

NOTE DITN/RES/2002/001 du 21 juin 2002

Description archivistique informatisée, XML et la DTD (EAD)

Madame Martine de Boisdeffre
Directrice des Archives de France

à
Mesdames et messieurs les présidents des
Conseils généraux (Archives départementales)
Mesdames et messieurs les maires (Archives
communales)
Mesdames et messieurs les présidents des
Conseils régionaux (Archives régionale)

Le besoin de normalisation

Le monde des Archives a compris tardivement l'intérêt de la normalisation, nationale et internationale, pour ses instruments de recherche : le thesaurus W a été produit en 1987, mis réellement en service en 1989, et le premier standard de description international, ISAD (G), n'a vu le jour qu'en 1994, alors que les formats de description bibliographique ont plus de 25 ans.

Aujourd'hui, la plupart des archivistes reconnaissent les bienfaits de la normalisation même s'ils restent sensibles aux contraintes et changements dans les pratiques qu'elle induit. Car dans un contexte de mise en réseau qui s'est généralisé depuis 1997, le public en tirera des bénéfices importants : il disposera d'instruments de recherche plus simples, de contenus plus homogènes et cohérents, mieux structurés, à la mise en forme plus claire et obtiendra pour une seule requête des réponses signalant des ressources localisées dans plusieurs départements ou communes françaises et dans plusieurs services étrangers, sans avoir à naviguer d'un site à l'autre.

Dans l'intérêt général, il faut donc 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 garantissant leur interopérabilité et assurant leur évolution future.

Les collectivités locales sont certes libres de mettre en œuvre les outils informatiques de leur choix. Il ne s'agit donc pas pour la direction des Archives de France de leur imposer des solutions en matière de logiciels ou de systèmes.

Le code général des collectivités territoriales attribue cependant à l'État un contrôle scientifique et technique sur les archives des collectivités territoriales visant à "la qualité scientifique et technique des instruments de recherche, la compatibilité des systèmes de traitement et la mise en valeur du patrimoine archivistique" (art. R 1421-1). C'est à ce titre que, dans le domaine de la description des fonds, les instruments de recherche produits par les services d'archives territoriaux sont soumis au visa de la direction des Archives de France, qu'ils soient sous forme papier ou sous forme électronique.

C'est dans le cadre de ce contrôle scientifique sur les procédures et les mises en forme des contenus, que la direction des Archives de France a décidé de formuler les quelques remarques qui suivent concernant les langages d'encodage informatique, afin de parvenir dans les meilleurs délais à une harmonisation des pratiques et, à moyen terme, à la mise en œuvre d'un portail qui guiderait le lecteur vers les ressources spécifiques de chaque service, vers leurs sites web notamment. Ce portail ne serait pas la centralisation sur un même serveur de toutes les données mais une porte plus large d'accès à ces données éparses. Il permettrait ainsi aux sites spécifiques d'avoir davantage de visiteurs, en jouant pour eux un rôle de labellisation et en faisant découvrir au visiteur une information qu'il n'aurait peut-être pas cherchée ou trouvée de lui-même.

Pour parvenir à ces fins, il faut veiller à utiliser les langages et les formats informatiques les plus ouverts et les plus standardisés possible. De nombreux services ont utilisé ou utilisent encore des outils propriétaires qui se sont révélés parfois totalement irrécupérables ou le sont difficilement dans un nouvel outil ou un nouveau système. Aujourd'hui, sans préjuger de l'avenir, dans le domaine de l'encodage des données, le XML semble le meilleur candidat.

Le XML

Le XML (eXtensible Markup Language) est un méta-langage de balisage permettant une modélisation informatique et une profonde structuration des données, leur diffusion sur le web et leur préservation à long terme. Il est indépendant de toute plate-forme matérielle et logicielle. Recommandation promulguée par le W3C (Consortium du web) en 1998, il peut être couplé à un langage de programmation pour la transformation (traitement, ordonnancement, conversion en HTML ou en PDF, mise en forme), dénommé XSL-T. Un même document XML ou deux documents XML de type équivalent, peuvent être mis à disposition sous plusieurs formes, pour plusieurs profils d'utilisateurs différents. On peut par exemple, dans plusieurs services, utiliser le même modèle XML pour des inventaires d'archives, et adopter des présentations et chartes graphiques différentes dans chacun de ces services. Cela ne perturbe pas la compréhension et l'échange des données et conserve à chaque producteur son image de marque, son label et son identité.

Le XML peut permettre l'encodage de tout type d'information textuelle et est ainsi aujourd'hui très utilisé pour le commerce électronique par exemple. Cela signifie que toute la gestion d'un service d'archives peut être écrite en XML. Cependant il n'en existe pas encore d'exemple pour la gestion matérielle des archives car l'usage du XML est particulièrement justifié lorsqu'il y a un besoin d'échanges. Or il est peu de données de gestion matérielle qui nécessitent un partage entre plusieurs services d'archives différents. En revanche, il en est autrement pour la description des documents.

On peut écrire des documents directement en XML ou avec des DTD (Document Type Definition) ou des Schémas pour des types de documents récurrents et modélisés. Il existe des DTD pour de nombreux types documentaires, par exemple pour le catalogage bibliographique. Pour certains types d'instruments de recherche archivistiques, il existe une DTD : l'EAD (Encoded Archival Description).

L'EAD

L'EAD, propriété de la Society of American Archivists, maintenue par la Library of Congress, a été testée à partir de 1998 par la direction des Archives de France, sur l'inventaire de la série M des Pyrénées-Orientales. Elle s'est avérée parfaitement compatible avec les pratiques archivistiques françaises et avec le standard de description international ISAD (G). Grâce au méta-langage XML, on peut en effet profondément structurer l'information et récupérer automatiquement les données du niveau supérieur au niveau inférieur. L'EAD permet de décrire avec ses 145 éléments et de façon non contraignante, étant donné que la plupart des éléments sont facultatifs, le nombre de niveaux hiérarchiques que l'on souhaite, en allant du fonds à la pièce. Elle permet l'intégration de fichiers numériques d'autres formats, par exemple de l'image numérisée du document. Cette structuration facilite la recherche dans un instrument de recherche ou dans plusieurs à la fois, soit sur des termes indexés, soit sur du texte libre, soit par les sommaires dans lesquels se déploient les

niveaux de description.

La DTD EAD est un bon outil pour les répertoires et les inventaires. Elle sera testée cette année pour les bordereaux de versements. En revanche, elle ne peut être utilisée pour d'autres types d'instruments de recherche : la direction des Archives de France va donc commencer à travailler à la modélisation des guides, de type thématique dans un premier temps.

Contrairement à d'autres domaines, il n'existe aujourd'hui, par chance pour le monde archivistique, qu'une seule DTD dont l'usage généralisé assurerait une première vraie cohérence et une véritable interopérabilité entre tous les répertoires et inventaires, une meilleure compréhension de leurs contenus pour le public et de puissantes combinaisons de recherche (par exemple pour des fonds épars entre plusieurs institutions).

Aujourd'hui, plusieurs instruments français sous XML en EAD sont mis à la disposition du public sur le web, parmi les ressources du Research Libraries Group (dans Archival Resources) ; ils le seront aussi progressivement sur les sites des institutions françaises concernées. Ce sont :

- l'état général des fonds des archives publiques du Centre historique des Archives nationales (encodage en cours),
- l'état général des fonds des archives privées du Centre historique des Archives nationales et du Centre des archives du monde du travail,
- l'inventaire du fonds Napoléon (Centre historique des Archives nationales), en ligne sur le site web du CHAN,
- l'inventaire du fonds de l'abbaye de Clairvaux (Archives départementales de l'Aube),
- le fonds Sarraut (Archives départementales de l'Aude).

D'autres institutions patrimoniales que les archives ont expérimenté et utilisent la DTD EAD : c'est le cas notamment de la Bibliothèque nationale de France pour le catalogue des manuscrits des Nouvelles acquisitions françaises et de la direction du Livre et de la Lecture pour la reprise des 104 volumes du catalogue imprimé des manuscrits des bibliothèques publiques de France. A l'étranger, les utilisations sont nombreuses. On en aura une idée en consultant le site officiel de l'EAD. Une idée des réalisations françaises sera donnée lors d'une journée d'information que la direction des Archives de France organise avec la direction du Livre et de la Lecture et avec la Bibliothèque nationale de France en octobre 2002. Vous serez informés dans les semaines qui viennent des modalités et date de cette journée.

La direction des Archives de France a aussi mis en place une chaîne de production XML pour la Base d'Orientation et de Recherche dans les Archives (BORA). Cette base sera en ligne prochainement, sur la plate-forme SDX de la direction des Archives de France.

BORA sera donc à terme un portail archivistique français couvrant tous les fonds d'archives. A la couche supérieure de description des fonds pourront être liés tous types d'instruments de recherche, dont des instruments encodés en EAD. Les liens seront faits directement sur les pages web des sites concernés. Dans les cas où les services ne disposent pas encore de sites web, la direction des Archives de France pourra envisager l'hébergement de leurs instruments de recherche numériques sur ses serveurs.

Pour les services d'archives qui ont déjà informatisé la description des fonds

Les progiciels archivistiques sont constitués de modules dont l'un sert en général à la description des fonds. Ce module tient compte en principe du standard ISAD (G) mais avec des adaptations très diversifiées et sur des systèmes technologiques eux aussi divers. Cette diversité des formats, des logiciels et progiciels propriétaires, des bases de données et des systèmes documentaires peut limiter les échanges entre les

partenaires potentiels que sont les services d'archives, obscurcir le paysage documentaire pour les utilisateurs et parfois même compromettre la conservation des données numériques.

Les concepteurs de certains progiciels d'archives envisagent donc d'évoluer vers le XML et ont avancé dans cette voie. Pour d'autres, la conversion vers XML et vers la structure des instruments de recherche en EAD peut s'avérer plus lourde et difficile. Il ne s'agit donc pas de faire brutalement table rase du passé mais de prendre certaines précautions et de faire évoluer peu à peu, dans la mesure du possible, les systèmes existants vers plus de compatibilité et de normalisation.

Pour les services qui utilisent déjà les modules de description d'un progiciel du marché, la plupart construits sur un système de gestion de base de données, il n'est pas besoin de remettre en cause leurs pratiques.

XML et l'EAD peuvent être utilisés en sur-couche dans laquelle seront exportées les données : grâce à cet interface en XML, elles seront ainsi plus aisément accessibles, par exemple dans un portail national. Toutes les données de bases bien structurées peuvent être aisément récupérées dans un encodage XML. Mais il est indispensable qu'elles respectent les préconisations d'ISAD (G), c'est-à-dire une structuration hiérarchique très logique et une description du général au particulier.

Ces services seront aussi attentifs au fait que, BORA devant passer sous XML, ils pourront à terme se faciliter la tâche si un export automatique des données descriptives des fonds est possible de leurs systèmes respectifs vers BORA et évite les ressaisies.

Pour les services d'archives qui n'ont pas encore informatisé la description des fonds

Dans un petit service, la description archivistique est le domaine le plus important à informatiser car il est le plus utile aux chercheurs, notamment pour la mise à disposition sur les réseaux.

Aujourd'hui la DTD EAD étant jugée parfaitement compatible avec les pratiques françaises, la direction des Archives de France recommande aux services qui n'ont pas encore informatisé l'analyse documentaire et la production des instruments de recherche, d'utiliser cette DTD EAD directement.

Ils pourront soit sous-traiter l'encodage et la mise à disposition des données à partir de fichiers texte bien structurés, soit les encoder en interne. Pour cela, ils pourront acquérir un éditeur XML du marché. Ils pourront créer ou se faire créer des feuilles de style pour rendre la saisie encore plus ergonomique et instinctive. Le Centre historique des Archives nationales a fait créer une feuille de style pour faciliter l'encodage en interne.

Il faudra définir également le(s) mode(s) d'affichage souhaité(s) sur les réseaux et faire créer des feuilles de style appropriés. On peut à partir du même contenu de données, fournir des présentations très diverses au visiteur, en allant du type formulaire au type édition traditionnelle.

Les services qui le préfèrent pourront aussi attendre la mise en ligne, en 2003, de la plate-forme de production XML, PROXIMUS, que le ministère de la Culture et de la Communication a décidé, en novembre 2001, de faire développer. Cette plate-forme sera composée de logiciels libres et tous les outils pourront être téléchargés. Elle est actuellement au stade de la maquette. Les outils pour les instruments de recherche archivistique, et en particulier la DTD EAD, seront sur cette plate-forme.

En revanche, un service plus important a besoin d'informatiser l'ensemble de sa gestion. Pour cela, la direction des Archives de France conseille d'éviter de faire programmer de nouveaux outils informatique de gestion archivistique. On peut considérer que les progiciels sont aujourd'hui suffisamment nombreux et correspondent aux besoins. Mieux vaut perdre un peu de spécificité que d'engager une collectivité dans des investissements et programmations lourds. Plus la diversité est grande, plus risquée est la survie des applications et leur maintenance, plus complexe est la démarche du public, plus compromis sont l'échange et la facilité de communication sur les réseaux.

Il est tout à fait envisageable de gérer les aspects matériels d'un service avec les modules d'un progiciel existant et la description des fonds avec la DTD EAD sous XML. L'élément de lien entre les divers systèmes est la cote. Or celle-ci est identifiée dans la DTD EAD par une balise <unitid> qui peut se trouver à tous les niveaux de description. Il est donc tout à fait possible d'envisager des relations de cette balise et de son contenu avec les tables des modules de gestion matérielle des progiciels archivistiques qui comportent cet identifiant cote. Si on fait ce choix, il faudra demander dans le cahier des charges la programmation de concordances entre les documents en EAD et leur gestion matérielle, par les identifiants (cotes).

Remarques complémentaires

Toute la documentation EAD est encore en anglais, ce qui peut être un obstacle au développement de cette DTD. La direction des Archives de France a donc souhaité mettre à disposition des versions françaises de cette documentation. Un groupe de travail a été constitué à l'AFNOR avec des représentants de services d'archives, de bibliothèques, de musées, du CNRS, avec des consultants. Il est présidé par la direction des Archives de France. La traduction du Dictionnaire des éléments ou balises de l'EAD sera disponible dans le courant de l'année 2002. Un manuel d'application sera ensuite élaboré.

La direction des Archives de France propose aussi, pour la première fois en juillet 2002, un stage pratique à l'EAD (avec travaux pratiques) dans son programme de formation.

La Directrice des Archives de France

Martine de BOISDEFFRE

Annexes

Site officiel de l'EAD (avec la DTD à télécharger, la Tag Library en ligne) : <http://www.lcweb.loc.gov/ead/>

Ressources sur l'EAD sur le site de la direction des Archives de France :

<http://www.archivesdefrance.culture.gouv.fr/fr/archivistique/DAFlangage.html>

Le Dictionnaire des éléments ou balises en version française y sera mis ligne dans le courant de l'année 2002.

Pour obtenir des noms de sociétés pouvant encoder des instruments de recherche en EAD :
s'adresser à Catherine Dhérent, catherine.dherent@culture.gouv.fr

Stages pratiques sur l'EAD :

La Direction des Archives de France propose un stage les 8 et 9 juillet 2002 à Paris, salle informatique (présentation de SGML/XML, présentation de l'EAD, exercices pratiques). Inscription sur bulletin inséré dans la brochure Formations 2002 à demander au bureau des Formations de la direction des Archives de France, 01 40 27 61 92 (Sylvain Facompré) ou 61 98 (Jean-Pierre Dubois).

Le Centre historique des Archives nationales organise aussi tous les deux mois environ un stage de deux jours "XML et EAD : découverte et exercices pratiques" ouvert à tous les archivistes. Se renseigner auprès de Florence Clavaud (01 40 27 62 84 ; m^él florence.clavaud@culture.gouv.fr)

Quelques ouvrages en français sur XML :

HAROLD Elliotte Rusty et MEANS W. Scott, *XML in a nutshell, Manuel de référence*, O'Reilly, 530 p., 2001.

MICHARD Alain, *XML, langage et applications*, Eyrolles, 499 p., 2000.