

RESEAU ARCHIVES INTER-ADMINISTRATIONS



COMPTE RENDU DE LA SIXIEME REUNION DU 28 NOVEMBRE 2003

Etabli par Béatrice WOZNICA

Participants

Claude AUBRIE INRA
claude.aubrie@inria.fr

Jean-Philippe BONILLI Centre Georges-Pompidou
01.44.78.48.49 / jean-philippe.bonilli@cnac-gp.fr

Patrick CAVALIÉ Ministère de l'Ecologie et du développement durable
01.42.19.18.03 / patrick.cavalié@environnement.gouv.fr

Catherine DHERENT Direction des Archives de France
01.40.27.60.00 / catherine.dherent@culture.gouv.fr

Damien FERRERO Agence de l'environnement et de la maîtrise de l'énergie
Damien.Ferrero@ademe.fr

Anne LACOURT École nationale des Ponts-et-Chaussées
01.64.15.34.69 / Anne.LACOURT@mail.enpc.fr

Henry de LANGLE Centre Georges-Pompidou
01.44.78.41.42 / henry.delangle@cnac-gp.fr

Daniel LÉLU Direction générale de l'Aviation civile
lelu.daniel@saf.dgac.fr

Aude NGUYEN Ministère des affaires sociales, du travail et de la solidarité/Ministère de la santé, de la famille et des personnes handicapées
01.40.56.48.91 / aude.nguyen@sante.gouv.fr

Vincent ROGER Ministère de l'Équipement
vincent.roger@equipement.gouv.fr

Pierre SALMERON Mairie de Bagneux
01.46.56.54.01/Pierre.Salmeron@wanadoo.fr

Béatrice WOZNICA Ministère de l'Outre-Mer
01.53.69.22.05/ b_woznica@hotmail.com

Les langages structurés, le XML et la DTD utilisée pour les instruments de recherche archivistique (EAD)

1- Introduction

La réunion a lieu au 56, rue des Francs-Bourgeois, à la Direction des Archives de France.

La réunion s'ouvre par une introduction d'Aude Nguyen :

Définition du langage : c'est l'emploi de codes gestuels, oraux ou écrits, ... structurés par une syntaxe et parfois une grammaire qui permet de donner un sens au message, donc de communiquer. Les langages informatiques ont évolué (langages structurés, objets, orientés objets) et permettent de faire dialoguer les applications, les faire inter-opérer, l'enjeu étant d'échanger de plus en plus de données dans de multiples domaines d'activités

Le rôle de l'archiviste : sa place se trouve avant le citoyen qui se situe au bout de la chaîne des procédures.

Comme les autres professionnels, il a besoin d'analyser des pré-requis fonctionnels et pas forcément informatiques, comme par exemple la mise à plat des procédures, la modélisation des données. Il se doit également de participer aux réflexions sur les formats d'échanges de données, les référentiels de données et d'applications, la sémantique du web et sa sécurité. Tout ceci nécessite un suivi, un accompagnement, une formation spécifique ; une bonne conduite de projet est essentielle.

Le langage HTML présente des limites, d'où la création d'XML, à la fois support de dialogue entre deux applications et langage de structuration de la description archivistique.

Catherine Dhérent fait ensuite une présentation plus détaillée de ces langages et de la DTD EAD

2- Les outils de description archivistique

2.1. Le rôle de l'archiviste

Avant tout, l'archiviste se doit de prendre en considération certains besoins :

- Pérenniser les instruments de recherche, qui maintiennent l'accès aux données (migration facilitées par des langages informatiques ouverts),
- Assurer leur partage et leur interrogation centralisée (en particulier pour les archives privées, dont les fonds sont souvent éclatés car versés dans des établissements différents, ce qui nécessite la création d'un seul et unique instrument de recherche), en fonction des souhaits et des intérêts de la personne productrice, ce que pourront permettre les nouveaux outils électroniques ,
- Recherche sémantique et multilinguisme.
- Fournir une aide plus efficace au public.

2.2. ISAD(G) (voir le compte rendu de la 4eme réunion RAIA du 19/11/2002)

- Norme internationale de description archivistique.
- Ses principales caractéristiques : Elle insiste sur la nécessaire structuration des données ; la description doit obéir à une hiérarchisation des informations (aller du général au particulier), identifier le fond et son producteur, retrouver la logique de sa production, s'interdire toute redondance.
 - ISAD(G) ne connaît aucune limite au nombre de niveaux de description.
 - Toutefois, cette norme induit une vision historicisante qui ne reflète pas l'organisation de l'administration.
 - Structure souple, elle laisse à l'archiviste toute latitude dans la désignation du fond, ce qui donne lieu à des interprétations diverses et variées et induit une libre évolution des notions (l'archiviste doit pouvoir disposer d'instruments souples pour ne pas rester figé dans un instrument de recherche. Selon ses capacités, il réalisera un état des fonds, un répertoire ou encore un inventaire). Les nouveaux outils informatiques peuvent même décrire un paragraphe dans une pièce d'archives.

2.3. Les outils électroniques traditionnels de description et leurs limites

- **Le traitement de texte Word** reste encore assez utilisé, cependant les documents qu'il produit doivent nécessairement être structurés, alors qu'il ne fournit pas de feuille de style. Le traitement de texte Word pose également un problème de format, de pérennisation, il ne permet pas le traitement des données ni la création partagée.

- **Le format PDF** rend possible une relative pérennisation des données ; cependant, il constitue une forme tellement immuable et peu souple qu'il n'est pas possible de le partager ni même de le faire évoluer. Il ne permet pas non plus le traitement des données.

- **Les bases de données** (sauf Arkhéia, les progiciels Avenio, Clara, Gaia, Thot reposent sur des SGBD) sont encore peu passées sur le Web (processus en cours). Elles rendent difficile l'application des principes de contextualisation et de hiérarchisation préconisés par ISAD(G) ; elles nécessitent une lourde programmation. Ces bases sont toutes différentes et multiples, leurs relations complexes, elles sont relativement rigides, ce qui explique qu'elles perdent actuellement du terrain face aux langages de balisage.

Patrick Cavalié se demande lequel de ces trois outils présente le moins d'inconvénients. En fait, tous ont leurs limites et nécessitent la présence d'une équipe informatique, mais il est avant tout très important de modéliser au maximum et se montrer extrêmement vigilant en amont avant de mettre en place l'application informatique.

Toutes ces limites des outils traditionnels expliquent qu'aujourd'hui les langages de structuration se révèlent les plus pertinents.

3- Les langages de structuration

3.1. Leurs avantages

Ces langages permettent :

- Une description à plusieurs niveaux
- D'affiner la description à volonté
- De déplacer les niveaux ou d'en ajouter

- D'utiliser et manipuler le contenu des documents et des données (le choix de l'instrument de recherche reste donc libre)
- De garantir la pérennisation des données
- D'avoir des présentations adaptées au contexte d'édition
- D'échanger au niveau international des données standardisées
- De constituer de vastes réservoirs d'information (voir par exemple le portail *Archival Resources* du RLG –<http://www.rlg.org>, l'accès est payant– qui met en ligne les guides et inventaires de millions de fonds)
- De faciliter la recherche au public
- d'avoir des points de contrôle.

3.2. Leur support technique

Les langages de structuration reposent sur un encodage de base de toute donnée électronique (= texte brut). L'association de caractères constitue un langage informatique : le standard ASCII (créé en 1965) devenu depuis UNICODE, qui prend en compte l'ensemble des langues de la planète et permet d'utiliser 64 000 caractères. Depuis deux ans, ce nouveau code ISO de langues a été mis en place, lorsqu'on a découvert qu'il existait 7 000 langues dans le monde.

Enfin, le texte se trouve enrichi par la reconnaissance des caractères placés entre chevrons. Ex. : <et>

<code> constitue une **balise** (ou mark) qui confère une structure, une grammaire, donc donne un sens à ce qu'on va écrire

3.3. Les familles de langage de balisage

- **SGML (Standard Markup Language)** est un langage de balisage pour la définition et l'utilisation de formats de documents ; il permet de réaliser tous types de documents (il est d'ailleurs largement utilisé par l'industrie, l'édition ou le milieu scientifique universitaire).

Le langage SGML permet de donner du sens à des contenus grâce aux balises, d'emboîter les balises les unes dans les autres (en structurant ainsi l'information, on peut lui conférer de la granularité). Il rend également possible une validation et un contrôle plus fins ainsi qu'une meilleure compréhension du contenu.

Cependant, SGML reste un langage lourd et complexe et n'est pas utilisé par les navigateurs web.

- En SGML, la **DTD (Document Type Definition)** est la transcription informatique d'un modèle de document, un ensemble de règles d'encodage, une grammaire qui permet de contrôler la validité d'un document. Elle est composée :

- **d'éléments** (= information essentielle), définis par des balises de début (< >) et de fin de contenu (</>). Ils ont un code donné entre les chevrons, identifié par un nom en clair : <nm> pour *nom*. Ces éléments sont décrits dans un dictionnaire, leur utilisation expliquée dans des manuels d'application. Ils peuvent contenir du texte ou d'autres éléments englobants, ou encore des liens vers des ressources électroniques (URI, PUR ou URL. NB : préférer l'URI et le PUR à l'URL qui change souvent, choisir les identifiants les plus pérennes).

Remarque : on pourrait réaliser une DTD par type de document ; il convient donc de vérifier si elle existe déjà : l'Agence pour le Développement de l'Administration électronique (ADAE) a répertorié sur son site les modèles déjà existants, à consulter (<http://www.adae.pm.gouv.fr/>).

- d'**attributs** : ils qualifient les éléments et en modifient le sens
- d'**entités : caractères spéciaux**

- **HTML** est une DTD écrite sur SGML, utilisée pour écrire des pages web. Sa structure est peu profonde : le nombre d'éléments est limité (une centaine).

Toutefois, cette application ne peut être modifiée. De plus, il s'agit d'une DTD et non d'un langage de structuration, qui ne distingue pas la forme du fond et perd les langages de balisage.

Devant les limites des langages existants est apparue la nécessité d'en créer un autre : XML.

4- XML (eXtensible Markup Language)

4.1. Définition

Langage créé en 1998, issu de SGML dont il est une version simplifiée adaptée au web. Extensible : on peut en faire ce qu'on veut. Pour plus d'infos, consulter le site <http://www.w3.org/XML>

XML permet de structurer les données, en séparant le fond de la forme. ; il a des outils pratiques (une feuille de style CSS), il possède XSLT (programme de transformation des données), permet de produire un document bien formé. L'utilisateur peut créer ses propres balises et ses propres DTD. Ce langage peut être intégré à toutes les applications existantes, et peut également être ré-injecté dans des bases de données. Des données de tout format (à condition d'être correctement écrites) peuvent être exportées et structurées en XML. Enfin, XML est indépendant des plates-formes matérielles et des logiciels.

4.2. Comment produire du XML ?

- En txt en dur (par exemple avec WordPad)
- ou - Avec un éditeur XML (par exemple Xmetal, XML-Spy)
- ou - Avec un outil libre (XMLOperator : <http://xmloperator.org>)
- ou - A partir d'une excellente feuille de style (exemple : Word) convertie en XML. (Exemple : certaines universités fournissent une feuille de style aux étudiants rédigeant leur thèse, celui étant ensuite converti en XML, puis inscrit directement dans un catalogue des thèses ; ainsi, l'étudiant pourra disposer d'un document propre qu'il pourra corriger à sa guise après sa soutenance, en vue de l'éditer par exemple).

5- Une DTD pour rédiger des instruments de recherche (répertoires et inventaires uniquement) : l'EAD (Encoded Archival Definition)

Remarque : cette DTD n'est pas appropriée pour rédiger un guide ou un bordereau de versement.

5.1. Définition

DTD écrite pour le SGML (et compatible XML) entre 1993 et 1998, issue du dialogue entre archivistes et informaticiens américains (groupe de travail de la Society of American Archivists, dont fait partie la Direction des Archives de France depuis septembre 2000). Le format MARC-AMC (Archival and Manuscripts Control) utilisé jusqu'ici, qui ne décrit que des pièces isolées, s'est révélé insuffisant quand est apparue la nécessité de structurer et contextualiser les données. C'est ainsi qu'a été créée la DTD EAD, qui bénéficie d'une très bonne maintenance (assurée depuis 1996 par le Bureau du Développement et des formats MARC de la Library of Congress, en partenariat avec la Société des Archivistes Américains).

5.2. Structure

- L'EAD est composée de 2 éléments :

- Des métadonnées et des informations sur le contenu de l'unité documentaire
- 145 éléments non obligatoires

- Les niveaux de description ont les mêmes balises :

* Haut niveau de l'EAD : 5 balises concernant les métadonnées (pour la conservation)

<ead> (balise ouvrante)

document ead :

<eadheader> (décrit l'instrument de recherche)

<frontmatter> (publication traditionnelle : intro, avant-propos...)

<archdesc> (description de l'unité documentaire)

</ead>

2 balises concernant les niveaux de description :

* Description de l'unité documentaire :

```
<archdesc>
  <did> (identification et description)
  <bioghist> (biographie, histoire de l'institution)
  <scopecontent> (présentation du contenu)
  <controllaccess> (tout ce que l'on souhaite indexer)
  <dsc> (description des composants : permet d'affiner
    la description au niveau du fond jusqu'au
    paragraphe, au choix)
  <c> (niveaux de description qui vont
    s'emboîter)
</archdesc>
```

* Identification et description de l'unité documentaire (balises les plus utilisées):

```
<did>
  <repository> (organisme responsable de l'accès
    intellectuel)
  <origination> (origine)
  <unittitle> (intitulé de l'unité documentaire)
  <unitdate> (date)
  <unitid> (identifiant)
  <physdesc> (description physique)
</did>
```

NB : Toutes les balises sont remplies.

* Éléments obligatoires : il y en a peu

```
<ead>
  <eadheader>
  <eadid> (identifiant EAD)
  <filedesc> (description du fichier)
  <titlestmt> (titre)
  <titleproper>
```

(NB : dans tout document électronique, on privilégie la métadonnée, pour la conservation, plutôt que le document lui-même)

```
<archdesc>
<did>
```

Langage peu contraignant, il permet à chaque service d'appliquer sa propre philosophie.

5.3. L'EAD à la Direction des Archives de France : travaux en cours

- Modélisation des guides de recherche
- Gestion de toute la chaîne de traitement des archives contemporaines (de la production à la description, avec à terme l'élaboration d'un inventaire automatique)
 - A ce jour, les bordereaux de versement, le Thésaurus W ainsi que les tableaux de gestion possèdent leur format XML (voir <http://ajlsm.com>)
- Sur le site internet de la DAF, on trouvera :
(<http://www.archivesdefrance.culture.gouv.fr/>, rubrique « Archivistique », puis « Description archivistique », puis « Informatisation de la description : la DTD EAD (Encoded Archival Description) »)
 - Des informations complémentaires sur l'EAD présentées par Catherine Dhérent
 - La traduction du dictionnaire américain des éléments EAD
 - Le bulletin des Archives de France consacré à la DTD EAD
 - Le référentiel des applications françaises de l'EAD
 - Des conseils pour rédiger un répertoire en EAD
 - Toutes les instructions concernant la modélisation des guides de recherche (voir rubrique « Circulaires et arrêtés :Instruction (DITN/RES/2003-001) du 17 octobre 2003sur « Guide de(s) sources et guides de recherche. Modélisation »)
- Pour faire une intéressante comparaison entre des catalogues anglais et américain, voir les sites <http://www.lib.duke.edu> (américain) et <http://catalogue.pro.gov.uk> (britannique)
- Enfin, le site <http://archive.org> offre un état des lieux sur l'archivage des sites web

Formation à l'EAD : il est possible de s'inscrire à un stage pour pratiquer l'encodage des instruments de recherche en EAD.

5.4. Conclusion

Les intérêts de la DTD EAD :

- Elle s'applique très bien à ISAD(G)
- Elle permet une intégration aisée de tous types de fichiers
- Elle combine 3 possibilités de recherche.
- Dans les bibliothèques, il sera nécessaire d'entamer une réflexion pour faciliter la recherche et la collecte des ouvrages sur un fonds structuré.

5.5 Prochaine réunion (dates, thèmes)

La prochaine réunion aura lieu en juin ou septembre 2004 et portera sur le projet de mise à plat des procédures d'archivage au ministère des affaires sociales, présenté par Jean-Pierre Brière.

Catherine Dhérent, qui prend très prochainement ses fonctions à la Bibliothèque Nationale de France, propose pour la réunion suivante (début 2005) de faire une présentation de l'installation d'un système de Records Management à la BNF.