Manuel de formation

Formation à l'encodage XML EAD

Par :

AJLSM
17 rue Vital Carles
33000 Bordeaux, France
Téléphone : +33 5 57 14 25 25 // Télécopieur : +33 5 56 44 08 47
Personne contact :
ARVERS, Jean-Luc : arvers@ajlsm.com

© AJLSM – 2012 : Diffusion interdite

Sommaire

1) Présentation 3
2) Objectifs 3
3) Introduction à XML / EAD 3
3.1 XML 3
3.2 L'EAD 4
3.2.1 Présentation de l'EAD 4
3.2.2 Structure de l'EAD 4
3.3 Caractères particuliers 5
3.3.1 Caractères réservés en XML 5
3.3.2 Caractères spéciaux 5
3.3.3 Les espaces 6
3.4 Consignes 6
3.4.1 Nommage des fichiers XML 6
3.4.2 Donner un identifiant 6
3.4.3 Normaliser 7
3.4.4 Être précis 7
3.5 Liens 8
3.5.1 Internes 8
3.5.2 Vers des ressources externes 8
3.6 Vérification 9
3.7 Problèmes courants 9
3.7.1 Encodage des caractères 9
3.8 Ajout de listes d'autorité finies personnalisées 9
4) Annexes 9

1) Présentation

Ce document sert de support documentaire pour la formation à l'encodage XML EAD destinée au personnel de la Bibliothèque municipale de Dijon en charge de l'édition de l'instrument de recherche EAD présentant les menus qui sera publié dans l'application Pleade.

Des exercices pratiques ont été réalisés au cours de la formation pour mettre en pratique les indications données par ce document.

Ce document peut être reproduit, mais à condition qu'il demeure entier et que les mentions d'auteur et de date soient bien indiquées.

2) Objectifs

Les objectifs de cette formation sont les suivants :

1. Se familiariser avec la norme et le vocabulaire XML.
   Cette norme ou recommandation n'est pas arrivée de nulle part en 1998 ; bien comprendre son origine aide à bien comprendre ses différentes utilités.
2. Se familiariser avec le format EAD
3. Connaître les avantages et les inconvénients de l'usage de XML.
4. Connaître les principaux concepts reliés aux documents XML : jeu de caractères, éléments, attributs, etc.
   Avec un vocabulaire commun, il sera plus simple de communiquer et d'interpréter les messages d'erreur renvoyés par l'éditeur XML
5. Présentation de l'éditeur XML qui va être utilisé et de sa configuration
6. Éditer l'instrument de recherche de la collection de menus de la BM de Dijon.
   Être capable d'interprétation les principaux messages d'erreurs susceptibles d'être rencontrés lors de l'édition des ressources.

3) Introduction à XML / EAD

3.1 XML

XML est un format ouvert qui :

- structure le contenu. Des programmes informatiques peuvent ensuite s'appuyer sur cette structure pour transformer le document en un autre format dédié, repérer et exploiter certains contenus, etc.
- facilite la réutilisation et l'échange d'information
- assure la pérennité de l'information documentaire
- met à disposition de l'information structurée sur les systèmes d'information d'aujourd'hui : Web / intranet
- est lisible par l'humain car il s'agit d'un format descriptif
- facilite le multilinguisme

La norme XML est présente partout et tout le monde l'utilise parfois sans le savoir : échange d'informations B2B, fils RSS, interface de logiciels, etc. Quelques exemples dans le monde documentaire :

- XHTML pour les pages Web
- DocBook pour l'édition de documentation technique
- Biblio-ML pour les catalogues bibliographiques

- TEI (Text Encoding Initiative) pour les documents relevant de la littérature, des sciences humaines et sociales
- EAD pour les documents d'archives
- Protocoles d'échange de données : SRU/SRW protocole d'interrogation à distance, OAI-PMH protocole d'échange de métadonnées

XML est donc un format pour la structuration, les échanges et un format d'archivage (car il ne dépend pas d'un logiciel).

A partir d'un document XML (donc d'un fichier électronique structuré) on peut produire divers documents destinés à la consultation, donc présentés de façon pertinente pour un usage et un type de média donné :

- fichier PDF pour l'impression
- fichier HTML pour la consultation au moyen d'un navigateur Web. C'est ce que fait Pleade : il reçoit en entrée des documents EAD et les rend consultables dans un navigateur Web.

3.2 L'EAD

L'utilisation de l'EAD pour décrire des documents d'archives permet :

- d'offrir un format normalisé pour la structuration d'instruments de recherche archivistiques
- une parfaite adéquation entre les normes de description archivistique (ISAD-G) et la représentation informatique de l'information
- d'exploiter les possibilités offertes par XML

3.2.1 Présentation de l'EAD

L'EAD (Encoded Archival Description)[1] est un vocabulaire XML qui sert à décrire des fonds tout comme les instruments de recherche traditionnels. Un document EAD est donc simplement un instrument de recherche encodé en XML et au format numérique.

Un document EAD respecte aussi des règles (norme XML, définition de type de document EAD) qui facilitent son échange et son traitement informatique. Il permet aussi d'exploiter plus finement les descriptions, par exemple au moyen d'index de navigation comme dans Pleade.

L'EAD respecte parfaitement la norme internationale de description archivistique ISAD(G). Sa première version officielle 1.0 a été diffusée par la Society of American Archivists en septembre 1998 puis une version "2002" a été publiée par la suite.

En France, c'est la Direction des Archives de France qui assure la promotion de la DTD EAD, dans le cadre de sa politique visant à favoriser l'émergence et le développement de systèmes d'information utilisant des formats d'échange et des protocoles reposant sur des standards, normes et recommandations internationales. Ces pratiques garantissent, en effet, l'interopérabilité et la pérennité des données.

3.2.2 Structure de l'EAD

L'EAD définit la structure de l'instrument de recherche mais pas son contenu.

La grammaire EAD est composée de près de 145 éléments non obligatoires (seuls certains le sont) et dont l'ordre est très souple. De plus, tous les éléments descriptifs utilisables au plus haut niveau sont également utilisables à chaque niveau subordonné.

On peut distinguer deux parties principales :

- l'en-tête `<eadheader>` que l'on peut assimiler à la description bibliographique de l'instrument de recherche
- la description archivistique du fond `<archdesc>`, contenant la description des sous composants `<dsc>`

[1] EAD : http://www.loc.gov/ead/ead.html

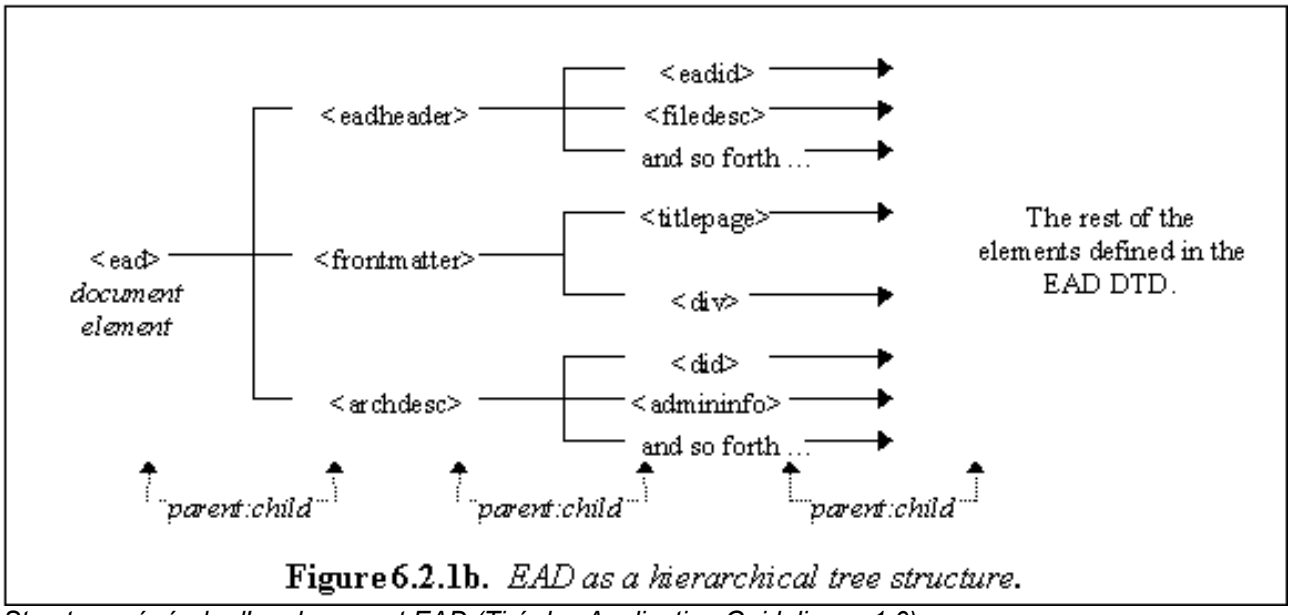


Figure 6.2.1b. *EAD as a hierarchical tree structure.*

Structure générale d'un document EAD (Tiré des Application Guidelines v1.0)

Viennent ensuite des niveaux hiérarchiques de description pour chaque composant archivistique. Pour ces niveaux, ce sont les mêmes balises qui peuvent être utilisées.

3.3 Caractères particuliers

3.3.1 Caractères réservés en XML

Certains caractères qui peuvent être utilisés dans le texte sont significatifs dans la norme XML. Ils ne peuvent donc pas être écrits dans le texte sous leur forme habituelle. Pour pallier ce problème, la norme XML recourt à des entités.

Les entités sont des substituts pour des séquences d'information ou de caractères. Une entité commence par le caractère `&` (ce caractère est donc significatif en XML) et se termine par un point virgule `;`.

Il existe 5 entités prédéfinies en XML qui permettent l'utilisation de caractères réservés au balisage dans les contenus. Les voici avec leur signification :

Entité	*Caractère*	*Description*	*Barre d'outils XMetal*
`&`	&	Esperluette ou ET commercial.	
`>`	>	Signe Supérieur à.	Symboles
`<`	<	Signe Inférieur à.	Symboles
`'`	'	Apostrophe	
`"`	"	Double apostrophe	

3.3.2 Caractères spéciaux

Pour écrire un caractère qui ne fait pas partie du jeu de caractères utilisé (par défaut UTF-8) ou que le clavier ne permet pas de saisir, on peut utiliser :

- les barres d'outils Symboles ou Caractères spéciaux de XMetal, disponibles dans les . Si le caractère que l'on souhaite écrire n'est pas présents dans ces barres d'outils, il faut passer par la méthode ci-dessous
- les entités caractères

En effet, dans un fichier XML, les caractères peuvent être directement encodés en Unicode. Il est donc possible d'inscrire tout caractère Unicode à l'aide des entités caractères, qui prennent la forme suivante :

```
&#xa9;
```

Dans ce cas, il s'agit du caractère qui occupe la position A9 en base hexadécimale (169 en base décimale) dans le jeu de caractères Unicode. Il s'agit de ce caractère ©. Par ailleurs, on peut aussi utiliser cette notation :

```
&#169;
```

Il s'agit exactement du même caractère, mais cette fois indiqué en base décimale. C'est le '`x`' qui suit le '`#`' qui indique la base hexadécimale.

Pour trouver le numéro d'un caractère, vous pouvez utiliser la table de caractère de Windows ou la fonctionnalité insertion de caractères spéciaux de votre traitement de texte.

Pour vérifier que les caractères spéciaux sont bien interprétés, vous pouvez regarder votre document dans les vues Normal ou Aperçu de la page ou dans un navigateur Web.

3.3.3 Les espaces

Bien qu'un document XML EAD puisse parfaitement valide selon la DTD, il se peut que des problèmes d'espace ou autres perturbent son affichage, par exemple dans un navigateur. L'encodage d'informations à l'intérieur d'une phrase (balise en ligne), par exemple pour souligner que telle partie de texte est une date ou un nom de personne, est sujet à ce genre d'erreurs. Les erreurs les plus communes sont les absences d'espace ou les espaces au mauvais endroit.

```
Correspondance1934
```

qui correspondrait à l'encodage suivant

```
Correspondence<date>1934</date>
ou
Bill T. Jones , 1899-1976
```

qui pourrait correspondre à l'encodage suivant

```
<persname normal="Jones, Bill T.">Bill T. Jones</persname> , <date normal="1899/1976">1899-
1976</date></unittitle>
```

C'est pourquoi il est recommandé d'effectuer le balisage en ligne dans un deuxième temps.

En XML, les balises ne définissent en rien comment doivent s'afficher les documents, elles ne font que structurer le document.

3.4 Consignes

3.4.1 Nommage des fichiers XML

Il est important de maintenir un système pour créer un nom unique pour chaque document XML EAD. Il est recommandé que les noms de fichier se compose de lettres et de chiffres sans espaces ni de caractères spéciaux. Ils doivent avoir pour extension. `.xml`.

L'usage est en général que le nom du fichiers XML commencent par `FR{identifiant de l'institution productrice}_`, soit pour la BM de Dijon `FR212316101_`.

3.4.2 Donner un identifiant

L'élément `<eadid>` est obligatoire en EAD. Son contenu correspond à l'identifiant de l'instrument de recherche. Chaque IR doit avoir un identifiant unique : unique chez son producteur mais aussi unique de manière plus large surtout dans un contexte de diffusion. C'est pourquoi pour définir cet identifiant on se base en général sur l'identifiant du producteur de l'IR. Donc pour la BM de Dijon, l'identifiant commencerait par `FR212316101_`.

Pleade se base sur l'élément `<eadid>` pour attribuer à chaque document publié un identifiant unique qui lui permet de distinguer un document d'un autre. Pour Pleade, deux documents ne sont qu'un seul est même document uniquement s'ils ont le même identifiant.

3.4.3 Normaliser

L'EAD définit dans sa grammaire un certain nombre de normes :

- iso639-2b : pour les codes de langues. Le code pour le français est `fre`.
  Ce code est utilisé comme valeur de l'attribut `@langcode` de l'élément `<language>`.
- iso8601 : pour les formats de dates.
- iso3166-1 : pour les codes de pays. Le code pour la France est `FR`.
  Ce code est utilisé pour renseigner l'attribut `@countrycode` de l'élément `<eadid>`. Cet attribut doit être renseigné pour rendre l'élément `<eadid>` compatible avec l'élément 3.1.1 de l'ISAD(G).
- iso15511 : pour le code ISIL (Identificateur international normalisé pour les bibliothèques et organisations liées) correspondant à l'organisme responsable de l'accès intellectuel. Pour la BM de Dijon, ce code est `FR-212316101`.
  Cette information est utilisée dans le document EAD pour l'attribut `@mainagencycode` de l'élément `<eadid>` ; elle indique le code de l'institution responsable de l'instrument de recherche (et qui n'est pas forcément l'institution qui conserve les unités documentaires décrites). Cet attribut doit être renseigné pour rendre l'élément `<eadid>` compatible avec l'élément 3.1.1 de l'ISAD(G).
  L'attribut `@repositorycode` de l'élément `<unitid>` reprend aussi ce code.
- Iso15924 : pour les codes de noms d'écriture. Le code pour les langues latines est `Latn` ou `215`.
  Ce code est utilisé pour renseigner l'attribut `@scriptcode` de l'élément `<language>`.

De même l'attribut `@normal` permet de donner la forme normalisée d'une chaîne. Par exemple :

```
<date normal="1899/1976">1899-1976</date>
<persname normal="Hugo, Victor">Victor Hugo</persname>
```

Pleade utilise les valeurs de l'attribut `@normal` quand elles existent pour la recherche et les index.

Dans tous les cas il est essentiel de :

- normaliser les dates au format ISO 8601

```
<unitdate normal="1791-02-01/1801-06-16">1er février 1791 - 27 prairial an IX</unitdate>
```

- normaliser certains éléments d'indexation, par exemple les noms de personnes suivant la norme AFNOR NF Z 44-061

```
<persname normal="La Fontaine, Jean de">Jean de La Fontaine</persname>
```

- élaborer des listes d'autorité pour contrôler le vocabulaire
  Parmi les descripteurs génériques les types de documents, les périodes, les actions et objets, les sujets, les concepts facilement compréhensibles par le public, pourraient faire l'objet de vocabulaires contrôlés. D'autant plus que l'EAD comprend des attributs permettant de préciser qu'un terme est issu de tel ou tel vocabulaire

```
<subject source="liste-periodes">Révolution française de 1789</subject> (dans ce cas répartir dans
Pleade le contenu des éléments <subject> dans plusieurs champs d'indexation ?)
<corpname source="fichier-autorites-BnF" authfilenumber="12345" normal="France. Sénat (1958-
….)">Sénat</corpname>
```

Pleade peut exploiter les attributs `@source`, `@role` pour constituer des index de recherche ou de navigation.

3.4.4 Être précis

Dans le cadre de l'encodage, il faut être le plus précis possible :

- utiliser les bons éléments XML. Le XML étant un format descriptif dans lequel le nom des balises est significatif, il convient de choisir le bon élément pour le bon contenu. Il en est de même pour l'EAD. De nombreuses institutions ont établies des recommandations pour l'encodage des documents EAD et l'usage des éléments

- encoder toute l'information susceptible d'être exploité. En EAD, cela signifie utiliser les possibilités de balisage en ligne. Par exemple, baliser les noms de personnes, les lieux géographiques, les dates, etc indiqués dans un contenu de type rédactionnel
- utiliser les attributs pour préciser l'information. Par exemple, pour préciser le rôle ou la fonction d'une personne

Mieux vaut plus de balises que pas assez. Quand le balisage est là, il peut ensuite ou non être exploité. Pleade peut exploiter tous ces niveaux de balisage.

3.5 Liens

Dans un document XML/EAD, plusieurs types de liens peuvent être créés. Plusieurs éléments existent d'ailleurs pour encoder ces liens. Ce document décrit les types de liens les plus courants.

3.5.1 Internes

Un lien interne est un lien d'un élément du document EAD vers un autre élément du même document. Pour créer un lien vers un autre élément, il faut que l'élément cible ait un identifiant. Cet identifiant est défini en utilisant l'attribut `@id`. La cible du lien, valeur de l'attribut `@target`, est donc l'identifiant de l'élément cible.

```
Le code définissant le lien
<ref target="c546">lien sur du texte</ref>

Quelque part dans le document, il faut un élément avec l'identifiant, par exemple
<c id="c546">
```

3.5.2 Vers des ressources externes

Pour que les documents numériques ou les documents associés puissent être correctement gérés puis affichés dans Pleade, il faut que les liens depuis le document EAD vers ces ressources externes (documents PDF associés, images, documents multimédias, etc.) doivent être correctement renseignés. Cela signifie que Pleade puisse accéder au fichier indiqué dans le lien.

Nous ne traitons dans cette partie que des ressources externes qui seront stockés sur des serveurs accessibles à Pleade et pour lesquelles les liens sont relatifs, c'est-à-dire qu'ils ne commencent pas par un protocole tel http://. Voilà quelques exemples de liens relatifs :

```
<daoloc href="DAFANCH01_NC120001820_V.JPG"/>
<daoloc href="ATLAS-8485/atlas8485-03.jpg">
<daoloc href="ATLAS-8485/"></daoloc>
<extref href="../../ead/pdf/FR0001.pdf"/>
```

Un lien relatif est une méthode qui permet d'écrire le chemin permettant de se rendre d'une ressource à l'autre en parcourant une arborescence de dossiers et fichiers. Il s'agit donc d'indiquer dans le lien comment aller du fichier dans lequel est écrit ce lien au fichier cible (destination du lien).

Écrire un lien relatif utilise les trois « commandes » suivantes :

- remonter dans le dossier parent : s'écrit `..`
- ouvre un dossier (ou rentrer dans un dossier) : s'écrit `/`
- ouvre un fichier : le nom du fichier

Ainsi, si nous reprenons le dernier exemple cité, il peut se lire ainsi : à partir du répertoire dans lequel est stocké le document EAD, il faut remonter de deux répertoires ou niveaux puis ouvrir les répertoires `ead` puis `pdf` dans lequel on trouvera le fichier `FR0001.pdf`.

Dans Pleade, tous les liens vers ces ressources externes doivent donc être relatifs mais ils peuvent être relatifs :

- au document dans lequel est écrit le lien. Valable pour tous les types de ressources externes.

- à un répertoire de stockage de ces ressources externes, dont l'emplacement a été renseigné dans un paramètre. S'applique pour les images, les séries d'images et les documents PDF.

Quelques exemples de liens vers des ressources externes autre que des images numérisées :

- Lien vers un autre IR publié dans Pleade

```
<otherfindaid><p>Descriptions détaillées dans le <archref
href="autredocumentEADdansPleade.xml">répertoire numérique</archref></p></otherfindaid>
```

- Lien vers des fichiers d'IR non intégrés à Pleade et dans d'autres formats (PDF, HTML, etc) (`<extref>` ou `<archref>`)

```
<extref href="ressources-externes">voir la ressource externe</extref>
<extref href="http://ajlsm.com">AJLSM</extref>
<archref href="http://demo.pleade.com/ead.html?id=frajlsm-pleade-test-01-fr">un IR en
ligne</archref>
```

3.6 Vérification

Pour vérifier facilement et rapidement qu'un document XML que vous venez de créer est valide et qu'il peut être publié dans Pleade, vous pouvez l'ouvrir dans un navigateur Web.

3.7 Problèmes courants

3.7.1 Encodage des caractères

L'encodage utilisé sur des postes Windows est le Cp1252. Il s'agit d'un encodage spécifiques à Windows qui possède des caractères très particuliers sans équivalents dans d'autres encodages et notamment ceux utilisés pour éditer des documents XML.

Du coup, lors de copier-coller de texte depuis un document Word vers un document XML, il se peut que certains caractères ne s'affichent pas correctement ou que la validation du document génère une erreur. Par expérience, nous évitons ces copier-coller.

C'est pourquoi il est préférable d'effectuer les copier-coller à partir d'un document au format TXT.

3.8 Ajout de listes d'autorité finies personnalisées

La version 2002 de la DTD a supprimé les listes pré-définies de valeurs pour certains attributs au profit de valeurs simples (chaînes de caractères sans espace de type code). Cependant, EAD 2002 permet malgré tout de définir des listes personnalisées de valeurs dans un fichier séparé `eadlocal.ent` (à ajouter à la DTD de base).

Ces valeurs définies seraient alors utilisées au cours des processus d'édition et d'analyse d'un document EAD, utilisant cette grammaire personnalisée.

Cette possibilité permet de s'assurer de la cohérence des valeurs entre les différentes personnes susceptibles d'encoder les documents XML EAD et d'éviter les fautes de frappe (A l'édition, on choisit la bonne valeur dans une liste).

4) Annexes

Quelques ressources utiles pour l'encodage :

[en] EAD Practices Working Group (2008). « EAD Best Practices at the Library of Congress », sur le site Library of Congress. Consulté le 15 octobre 2009.
http://www.loc.gov/rr/ead/lcp/index.html

« Informatisation de la description : la DTD EAD (Encoded Archival Description) », sur le site *Archives de France*. Consulté le 15 octobre 2009.
http://www.archivesdefrance.culture.gouv.fr/gerer/classement »/normes-outils/ead/

« Dictionnaire des éléments de l'EAD, version 2002. Traduction française de la version anglaise réalisée par le groupe AFNOR CG46/CN357/GE3 », sur le site *Archives de France*. Consulté le 15 octobre 2009. (fichier

pdf, 266 pages)
http://www.archivesdefrance.culture.gouv.fr/static/1066

Groupe d'experts AFNOR CG46/CN357/GE3 (juin 2005). « Faire un répertoire ou un inventaire simple avec l'EAD », sur le site *Archives de France*. Consulté le 15 octobre 2009. (fichier pdf, 69 pages)
http://www.archivesdefrance.culture.gouv.fr/static/1068

Guide des bonnes pratiques de l'EAD en bibliothèque [site Web]. Consulté le 15 octobre 2009.
http://bonnespratiques-ead.net/