de compression, type de codage par exemple. Le format peut se référer à une norme mais nous avions vu que les normes sur les formats se présentaient comme des poupées russes et que la connaissance de la poupée la plus grande ne permettait pas pour autant que connaitre les caractéristiques des plus petites.

Avec une définition précise et complète du format de l'information, il devient possible de valider le format c'est-à-dire :

• de s'assurer que les caractéristiques techniques définies par les spécifications sont bien vérifiées (conformité par rapport à la norme ou au standard),

• mais également de vérifier que ces caractéristiques respectent une spécification particulière liée à une application spécifique (conformité par rapport à des règles d'utilisation ou des conditions restrictives qui ont été décidées pour l'archivage).

De plus, la définition des caractéristiques est souvent nécessaire pour développer les outils de transformation mis en œuvre lorsqu'une migration de format est à effectuer.

Les métadonnées techniques ne se limitent pas à la définition des formats. En prévision d'opérations de migration de format ou d'utilisation d'outils d'émulation, il faut définir l'environnement technique tant logiciel que matériel, de création ou de restitution de l'objet numérique.

Exemple : Exemples :

- Systèmes d'exploitation : « Windows 98 et supérieur » : supérieur jusqu'où ???

– Environnement logiciel : « Word 98 et supérieur » : supérieur jusqu'où ??? Compatibilité avec d'autres logiciels (Open office) ? Risques de pertes de fonctionnalités ?

– Environnement matériel : les périphériques : comment émuler le comportement d'un joystick ou d'un crayon optique sur un PC ? Problèmes de vitesse de traitement ?

Complément : les métadonnées techniques des formats image

Il existe plusieurs formats de **métadonnées internes** :

- EXIF (Exchangeable Image File) : ensemble de métadonnées essentiellement techniques relatives à la prise de vue et fournies automatiquement par l'appareil numérique (fabricant et modèle de l'appareil, hauteur et largeur de l'image, date et heure de la prise de vue, orientation, résolution, temps d'exposition, ouverture, présence d'un flash, etc.), qu'il est possible d'intégrer dans des images JPEG/JFIF notamment. Le format EXIF a été développé en 1995 par la JEIDA (Japan Electronic Industry Development Association) ; la version 2.2 actuelle date de 2002 ;

- IPTC-NAA/IIM (International Press and Telecommunications Council - Newspaper Association of America / Information Interchange Model) : ensemble de métadonnées essentiellement sémantiques de l'image et nécessitant l'intervention d'un opérateur humain pour être renseignées (identifiant, titre, auteur, copyright, date de création, mots-clés, lieu, etc.), qu'il est possible d'inclure dans des images JPEG/JFIF ou TIFF. Les informations saisies dans le fichier de récolement ou un fichier d'indexation peuvent être réutilisées automatiquement pour alimenter les champs IPTC. La première version du modèle IPTC a été publiée en 1991 ; la version 4.1 actuelle date de 1999 .

- XMP (Extensible Metadata Platform) : format de métadonnées basé sur XML, créé par Adobe en 2001, utilisé à l'intérieur des fichiers d'images (JPEG/JFIF, TIFF, GIF, PNG, PDF, SVG...). Même s'il prédéfinit la façon de stocker un certain nombre d'informations les plus courantes, en reprenant en particulier des éléments de Dublin Core et d'EXIF, XMP est ouvert à tout type de métadonnées XML. Il est possible d'exploiter les métadonnées XMP même en l'absence des applications d'origine.

Ainsi qu'un format normalisé de métadonnées externes

Il existe un format de métadonnées spécifique pour la caractérisation technique des images fixes numériques : le Data Dictionary - Technical Metadata for Digital Still Images, norme ANSI/NISO Z39.87 publiée en décembre 2006 . Ce dictionnaire de données possède une déclinaison sous forme de schéma XML : MIX (Metadata for Images in XML) . Les 200 éléments prévus par le dictionnaire de données sont répartis en cinq familles : information de base sur l'objet numérique (identifiant, taille, format, compression, fixité), informations de base sur l'image (dimensions, couleur...), métadonnées de capture de l'image (taille de la source, date de capture, informations sur le scanner, informations sur la caméra numérique, coordonnées géographiques), métadonnées d'évaluation de l'image (échantillonnage de capture, échantillonnage colorimétrique...), historique des modifications.

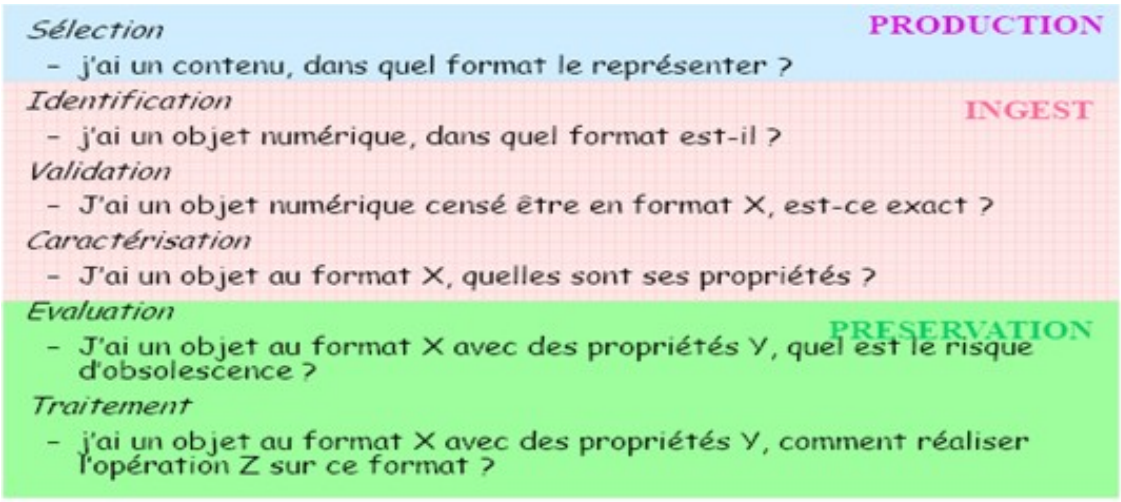
1-Les spécifications EXIF sont consultables à l'adresse : http://www.exif.org/

2-Le modèle IPTC est consultable à l'adresse : http://www.iptc.org/IIM/

3-Le dictionnaire de données est consultable à l'adresse : http://www.niso.org/standards/index.html

4-La documentation sur MIX est consultable à l'adresse : http://www.loc.gov/standards/mix/

B. 5.2 – Informations à rassembler concernant un format



Informations à rassembler concernant un format (d'après Emmanuelle Bermes)

- Noms canoniques et variantes : par exemple PDF, Adobe PDF, Portable Document Format
- Signatures internes et externes : « .pdf », magic number
- Spécifications du format : http://partners.adobe.com/public/developer/en/pdf/PDFReference16.pdf

  Attention, on ne peut pas se satisfaire d'un lien externe pour disposer de la spécification technique complète du format. La pérennité de ce lien n'est nullement garantie. On devra donc soit récupérer ces spécifications et les archiver également, soit s'assurer de la pérennité du site qui dispose de ces spécifications ainsi que de la pérennité de l'adresse de ce site.

- Auteurs, titulaires de droits, maintenance : Société Adobe
- classifications et relations : PDF <has subtype> PDF 1.4, PDF 1.7, PDF/A, PDF/X…
- systèmes, services et outils permettant de lire les données conformes à ce format : Adobe Acrobat Reader...

Complément : DROID et JHOVE, des outils par identifier, caractériser, valider les formats

DROID (Digital Record Object Identification) est un outil Open Source sous licence BSD fourni par les Archives nationales du Royaume-Uni. Il s'appuie sur le registre de format PRONOM également supporté par les Archives nationales du Royaume-Uni. PRONOM cherche à fournir les caractéristiques et des références sur la documentation de chaque version de format. Il définit un identifiant unique pour chacune de ces versions, appelé PUID (Pronom Unique Identifier). DROID identifie le format d'un objet numérique en fournissant son PUID.

http://www.nationalarchives.gov.uk/aboutapps/fileformat/pdf/automatic_format_identification.pdf

JHOVE (JSTOR/Harvard Object Validation Environment) est un outil Open source sous licence GNU Lesser General Public de validation de formats fourni par JSTOR et l'université de Harvard.

http://hul.harvard.edu/jhove/

Le CINES (Centre informatique national de l'enseignement supérieur) Montpellier en France, offre un service à distance de validation des formats, basé sur l'outil JHOVE : http://facile.cines.fr/

En septembre 2009, ce service permettait de valider les fichiers aux formats suivants : AIFF, FLAC, GIF, HTML, JPEG, MPEG-4, OGG, ODT, PDF, PNG, SVG, TIFF, TXT, WAVE, XML

Pour chacun de ces formats, vous trouverez des informations disponibles sur Internet : signification du sigle, type de données auxquelles ce format s'applique, caractéristiques techniques, etc.

Chapitre 6 - Les métadonnées administratives

Les métadonnées administratives sont les métadonnées qui servent à gérer la vie de l'objet numérique. Dans le modèle OAIS, les métadonnées administratives représentent l'Information de pérennisation qui sert à identifier, authentifier, à définir la provenance et le contexte de l'information à pérenniser ainsi que les droits attachés à cette information OAIS (voir les définitions données dans la partie 4 sur les normes et standards au terme Information de pérennisation).

Naturellement, une partie de ces métadonnées administratives sera également utilisée dans le processus de recherche d'information.

Elles regroupent les métadonnées d'identification, de contexte, de provenance, d'intégrité et de gestion de droits.

A. 6.1 – Les métadonnées d'identification

Les métadonnées d'identification permettent l'identification univoque des objets archivés. Généralement, il doit s'agir d'un identifiant pérenne. Le choix d'un type d'identifiant est loin d'être une opération anodine. Ce choix est stratégique dans le cadre de la pérennisation. Les types d'identification sont nombreux. L'objectif est d'adopter un identifiant qui résistera à l'épreuve du temps, aux évolutions de classement intellectuel des contenus, aux changements d'organisation physique des données. Il doit être adaptable. Il doit également supporter l'évolution de la taille de l'Archive. Imaginez les conséquences que pourrait entraîner un choix d'identifiant qui trouverait une limite après quelques dizaines d'années ? Il doit être extensible. Il doit pouvoir éventuellement être capable de permettre d'identifier plusieurs niveaux d'information d'un objet ou d'une collection. Il doit être « granulaire ».

Jusqu'à ces dernières années, des identifiants signifiants étaient préconisés :

Exemple : Exemple aux Archives nationales en France :

AP : archives privées

- 400 AP : archives Napoléon
  - 400 AP 106 à 167 : correspondances et pièces diverses classées par ordre alphabétique
    - 400 AP 106 : Aali Pacha-Alexandre Jean, prince de Roumanie

Toutefois, aujourd'hui, d'une part, la gestion de fichiers numériques rend les systèmes de cotation traditionnels complexes à mettre en œuvre et, d'autre part, rares sont les Archives qui soient en complet circuit fermé sans ouverture sur le réseau Internet. Il convient par conséquent de re-considérer la capacité des identifiants à être diffusés et « citables » sur le web.

Complément : Un exemple d'identifiants « opaques » : les identifiants ARK adoptés à la bibliothèque nationale de France

ARK « Archival Resource Key » est un système d'identifiants pérennes créé et maintenu par la California Digital Library.

http://www.cdlib.org/inside/diglib/ark/

La nécessité de disposer d'un moyen d'identification pérenne pour les objets accessibles sur le réseau n'est pas nouvelle. Dès le début des années 90, est apparue la nécessité d'avoir un moyen d'identification qui ne soit pas dépendant du nom de l'ordinateur dans lequel les objets sont stockés et qui réponde aux besoins de gestion des noms ; par exemple en évitant d'avoir à renommer les identifiants lorsque la localisation change. Une réponse a été les URN (Uniforme Ressource Name) bientôt suivie des mécanismes et services d'affectation d'identifiants comme les DOI (Digital Object Identifier) et PURL (Persistent Uniform Ressource Locators). Ces mécanismes s'appuient sur la redirection pour garantir la persistance d'accès à un objet. Le problème de ce type d'approche est lié à la gouvernance des identifiants. Elle repose sur une vision centralisée des services de résolution de nom et certains mécanismes sont propriétaires. Sur le long terme, le fait d'être lié à un organisme est un risque. C'est une problématique identique qui a mené à la mise en place des DNS (Domain Name System) pour décentraliser et distribuer la résolution des noms de domaine. Les identifiants ARK s'appuient sur des mécanismes similaires pour limiter les dépendances.

Leur mise en oeuvre permet :

* d'afficher l'identifiant pérenne dans la barre d'URL lors de la consultation d'un document numérisé ;

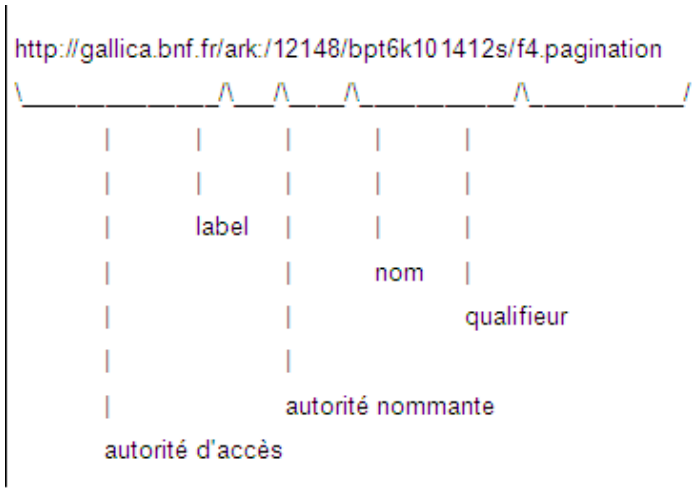
* de conserver dans l'URL le nom de domaine du contexte de visualisation ;

* d'appeler chaque service de visualisation (pagination, table des matières, etc.) dans l'URL à l'aide d'un paramètre simple, nommé "qualifieur" ;

* d'obtenir plus facilement qu'auparavant l'URL d'une page précise au sein d'un document par exemple numérisé.

Regardons de plus près la structure d'un identifiant ARK. C'est une suite de caractères qui comprend toujours le label ark:/ Il est composé de cinq parties : un préfixe qui définit l'autorité d'accès, un label qui est toujours « ark: », un identifiant d'autorité nommante, un nom qui désigne l'objet à identifier et finalement un « qualifier » qui permet de référencer plus finement les éléments constitutifs d'un objet.

L'autorité d'accès et le « qualifier » sont optionnels. Un exemple d'identifier ARK peut être :



Structure identifiant ARK

Complément

L'autorité nommante n'est pas obligatoirement la même que l'autorité d'accès. Ainsi, à la manière des DNS, une autorité d'accès peut résoudre les identifiants ARK qu'elle n'a pas créés en redirigeant l'identifiant vers l'autorité d'accès connue pour l'autorité nommante.

Une autorité nommante est libre de nommer ses objets comme elle le désire. Néanmoins, le nom et le qualifier de l'identifiant ARK ne doit pas dépasser 128 caractères et seuls certains caractères sont autorisés (lettres, chiffres et quelques caractères spéciaux). Pour la préservation, il est fortement conseillé de respecter une syntaxe qui interdit les noms signifiants. Un nom non-signifiant est à priori plus pérenne : ce que je nomme d'une façon aujourd'hui n'est peut-être pas valable demain. De plus, cela permet de s'absoudre de la problématique des langues. Enfin, cela facilite la génération automatique des noms. Une autorité nommante s'engage sur la pérennité d'accès à l'objet par ce nom. La dernière partie d'un identifiant ARK qui décrit les sous-hiérarchies et les variantes est optionnelle puisque que par définition non-pérenne.

Ainsi, la spécification ARK permet la définition d'identifiants qui ont la propriété d'établir un lien indépendant des systèmes et des organisations avec un objet. Elle définit également les services permettant d'obtenir de l'information descriptive sur l'objet.

Les identifiants ARK ont une bonne approche du problème. Néanmoins, l'adoption de ce mécanisme reste limitée à quelques organisations. La résolution partagée des identifiants n'est pas véritablement effective.

B. 6.2 - Les métadonnées de provenance et de contexte

Les métadonnées de provenance et de contexte informent sur la vie de l'objet numérique. C'est le comment et le pourquoi : d'où vient l'objet numérique ? Comment l'information a-t-elle été collectée ? Par quel moyen technique (capteur de satellite, réception de signal, numérisation) ? Quels traitements ont été effectués ? Par qui ? Pourquoi ? Quand ? Quelles sont les raisons qui ont motivé sa création, sa collecte, sa réception ? Quelle confiance pouvons-nous avoir dans cette source ?

C. 6.3 – Les métadonnées d'intégrité

Les métadonnées d'intégrité permettent de disposer d'informations relatives au respect de l'intégrité des objets. Ces informations sont à usage interne (surveillance et contrôle des objets) et/ou à usage externe (prévenir des litiges, des contentieux). Il s'agit par exemple d'empreintes des fichiers à conserver, obtenus sur la base de procédés cryptographiques.

D. 6.4 – Les métadonnées de droits

Elles permettent de gérer le statut légal de l'objet numérique. Il ne s'agit pas ici des DRM (Digital Right Management) au sens de Moyens Technique de Protection (MTP) qui visent à contrôler, contraindre, empêcher des usages qui peuvent être faits par des utilisateurs. Rentrent dans cette catégorie, pour les archives publiques, les délais légaux de communicabilité fixés par la réglementation. Ces mécanismes sont mis en œuvre soit par le système de diffusion soit par l'Archive elle-même au moment de l'accès.

Il s'agit ainsi des métadonnées qui vont permettre de définir la politique de diffusion en fonction de contraintes légales ou imposées par l'entité détentrice des droits. La concrétisation formelle de cette politique peut être une licence. On parle de droits d'usage et de droits d'accès. Ils définissent en fonction du contexte d'utilisation : Qui ? Avec quoi ? (baladeur numérique, télévision, ordinateur) ce qu'il est permis de faire : trouver, voir, imprimer, copier, modifier, détruire, et les contraintes d'usage : qualité, quantité, prix...

Les métadonnées de droit peuvent être incluses dans un format de métadonnées plus général comme le format TEF. Elles peuvent aussi, surtout lorsqu'elles sont complexes, être décrites au moyen de langages spécifiques à ce type de métadonnées. Ces langages doivent permettre à tout moment, de déduire les contraintes de communicabilité d'un objet numérique à partir de règles définies. Citons ici ODRL (Open Digital Rights Language) dont l'ambition est de devenir un standard en matière d'expression des droits.

Complément : dictionnaire de métadonnées PREMIS

Bibliographie :

OCLR - RLG, Data dictionary for preservation metadata, Dublin, Ohio, 2005 : http://www.oclc.org/research/projects/pmwg/premis-final.pdf

Elaboré à l'initiative de OCLC (Online Computer Library Center) et RLG (Research Library Group), PREMIS vise à identifier les métadonnées nécessaires pour assurer les principales fonctions dans un contexte de pérennisation, définir des métadonnées dont la fourniture ou l'analyse ne nécessite pas d'intervention humaine, élaborer un dictionnaire de données totalement indépendant de l'implémentation et reposant sur des « unités sémantiques » (semantic units). PREMIS n'est pas un format de type « capsule » qui servirait à construire les paquets d'information, mais un dictionnaire définissant des éléments de métadonnées jugés indispensables à une fonction d'archivage pérenne.

Les principales unités sémantiques

Forces

Le principal avantage de PREMIS, par rapport aux autres formats, est qu'il est conçu pour prendre en compte les besoins en matière de préservation du numérique, et qu'il est maintenu par la Bibliothèque du Congrès. Il s'appuie sur des pratiques qui tiennent compte de l'existant et sur une forte communauté d'utilisateurs. Enfin, PREMIS accorde une large place à l'extensibilité.

Faiblesses

PREMIS ne répond pas à tous les besoins en termes de métadonnées de pérennisation, puisqu'il ne prévoit que les métadonnées communes à tous les types de fichiers numériques, en laissant à l'utilisateur le soin de choisir d'autres formats pour les métadonnées techniques plus spécifiques. Ainsi, l'interopérabilité ne sera nécessairement que partielle avec les autres utilisateurs de ce format.

E. 6.5 - Les métadonnées orientées gestion des archives courantes et intermédiaires

La question des métadonnées est abordée dans les normes afférentes des archives courantes et intermédiaires (voir les définitions données dans la partie 4 sur les normes et standards).

Les métadonnées sont effet essentielles pour permettre une qualification certaine des «archives courantes et intermédiaires » et pouvoir ainsi prouver leur authenticité et intégrité.

C'est ainsi que dans cette famille de normes, on peut se reporter à la norme ISO 23081-1, Information and documentation, Record management processes, Metadata for records, part 1 : principles, Genève, 2006. Elle définit par exemple, outre les principales classes définies dans les parties précédentes, une catégorie supplémentaire de métadonnées dédiée au commerce électronique.

On peut également citer les métadonnées listées dans le modèle MoReq2

Bibliographie

MoReq2 specification – Model Requirements for the Management of Electronic Records, Update and Extension, 2008. Traduction française : Exigences-types pour la maîtrise de l'archivage électronique. Mise à jour et extension - 2008. Spécifications MoReq2 (chapitre 0, texte principal, annexes) : site de la direction des Archives de France, *http://www.archivesdefrance.culture.gouv.fr/static/2094*

On retrouve des jeux de métadonnées différents pour le plan de classement, les séries, dossiers, sous-dossiers, volumes et documents, les règles de conservation/destruction, les composants, les métadonnées témoins (après destruction), les types de documents archivés, les extraits, les acteurs, les entités/acteurs.

Les métadonnées incluent les données d'indexation et d'autres données indispensables à un archivage efficace, notamment les droits et restrictions d'accès. A noter que MoReq2 ne considère pas les données d'historique des événements comme des métadonnées à part entière.

La nature des métadonnées, leurs modalités de production et leurs caractéristiques sont détaillées dans l'annexe 9 : 157 métadonnées distinctes sont ainsi analysées.

Accès à la version française de Moreq2 : *http://www.archivesdefrance.culture.gouv.fr/gerer/publications/manuels/#moreq2*[2]

2 - http://www.archivesdefrance.culture.gouv.fr/gerer/publications/manuels/#moreq2

Chapitre 7 - Les métadonnées de structure

Elles donnent les moyens de gérer l'arborescence des objets complexes et de les restituer. En effet, même si un document simple ne pose pas de problème de structure en soi, on peut tout de même avoir plusieurs niveaux d'accès. Elles permettent ainsi de connaître tous les fichiers qui composent un document ainsi que leurs relations entre eux.

A. 7.1 – Définitions et exemples de formats de métadonnées de structure

Les métadonnées de structure servent à connaître l'organisation de l'information contenue et des objets numériques.

Il y a deux niveaux de structure : un niveau logique et un niveau physique :

- Le niveau logique définit les liens entre des éléments qui ont du sens pour l'utilisateur : numéro de page, de plages audio, titre de chapitres, d'articles, etc.
- Le niveau physique définit comment sont enregistrés les objets numériques : dans quel fichier ? Dans quel répertoire ? Sur quel support ? On parle également de carte de structure logique, de structure physique.

Dans le modèle OAIS, les métadonnées de structure représentent l'Information d'empaquetage OAIS

(voir les définitions données dans la partie 4 sur les normes et standards au terme Information d'empaquetage).

Légende du tableau : Les métadonnées de structure donnent le moyen de gérer l'arborescence des objets complexes et de les restituer.

Complément : Les formats d'empaquetage METS ; XFDU et MPEG21

Les formats d'empaquetage ne définissent pas seulement des métadonnées de structure, ils définissent des paquets pouvant contenir les données à préserver, les métadonnées

associées à ces données (descriptives, techniques, administratives) ainsi que des métadonnées (de structure) décrivant l'organisation logique et physique de cet ensemble. Les paquets en question peuvent être réels (toutes les données et métadonnées sont stockées ensemble), dans ce cas, le paquet à une réalité physique. Les paquets peuvent être également virtuels au sens où l'on définit essentiellement des pointeurs vers des données et des métadonnées stockées à des endroits différents.

Trois formats d'empaquetage méritent d'être présentés : METS, XFDU et MPEG21-DIDL

Complément : METS (Metadata Encoding and Transmission Standard).

Bibliographie

Digital Library Federation, Metadata Encoding and Transmission Standard : primer and Reference Manual, version 1.6, 2007.

http://www.loc.gov/standards/mets/

Maintenu par la Bibliothèque du Congrès, METS est un format conçu pour gérer tout type d'objet numérique, simple ou complexe. C'est un format de type « capsule », qui peut donc intégrer tout autre format de métadonnées descriptives ou techniques. Il a pour principale caractéristique d'être modulaire (une instance METS est composé de sept sections) et de séparer la structure de l'objet et les métadonnées (les métadonnées sont regroupées dans des sections spécifiques, et associées aux objets correspondants par le biais de liens ou pointeurs). Il permet de définir à la fois une carte de structure physique et logique. METS est ainsi une enveloppe, un conteneur de métadonnées. Des métadonnées sur les métadonnées en quelque sorte. Il ne définit pas quelles sont les métadonnées à utiliser mais permet de catégoriser les métadonnées et de les lier entre elles. Une partie du modèle consiste à inclure des métadonnées ou à pointer vers des métadonnées externes. La description logique et la description physique des métadonnées sont séparées dans des parties distinctes de l'enveloppe. Le standard définit le mécanisme de gestion et d'organisation du système de liens entre les différents éléments. Ce principe est extrêmement puissant car il permet de s'adapter à tous les types d'organisations de données. La contrepartie réside dans la complexité du réseau de liens créés.

METS propose d'organiser la modélisation de ces métadonnées en sept parties :

• l'entête METS identifie la date de création et le créateur des métadonnées,

• les métadonnées descriptives (par exemple, Dublin Core ou encore EAD) ; c'est la carte d'identité de l'objet ou partie d'objet numérique référencé,

• les métadonnées administratives sont subdivisées en quatre sous parties :

o les métadonnées techniques : informations sur les caractéristiques des fichiers (taille, date de création, type de fichier, etc.),

o les métadonnées des droits intellectuels : informations sur les droits d'accès et d'usage,

o les métadonnées de la source analogique : informations sur l'objet d'origine sous forme analogique, s'il y a lieu (forme, taille, type de papier ou de film, notice bibliographique d'origine, référence de l'original)

o les métadonnées de la provenance numérique : informations qui décrivent les processus de création/migration/transformation de l'objet numérique.

• la partie fichier décrit l'organisation physique des fichiers,

• la carte de structure décrit l'organisation logique (hiérarchique) des objets numériques (dossier, rapport, page, article, paragraphe). C'est un élément central. Elle établit les liens avec les métadonnées et l'organisation physique des fichiers,

• les liens de structure contiennent les hyperliens entre les différents niveaux de la carte de structure. Cette partie est utilisée pour la description d'objets provenant du web,

• la partie comportement décrit les outils nécessaires à l'exploitation des objets numériques.

Forces

Le respect des concepts de l'OAIS fait du format METS un bon moyen d'accueillir les métadonnées retenues pour l'archivage numérique. Le standard prévoit d'inclure, ou de référencer, non seulement les métadonnées au format XML mais prévoit aussi un mécanisme pour inclure les autres formats en les considérant comme des objets binaires. Cette capacité lui permet de s'adapter à la plupart des besoins dans ce domaine.

METS est un format qui a acquis une certaine maturité et sur lequel la communauté des bibliothèques numériques possède une bonne visibilité. La plupart des projets de préservation numérique déclarés l'utilisent. Du point de vue technique, le fait de pouvoir intégrer à la fois une carte de structure physique et logique est un avantage, car il permet de préserver des données importantes pour l'accès et pour la « représentation » du document, comme les tables des matières. METS permet en outre de catégoriser les métadonnées.

Faiblesses

Pour des objets numériques comportant un grand nombre de fichiers, l'instance METS est difficile à lire et à comprendre d'emblée (pour un humain), du fait d'une gestion par liens complexe. Par ailleurs, METS ne possède pas de modèle conceptuel, ce qui rend sa migration vers un autre format plus difficile. Enfin, ce format est peu documenté et certains éléments sont difficiles à interpréter précisément.

Les métadonnées à utiliser restent à définir au cas par cas. Et même si de nombreuses initiatives ont abouti à l'élaboration de métadonnées adoptées par une large communauté d'utilisateurs (Dublin Core, EAD, PREMIS, etc.), la plupart des métadonnées nécessaires à la pérennisation n'ont pas encore trouvé de standard largement diffusé.

XFDU (XML Formatted Data Unit) [XFDU 08]

XFDU (XML Formatted Data Unit) [XFDU 08]

Bibliographie

CCSDS, CCSDS 661.0-B-0, XML Formatted Data Unit (XFDU) Structure and Construction Rules, 2008.

http://public.ccsds.org/publications/MOIMS.aspx

Le format XFDU a été créé par le Comité consultatif pour les Systèmes de Données spatiales (CCSDS), pour faciliter l'implémentation du modèle de référence OAIS. Les concepts qu'il utilise, et sa terminologie, sont donc ceux de l'OAIS. XFDU est un format de type « capsule » qui permet de créer l'enveloppe du paquet d'information. Il est très proche de METS, dont il reprend les mécanismes de base (distinction entre section de métadonnées / section d'objets de données ; forte utilisation des liens et des pointeurs) et jusqu'à certains noms d'éléments (flocat, dmdsec, mdwrap). Il ne définit pas les éléments de métadonnées particuliers, mais offre un moyen de les catégoriser selon la terminologie OAIS. La version actuelle est figée depuis septembre 2008. Elle a été adoptée comme standard du CCSDS et sera proposé à l'ISO.

Forces

XFDU est un candidat très sérieux pour l'organisation de l'information d'empaquetage. Il est ainsi utilisé pour l'archivage des données d'observation de la Terre de l'Agence Spatiale Européenne.

Faiblesses

XFDU est un format qui a été approuvé dans sa version définitive en septembre 2008. Il est encore très récent et les réalisations ou les expérimentations sont encore limitées. La documentation est assez lacunaire (elle contient peu d'exemples mais un tutoriel doit être publié par le CCSDS). Un certain nombre d'outils ont été développés par les agences spatiales impliquées dans la normalisation du XFDU mais ces outils n'ont pas vocation à devenir des logiciels libres.

Complément : MPEG21-DIDL (Digital Item Declaration Language)

MPEG21-DIDL (Digital Item Declaration Language)

MPEG21-DIDL (Digital Item Declaration Language) constitue une partie de MPEG21 qui est un ensemble de normes élaborées par l'industrie des contenus numériques, visant à réaliser un cadre interopérable pour la diffusion et l'échange d'objets numériques. DIDL est un format libre et ouvert, dont le potentiel n'est optimal que si l'on utilise aussi les autres parties de MPEG21, diffusées par l'ISO.

Du point de vue des caractéristiques techniques, DIDL est un format de type « capsule », dans lequel on peut intégrer tout type de métadonnées. Son modèle de données permet de gérer un nombre non limité de niveaux de granularité (container / item / (sous-)item /…/ component). À la différence de METS, les métadonnées se situent au même endroit que l'entité à laquelle elles s'appliquent. La carte de structure physique est déduite de l'arborescence de l'instance DID et il n'y a pas de moyen aisé de faire apparaître une carte de structure logique. Les métadonnées sont encapsulées dans des éléments « Descriptor », qui sont répétables et peuvent donc servir de blocs modulaires.

Forces

Le principal avantage de DIDL est sa relative simplicité, qui se traduit par une bonne lisibilité des instances DID. On repère facilement les métadonnées et les entités auxquelles elles s'appliquent. Bien que ce format n'ait pas été conçu pour la préservation à long terme de l'information numérique, il peut être utilisé dans cette optique, y compris dans le respect du modèle de référence OAIS. Le fait qu'il repose sur un modèle conceptuel est un atout permettant, si les besoins s'en font jour, de le convertir facilement dans un autre format.

Faiblesses

Utiliser DIDL est relativement audacieux, dans la mesure où il n'a pas fait encore la preuve de sa maturité et que son utilisation dans la communauté des bibliothèques numériques reste très marginale (à signaler tout de même : le Laboratoire national de Los Alamos, qui a pris part à l'élaboration de la spécification DIDL et l'Université Old Dominion de Norfolk, dans le cadre du programme américain National Digital Information Infrastructure and Preservation Program (NDIIPP) piloté par la Bibliothèque du Congrès). Du point de vue technique, DIDL a deux inconvénients importants : l'impossibilité de donner des attributs aux entités Container/Item/Component pour les catégoriser finement et le fait de devoir se limiter à une seule carte de structure (physique ou logique).

B. 7.2 – Un format orienté transfert : le standard d'échange de données pour l'archivage français (SEDA)

Site de la direction générale de la modernisation de l'Etat

https://www.ateliers.modernisation.gouv.fr/ministeres/projets_adele/a103_archivage_elect/public/standard_d_echange_d/folder_contents

Site de la direction des Archives de France

http://www.archivesdefrance.culture.gouv.fr/gerer/archives-electroniques/standard/

7.2.1 – Objectifs et description générale

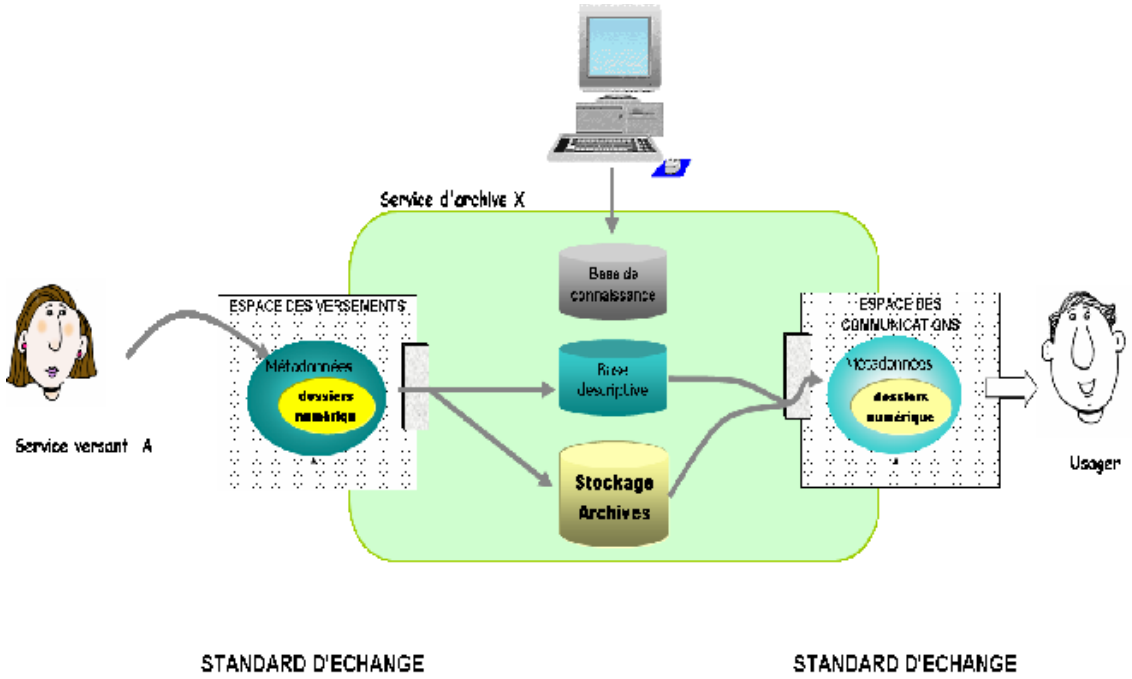
Son objectif a été dans un premier temps très concrètement de pouvoir préciser aux services producteurs le format dans lequel les objets d'archives devaient être transférés accompagnés de leurs métadonnées descriptives, informations de pérennisation et informations de représentation (bordereau de versement) pour ensuite pouvoir être intégrés dans une plateforme d'archivage électronique.

Il s'agissait par conséquent de définir un format pivot permettant de faciliter les échanges entre un service d'archives et ses interlocuteurs et de permettre l'automatisation des transferts et de l'élaboration des bordereaux de versement, ces derniers étant automatiquement « nourris » des métadonnées métier présentes dans les systèmes d'information des producteurs (applications métier sous forme de bases de données, gestions électroniques de documents, flux de données dématérialisés). Il s'agissait ainsi d'éviter les ruptures de charge entre chacun des partenaires.

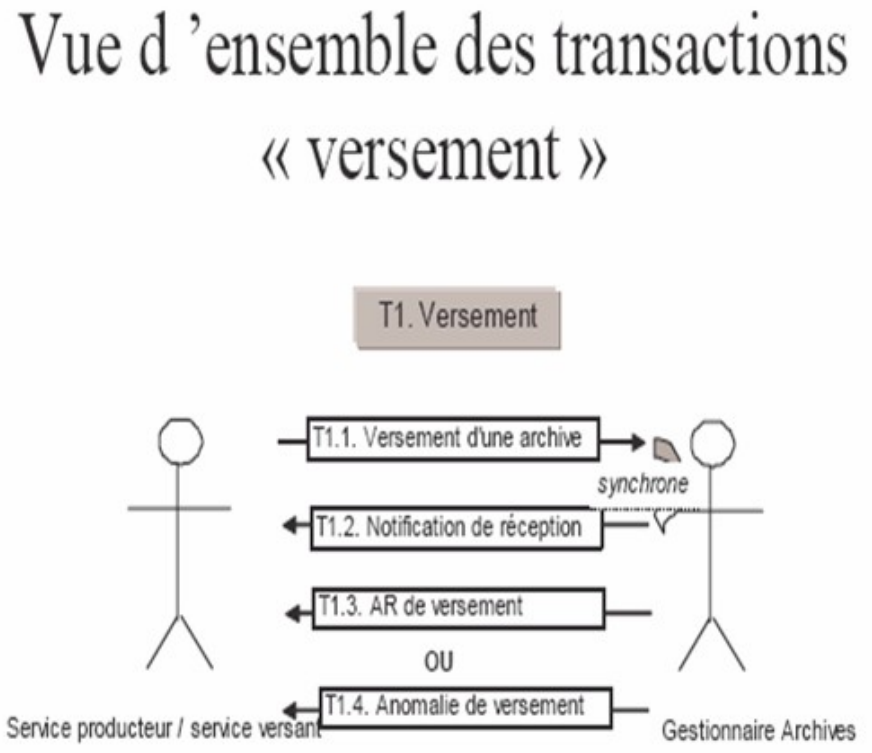
Ainsi, le SEDA définit à la fois :

- la structure et le contenu des messages échangés entre les acteurs du processus, qu'il s'agisse d'un transfert, d'une élimination, d'une communication, d'une restitution
- la structure et le contenu des bordereaux de versement accompagnant les objets d'archives.

La structure de ces bordereaux est conforme à la norme de description des documents d'archives (ISAD-G –précisément présentée dans le module Traitement des archives définitives) et l'ensemble du SEDA est conforme aux règles de l'organisation de normalisation internationale des Nations-Unies : l'UN-Cefact : *http://www.unece.org/cefact/*



Standard échange



Vue d'ensemble des transactions "versements"

7.2.2 – Mise en œuvre du SEDA

Lorsqu'un service producteur souhaite transférer des objets à archiver pour la première fois, la première tâche consiste à élaborer un modèle de description pour cette catégorie documentaire, dérivée du SEDA, de manière à pouvoir spécifier précisément la forme des exports qui seront ensuite mis en œuvre à partir du système d'information du producteur (application métier, gestion électronique de documents, plateformes sécurisées de télétransmission avec signature électronique des transactions….), quel que soit les modalités de l'export (par réseau, automatique, manuel).

La structure et la granularité du classement sont définies (description à un ou plusieurs niveaux : par exemple versement des dossiers de personnels clos en 2009 et au niveau de description inférieur, dossier de personnel de monsieur Durand). Pour chaque niveau, on définit les éléments de description associés suivant la structure du SEDA :

- la plupart notamment aux niveaux inférieurs de la description seront récupérés automatiquement à partir du système d'information (dans notre exemple, des éléments concernant l'identité de l'agent et les principales étapes de sa carrière) ;
- d'autres seront définis a priori (par exemple des mots clés issus des thesaurii des services des archives) ;
- d'autres encore ne seront définis qu'au moment du transfert (par exemple la date du versement).

A partir de ce moment-là, les spécifications sont définies qui permettront d'intervenir informatiquement sur le système d'information du producteur pour préparer concrètement les versements (afficher les dossiers clos en 2009) et générer automatiquement les exports.

Jusqu'à présent, les modèles de description appelés « profils » sont élaborés manuellement.

Exemple

Une application est actuellement en cours de développement au sein de la direction des

Archives de France afin d'assister l'utilisateur à l'élaboration de ces profils et générer ensuite des schémas XML pour chaque profil. Si le profil concerne une catégorie qu'on retrouve sur l'ensemble du territoire, il est élaboré et publié avec l'assistance de la direction des Archives de France, de manière à ce que chaque service d'archives puisse l'exploiter pour ses propres besoins.

Un export (SIP) lors d'un transfert d'un service producteur vers un service d'archives est par conséquent constitué d'un fichier conteneur dans lequel on trouve

- à la fois le bordereau de versement .xml
- et les fichiers de données associés.

Le lien entre les deux se fait par l'intermédiaire d'identifiants URL.

Lors de l'arrivée dans le service d'archives, le fichier est analysé, contrôlé, validé ou non, des messages sont échangés.

Lors de la prise en charge emportant transfert de responsabilité, les informations du bordereau de versement sont intégrées et éventuellement enrichies dans l'outil de gestion de recherche du service d'archives.

Le SIP est transformé en AIP suivant le format de métadonnées pour la conservation choisi par le service d'archives. Dans les cas les plus simples, il peut avoir la même forme et structure que le SIP mais avec enrichissement du bordereau de versement (résultat des contrôles, identification des formats, enrichissement de la description…) et la génération de fichiers issus des conversions de format des fichiers reçus, si des transformations de cette nature sont programmées.

L'AIP est ensuite ingéré dans l'infrastructure de stockage du service d'archives, avec lien avec le système de gestion documentaire.

Exemple : Des exemples de profils et fichiers au format du SEDA

- Profil pour les actes réglementaires des collectivités territoriales soumis au contrôle de légalité (par les préfectures) issus de plateformes de télétransmission: faire le lien avec le fichier profil_contrôle_legalite_exemple
- Profil pour les données issues de l'application métier et de l'outil de gestion électronique de documents gérant les prestations accordées par les maisons départementales des personnes handicapées (MDPH) dans les départements : faire le lien avec le fichier profil_MDPH_exemple

Exemple : Bordereau de versement, partie « entête » pour un versement de dossiers concernant l'attribution du revenu minimum d'insertion (RMI) dans le département du Finistère

<ArchiveTransfer
xsi:schemaLocation="fr:gouv:ae:archive:draft:standard_echange_v0.1
../../ANFontainebleau/schemas/archives_echanges_v0-1_archivetransfer.xsd">

–

<Comment>

Transfert de dossiers individuels de revenu minimum d'insertion

</Comment>

<Date format="ISO 8601">2008-06-18T00:00:00.0Z</Date>

<TransferIdentifier schemeAgencyName="Direction de l insertion">2008-00025

</TransferIdentifier>

–

<TransferringAgency>

<Identification schemeName="SIREN">org_submitter</Identification>
–
<Name>
Direction de l'insertion et de la lutte contre les exclusions
</Name>
–
<Contact>
<JobTitle>Chef de service</JobTitle>
<PersonName>M. DUPONT</PersonName>
<Responsibility/>
</Contact>
–
<Address>
<CityName>QUIMPER</CityName>
<Country>France</Country>
<LineOne>Cité administrative de Ty Nay</LineOne>
<LineTwo>BD du Finistère</LineTwo>
<Postcode>29000</Postcode>
</Address>
</TransferringAgency>
–
<ArchivalAgency>
<Identification schemeName="Identifiant de la direction des archives">org_archivist</Identification>
<Name>Archives départementales du Finistère</Name>
–
<Contact>
<JobTitle>Responsable des archives contemporaines</JobTitle>
<PersonName>Mme DUPONT</PersonName>
<Responsibility>correspondant archives électroniques</Responsibility>
</Contact>
–
<Address>
<CityName>QUIMPER</CityName>
<Country>France</Country>
<LineOne>Cité administrative de Ty Nay</LineOne>
<LineTwo>BD Finistère</LineTwo>
<Postcode>29000</Postcode>
</Address>
</ArchivalAgency>

Exemple : Bordereau issu d'une transformation par une feuille de style, pour un versement de dossiers d'étrangers : description du versement dans son ensemble

Intitulé : Délivrance de titres de séjour : dossiers individuels des étrangers clos en 1997

(décédés, naturalisés, expulsés, inactifs depuis plus de 2 ans)

Cote service d'archive: W1234

Convention: contrat-GED_etrangers

Niveau de description: Groupe de documents

Historique: En 2006, un applicatif Ged a été mis en œuvre par la préfecture des Ardennes : Gargantua, version XX. Les données de description permettant au service des étrangers de retrouver un dossier scanné ont été toutes extraites de la base nationale des ressortissants étrangers en France (AGDREF) et injectées dans Gargantua. Les dossiers papier ont été ensuite numérisés par le logiciel de gestion électronique, et transmis au service d'archive après transformation au format français d'échange des données pour l'archivage électronique, publié dans le référentiel général d'interopérabilité par le logiciel XXXX élaboré par le service informatique de la préfecture en 2008. A noter qu'il peut exister des métadonnées sans dossier numérisé associé, la préfecture ayant opté pour une reprise partielle des dossiers

Description: Fiche cerfa, pièces justificatives d'état civil et de la situation familiale et professionnelle

Dates extrêmes : 17 décembre 1976 au 30 décembre 2006

Service producteur: Préfecture des Ardennes-Direction de la réglementation et des libertés publiques

Lieu de dépôt: Archives départementales des Ardennes

Indexation:

Mot-clé: étranger

Confidentialité: La lecture des métadonnées est libre

Identifiant: code du mot étranger dans le thesaurus W

Type: Collectivité

Mot-clé: titre de sejour

Confidentialité: La lecture des métadonnées est libre

Identifiant: code du mot titre de séjour dans le thesaurus W

Type: Collectivité

Sort final: 2 ans puis Conserver

Date de départ du calcul: 18 novembre 2006

Restriction d'accès: 75 ans à compter de la date de l'acte ou de la clôture du dossier

Date de départ du calcul: 18 novembre 2006

Langue du contenu: français

Confidentialité de la description: La lecture des métadonnées est libre

Exemple : Description au niveau d'un dossier anonymisé: les éléments de description ont été récupérés automatiquement à partir des index du système d'information du producteur

Intitulé : 0000000186 L. B.

Niveau de description: Dossier Description: numero AGDREF=0000000186;nom patronymique=L.;Prenom=B.Nom marital=non renseigne;Date de naissance=1924-07-06T00:00:00;Lieu de naissance=AZAZGA;Pays de naissance=ALGERIE;Sexe=M;Nationalite=ALGERIENNE;Document actuel=Carte resident algerien;Date de debut de validite=1995-02-27T00:00:00;Date de fin de validite=2005-02-26T00:00:00;Date de delivrance=1995-07-01T00:00:00;Reference reglementaire=non reference;Statut=N;Numero de dossier=IBA0000002;Code de mouvement=non renseigne;Indicateur dossier archive=non renseigne;numero dossier reference=non renseigne;Commune d'habitation=VIVIER AU COURT;

Dates extrêmes : 01 juillet 1995 au 26 février 2005
Langue du contenu: français
Confidentialité de la description: La lecture des métadonnées est restreinte

Chapitre 8 - Conclusion sur les métadonnées

A. Conclusion

Les métadonnées jouent un rôle essentiel sur de multiples plans qu'il faut prendre en compte. Les métadonnées descriptives doivent permettre à l'utilisateur de trouver les objets numériques intéressants parmi des millions d'autres, d'évaluer leur intérêt et de restituer toute la sémantique nécessaire à leur utilisation.

L'Archive, de son côté, utilisera largement les métadonnées techniques, de structure, de droit, d'empaquetage. Les formats de métadonnées s'appuient pour l'essentiel sur les technologies XML. La difficulté principale réside et résidera dans leur processus de création qui est loin de pouvoir être toujours automatisé.

Bibliographie

[Premier ouvrage de synthèse sur l'archivage numérique en langue française.]

• BANAT-BERGER F., HUC C., DUPLOUY L., L'Archivage numérique à long terme, les débuts de la maturité? Paris, La Documentation française, 2009.

[Norme de référence essentielle pour comprendre le problème posé par l'archivage numérique]

http://public.ccsds.org/publications/archive/650x0b1(F).pdf