Ekaterina NOVOKHATKO

LE CATALOGUAGE DES MANUSCRITS: DIALOGUE DES FORMATS

De nos jours, les campagnes de numérisation du patrimoine deviennent de plus en plus fréquentes, et différents points sont à considérer: comment traiter au mieux un fond historique concret pour garder son authenticité et améliorer son interopérabilité? Certes, l'hétérogénéité des formats permet de présenter sous différents formats numériques les métadonnées de sources historiques, mais cette diversité incite toujours à créer des modèles multiples pour leur traitement. La question d'une normalisation du format bien réfléchie se pose, afin de pouvoir communiquer entre les différentes institutions scientifiques de manière globale et de pousser donc à une meilleure interopérabilité des données.

À l'exemple des fonds anciens de la bibliothèque d'agglomération de Saint-Omer, contenant les manuscrits médiévaux de l'abbaye de Saint-Bertin, cet article met en avant la réflexion sur les solutions proposées pour la conversion des formats[1]. La problématique principale concerne la conversion des métadonnées dans deux formats différents en gardant leurs spécificités, et en faisant attention à l'encodage des notices et à leur structuration. Le sujet présenté sera développé en trois étapes: la contextualisation du projet Saint-Bertin; l'évocation de certaines notions sur les formats numériques; les extraits du schéma de conversion illustrant les différents objectifs des formats, ce qui a finalement constitué un défi numérique pour la conversion du catalogue de Saint-Bertin.

1. Le projet Saint-Bertin

Le projet »Saint-Bertin, centre culturel du VII^{e} au XVIII^{e} siècle : constitution, conservation, diffusion, utilisation du savoir« est mené depuis 2013 par l'Institut de recherche et d'histoire des textes (IRHT-CNRS), l'Équipex Biblissima, la bibliothèque d'agglomération de Saint-Omer (BASO) et la bibliothèque de Boulogne-sur-Mer (BSM). Ce projet fait partie des maints projets numériques français de nos jours dont le but est la reconstitution des bibliothèques médiévales; en

[1] Cette conversion a été réalisée au cours d'un stage au sein de la section de la paléographie latine de l'IRHT, qui s'est inscrit dans le projet »Saint-Bertin, centre culturel du VII^{e} au XVIII^{e} siècle: constitution, conservation, diffusion, utilisation du savoir« (http://saint-bertin.irht.cnrs.fr/).

effet, l'abbaye Saint-Bertin à Saint-Omer et sa bibliothèque de manuscrits sont une ressource de premier ordre pour l'histoire intellectuelle européenne[2].

Le but de convertir les données numériques du format TEI vers EAD s'inscrit dans le programme du projet »Saint-Bertin«, visant à numériser les manuscrits médiévaux provenant de l'abbaye; enrichir les métadonnées disponibles; éditer les catalogues médiévaux et modernes de l'abbaye de Saint-Bertin; identifier de nouvelles sources sur l'histoire des bibliothèques; analyser les écritures, les aspects codicologiques ainsi que les marques d'usage et de provenances. Les images produites lors de la campagne de photographie numérique ainsi que les notices des manuscrits encodées dans les formats TEI et EAD ont vocation à être intégrées dans les bibliothèques numériques françaises et les bases de données[3].

En parlant de la conversion des données numériques des manuscrits conservés à Saint-Omer, on sous-entend qu'elles représentent les métadonnées, c'est-à-dire des données servant à définir ou décrire les autres données quel que soit leur support, papier ou électronique. Plus précisément, il s'agit d'un ensemble structuré d'informations décrivant une ressource quelconque[4]. Il existe nombre de formats de métadonnées, dépendants des domaines ou filières thématiques concernés, ainsi que des objectifs donnés à l'encodage.

Dans le cadre du projet Saint-Bertin, on peut considérer comme métadonnées le catalogue des manuscrits de la BASO donnant la description précise des manuscrits; les notices des chercheurs (sous différents formats) lors de leurs plusieurs missions à Saint-Omer; le catalogue de H. Michelant sous forme papier et publié en ligne sur le CCFr[5] en EAD. Cette variété de métadonnées, leur complexité, leur convergence et leur répétition, d'une part, et le besoin de les fusionner, d'autre part, impliquent de réfléchir sur différentes possibilités de leur traitement. Cela concerne, notamment, la création du catalogue en format numérique contenant les métadonnées; la

---

[2] Étant donné que cette abbaye a fourni des textes de toutes natures aux savants dès le Moyen Âge et jusqu'à la philologie contemporaine, en passant par la Renaissance. Cf. la présentation du projet sur le site de »Biblissima«, http://www.biblissima-condorcet.fr/fr/appels-a-projets/projets-retenus/saint-bertin-centre-culturel-viie-xviiie-siecle-constitution.

[3] »Medium, le répertoire des manuscrits reproduits et recensés«: http://medium.irht.cnrs.fr/. Une nouvelle version, qui est en cours, est disponible sur ce site: http://medium-avance.irht.cnrs.fr/index/index. »FAMA, la base des œuvres latines médiévales à succès«: http://fama.irht.cnrs.fr/fr/. »Initiale. Catalogue informatisé des manuscrits enluminés des bibliothèques publiques de France«: http://Initiale.irht.cnrs.fr/accueil/index.php.

[4] Patrick PECCATE, Soft Experience. Métadonnées, une initiation: Dublin Core, IPTC, Exif, RDF, XMP, etc., première version: août 2002, dernière mise à jour: 13 décembre 2007, http://peccatte.karefil.com/Software/Metadata.htm.

[5] Catalogue collectif de France, http://ccfr.bnf.fr/portailccfr/jsp/public/index.jsp?failure=/jsp/public/failure.jsp&success=/jsp/public/index.jsp&profile=public.

fusion de toutes les métadonnées et leur enrichissement; les normalisation, standardisation et création des liens dans les buts d'interopérabilité du catalogue produit; et finalement la conversion du catalogue en TEI vers un catalogue en EAD conforme aux »bonnes pratiques« (d'après les standards de la BnF) pour la description des manuscrits médiévaux.

2. 1. Langage XML

Le métalangage XML (*Extensible Markup Language* – »langage de balisage extensible«), développé à partir de 1996, a plusieurs avantages informatiques. En premier lieu, il est propice à l'interopérabilité des systèmes. Ensuite, il permet de garder les données à long terme, ce qui aide à assurer une certaine pérennité des documents. Enfin, au niveau des spécificités structurelles du XML, il est basé sur l'encodage du contenu logique du document et non pas sur celui de sa mise en forme. Par ailleurs, les décisions concernant cette dernière appartiennent complètement à l'éditeur, ce qui contribue à la présentation personnalisée du texte encodé[6]. Toutes ces caractéristiques permettent de faciliter l'exploitation et l'échange des données encodées dans le langage XML. Ces données deviennent donc accessibles pour des utilisateurs multiples et sont ouvertes à la modification.

Ces propriétés ont rendu le XML de plus en plus répandu dans les projets numériques, ce qui a mené à la création de schémas qui aident à modéliser les sources similaires[7]. Parmi les exemples de tels schémas, on peut mentionner Docbook (apprécié pour l'encodage des documents techniques), l'EAD (pour la description des fonds et des documents d'archives) et la TEI (standard pour l'édition des textes, en particulier historiques et littéraires).

2. 2. Format TEI: »une sorte d'encyclopédie sur des notions textuelles largement acceptées[8]«

La communauté TEI (*Text Encoding Initiative*)[9], créée en 1987, a développé et proposé des standards et des recommandations pour la représentation du matériel textuel en version numérique

---

[6] Marjorie BURGHARDT, Editer des sources historiques en ligne grâce à XML – Un guide pratique, p.5, http://mutec.huma-num.fr/sites/www.mutec-shs.fr/files/Guide%20Editer%20des%20sources%20historiques%20%20gr%C3%A2ce%20a%20XML.pdf.

[7] Ibid., p. 13–14.

[8] »The TEI framework provides a useful way of thinking about the nature of text: it constitutes a kind of encyclopedia of generally-agreed textual notions.« Lou BURNARD, What is the Text Encoding Initiative? How to add intelligent markup to digital resources, Marseille 2014 (Encyclopédie numérique), http://books.openedition.org/oep/680.

à travers l'encodage des textes. Le format d'encodage des textes XML-TEI est devenu très répandu et est le plus fréquent parmi les utilisateurs actuels. C'est ce format qui a été utilisé pour l'encodage du catalogue des fonds de la BASO dans le cadre du projet Saint-Bertin. Étant donné que ce langage »permet de décrire les caractéristiques sémantiques d'un texte plutôt que sa présentation«[10], la TEI aide à capter une gamme beaucoup plus large de détails textuels et métatextuels que d'autres formats qui ciblent aussi l'encodage des caractères et de structure de base de textes (par exemple ePub).

Un autre grand avantage de la TEI est sa capacité à suggérer nombre de décisions structurelles qui permettent de représenter au mieux le texte, quel que soit son support. La TEI sert également à créer une édition critique[11]. De plus, elle offre les éléments nécessaires pour garder la version primitive du texte en proposant également une (ou plusieurs) version modifiée (qui tient compte des ajouts, suppressions, corrections et endommagements possibles)[12].

Enfin, il faut mentionner les avantages suivants du format TEI: la possibilité d'effectuer la publication électronique sur la base du document XML-TEI; l'application de ce dernier en tant que base de données exploitable[13]; et les différents niveaux possibles de granularité de la description du document (c'est-à-dire le niveau d'accès à l'information). Grâce à toutes ces caractéristiques de la TEI, son application permet d'englober toute la richesse de chaque manuscrit ainsi que d'encoder les nuances et les observations les plus minutieuses lors de l'expertise scientifique.
Le

Tableau **1** montre le schéma-extrait d'un document TEI-XML qui présente la structure de l'encodage d'un manuscrit du catalogue de Saint-Bertin:

---

[9] Cette communauté des chercheurs en humanités, sciences sociales et linguistique est organisée en TEI Consortium (TEI-C http://www.tei-c.org) qui maintient et développe les standards documentés en *Guidelines (Recommandations pour l'encodage et l'échange de textes électroniques*), TEI P5, élaborées en 2003. Cf. Tutoriel TEI, http://teibyexample.org/modules/TBED00v00.htm.

[10] »Qu'une séquence de mots exprime un titre de livre, par exemple, et non seulement un bloc quelconque à représenter en italique«; cf. Stéfan SINCLAIR, Geoffrey ROCKWELL, Les potentialités du texte numérique, chap. 12, dans: Marcello VITALI-ROSATI, Michael E. SINATRA (dir.), Pratiques de l'édition numérique, Montréal 2014, p. 191–204, http://books.openedition.org/pum/337?lang=it.

[11] BURNARD, What is the Text Encoding Initiative (voir n. 8), cf. chap. »Varieties of textual structure«, http://books.openedition.org/oep/688

[12] Tutorial TEI, http://teibyexample.org/modules/TBED06v00.htm.

[13] Cf. les arguments de Laura Lebarbey (qui a créé le catalogue TEI pour le projet Saint-Bertin en 2015): L. LEBARBEY, Reconstituer et étudier une collection ancienne au-delà de la pluralité des bases et des formats, École nationale des chartes, mémoire de master, 2015, p. 34.

```
<TEI>
<teiHeader> </teiHeader>
<text>
<msDesc xml:id="Saint-Omer_n°" resp="#NN">
        <msIdentifier> l'identifient du manuscrit contenant les données sur sa conservation actuelle </msIdentifier>
        <msContents> : contenu textuel du manuscrit qui peut consister en plusieurs décomposants </msContents>
        <physDesc> caractéristiques codicologiques, paléographiques et iconographiques du manuscrit (donc la description matérielle complète, peut contenir plus de balises) </physDesc>
        <history> datation, origine et provenance du volume </history>
        <additional> toutes les informations complémentaires (bibliographie pour chaque manuscrit précis, références et informations administratives) </additional>
        <msPart n="1°"> (unité codicologique d'un manuscrit) contient des informations sur une partie d'un manuscrit, distinct à l'origine, qui fait aujourd'hui partie d'un manuscrit composite. </msPart>
</msDesc>
</text>
</TEI>
```

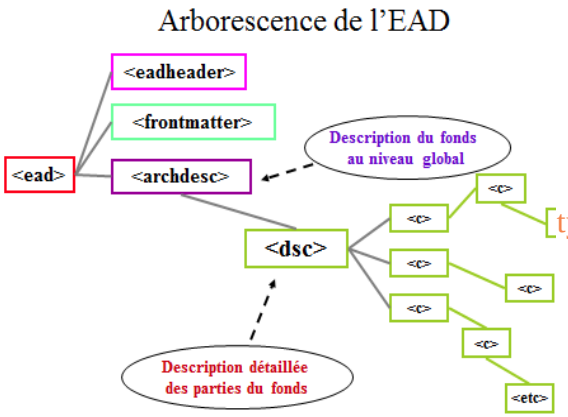
Tableau 1: Structure de la partie concernant le manuscrit (<msDesc>)

La structure choisie semble être précise, et adaptée à la création du catalogue électronique des manuscrits des fonds anciens. Le regroupement des métadonnées fait partie d'une balise <teiHeader> (dans la structure du fichier, elle est un élément nécessaire et se trouve tout au début) et toutes les notices des manuscrits se trouvent dans le <text> qui représente une liste contenant plusieurs <msDesc>, ce qui assure l'ergonomie, l'exploitation facile et la rédaction accessible du fichier.

Certes, la structure globale du fichier TEI (avec la division en <teiHeader> et <text>) est intangible, mais la façon d'encoder les données varie d'un document à l'autre. L'avantage de la TEI est donc sa malléabilité, qui permet de l'utiliser en l'adaptant au mieux à l'information disponible. D'un côté, cette caractéristique de la TEI aide à élaborer le meilleur modèle pour la source historique à encoder. Mais d'un autre côté, ceci incite à créer de multiples solutions d'encodage pour différents projets, ce qui peut rendre l'interopérabilité plus difficile.

2. 3. Le format EAD

Le format EAD (*Encoded Archival Description* ou Description archivistique encodée) représente un standard d'encodage des instruments de recherche archivistique destinés à être mis en ligne, qui est basé sur le langage XML[14]. Créé dans les années 1990, l'EAD a été spécialement élaboré pour le traitement des données des fonds d'archives. Son avantage par rapport aux formats existants (par exemple, MARC) consistait en une meilleure adaptation à la structuration des ensembles documentaires organisés sur plusieurs niveaux hiérarchiques. Ce format est donc beaucoup utilisé dans les archives françaises, car il permet de structurer au mieux la description des documents, grâce à l'arborescence développée de la hiérarchie des fonds, systématisé dans les balises EAD:




```
<ead>
<eadheader> </eadheader>
<frontmatter></frontmatter>
<archdesc level="collection"
type="catalogue">
<did></did>
<bibliography ></bibliography>
<dsc>
<c level="otherlevel" otherlevel="notice">
<did></did>
<custodhist></custodhist>
<scopecontent></scopecontent>
<c level="otherlevel"
otherlevel="sous-notice">
<did></did>
<scopecontent></scopecontent>
</c>
</c>
```

[14] L'EAD se fonde sur les principes de la DTD EAD 2002 (Document Type Definition) qui permet un respect complet de la norme internationale de description archivistique ISAD (G). Cf. Historique de l'EAD, http://bonnespratiques-ead.net/guide/intro/historique.

	</dsc> </archdesc> </ead>

Tableau 2: Arborescence de l'EAD[15] (gauche); L'EAD en code XML (droite)

On voit sur le Tableau 2 à gauche que la structure de l'EAD est très visible et claire. De plus, sur le même tableau à droite elle est comparable avec celle de la TEI. Si, dans la TEI, la division globale se présente comme <teiHeader> (métadonnées sur le document électronique) et <text> (données sur les objets encodés), dans l'EAD cette fonction est remplie par <eadheader> (pour les métadonnées) et <archdesc> (pour la description du contenu). L'élément <frontmatter>[16] n'est pas obligatoire mais il permet d'encoder les informations pour la page de titre. Tous les éléments décrits qui se trouvent dans le fonds sont encodés à l'intérieur des balises <c>[17] (composants du fonds) qui peuvent descendre dans la hiérarchie jusqu'au dernier dossier dans lequel les documents sont conservés.

Néanmoins, la Bibliothèque nationale de France utilise le format EAD également pour l'encodage des manuscrits. Son »Catalogue général des manuscrits des bibliothèques publiques de France« (CGMBPF – ou plus raccourci CGM) existe depuis 1833 et est passé depuis 1993 sous

[15] Récupérée du Guide pour la mise à jour des fichiers issus de la conversion rétrospective du CGM, p. 5, http://www.bnf.fr/documents/cgm_retroconverti_mise_a_jour.pdf.

[16] Cet élément donne des informations sur la création, la publication ou l'utilisation de l'instrument de recherche lui-même et non sur les documents en cours de description. Cf. Encoded Archival Description, Tag Library, Description Archivistique Encodée, Dictionnaire des balises, Society of American Archivists, traduit de l'anglais par le groupe AFNOR CG46/C N357/GE3, octobre 2004, p. 137, http://www.archivesdefrance.culture.gouv.fr/static/1066.

[17] Élément englobant qui désigne une sous-partie des documents en cours de description. Un composant <c> fournit des informations sur le contenu, le contexte et l'importance matérielle d'un sous-ensemble de documents. Ibid., p. 59.

forme électronique[18]. Une partie de ces notices a été reçue par le CCFr, qui vise à améliorer la visibilité des collections de manuscrits par la mise en ligne des notices descriptives, ainsi qu'à relancer la dynamique de signalement en permettant les mises à jour et les ajouts[19].

L'EAD, quant à lui, présente beaucoup mieux la structure du fonds sans se lancer dans la description complète de la source. Cette caractéristique en fait un bon outil au sein des archives. La souplesse qu'offre ce format et sa capacité à replacer les documents dans leur contexte en restituant l'arborescence des fonds conduisirent le comité de pilotage à le retenir[20]. Mais au sein du CCFr une autre approche a été retenue: celui du signalement de l'information. Dans le cadre de CCFr, l'EAD est donc utilisé pour structurer l'ensemble du texte et pour l'indexer. Cette indexation est faite de telle façon que le texte soit rapidement et facilement repérable lors du lancement de la requête de recherche en ligne.

3. Exercice de conversion

Les objectifs numériques différents des formats TEI et EAD nécessitent donc non seulement un travail technique pour la conversion, mais aussi une réflexion sur la restructuration des métadonnées. La conversion est souvent réalisée à l'aide d'une feuille de style XSLT[21]. Étant donné la multiplicité des options possibles lors de la conversion, les choix doivent être faits en fonction des deux aspects suivants. En premier lieu, cette conversion tient à respecter la particularité du format cible, l'EAD, avec sa hiérarchie de niveaux qui peut être utile pour rester fidèle à la nature de la source historique encodée. En second lieu, il était prévu de garder les informations encodées en leur richesse sans pourtant trop alourdir le fichier de sortie. Afin d'atteindre ces objectifs, il était donc indispensable lors du travail de revoir toute la structure du document encodé et de suivre les correspondances entre les éléments utilisés dans ces deux formats.

---

[18] *Guide pour la mise à jour des fichiers issus de la conversion rétrospective du CGM*, p. 51–57, http://www.bnf.fr/documents/cgm_retroconverti_mise_a_jour.pdf.

[19] Florent PALLUAULT, Le catalogue général des manuscrits des bibliothèques publiques de France: informatisation et avenir, dans: Bulletin des bibliothèques de France (BBF), 2009, n° 1, p. 68–72, http://bbf.enssib.fr/consulter/bbf-2009-01-0068-010.

[20] Guide pour la mise à jour des fichiers issus de la conversion rétrospective du CGM, p. 53, http://www.bnf.fr/documents/cgm_retroconverti_mise_a_jour.pdf.

[21] XSLT (EXtensible Stylesheet Language) est un langage de transformation de la structure et du contenu d'un document XML. Cf. M. R. KAY, XSLT 2.0 and XPath 2.0: Programmer's Reference, 4th ed, Indianapolis 2008, http://197.14.51.10:81/pmb/INFORMATIQUE/XSLT2.0%20and%20XPATH%202.0.pdf.

3.1. La structure du document

Sachant que les trois partitions principales de l'EAD sont <eadHeader>; <frontmatter> et <archdesc>, il faut de prime abord réfléchir sur leur contenu en partant du fichier en TEI. La balise <eadHeader> contenant les métadonnées du catalogue des manuscrits a été principalement créée à partir de la balise <teiHeader> du document TEI. <frontmatter> n'a pas été utilisé (en tant qu'élément non obligatoire, mentionné plus haut). Ici, les transformations de la balise <archdesc> seront donc abordées.

La balise <archdesc>, quant à elle, indique la collection, c'est-à-dire les fonds anciens de la BASO. Elle décrit également les objets eux-mêmes, les notices des manuscrits proprement dits. La description générale de la collection n'est pas présentée dans un fichier TEI sous une balise unique en tant qu'élément séparé; c'est pourquoi, lors de la conversion, les informations du fichier TEI ont été extraites de différents endroits et réunies ensuite dans <archdesc> de l'EAD. Cela concerne notamment la bibliographie sur le fonds de la BASO et les informations sur la BASO en tant qu'organisme responsable de l'accès intellectuel aux manuscrits. Ces dernières ont finalement été reprises du fichier EAD initialement encodé par le CCFr et mises directement dans un fichier XSLT, à part certaines balises dont le contenu était converti depuis la TEI. Cette description se trouve donc encodée au niveau de la collection avec un attribut *@type* proposé par le CCFr comme »catalogue« (cf. Annexe 2): <archdesc level="collection" type="catalogue">. En ce qui concerne la bibliographie, ces informations sont encodées dans la balise <bibliography> signalant les publications pertinentes liées aux documents décrits[22]. Or, dans la TEI, les informations générales (toutes les métadonnées sur l'ensemble du fonds décrit) seront automatiquement placées à l'intérieur de la balise <teiHeader>, dans le <sourceDesc> qui représente une description bibliographique pour un texte numérisé[23]. À cet exemple, on peut voir un fonctionnement différent des métadonnées en TEI et en EAD.

La possibilité de réunir les informations identiques concernant tout l'ensemble des objets décrits dans le catalogue au niveau de la balise <archdesc> devient un avantage primordial de l'EAD. Cette option illustre bien la capacité du format EAD à décrire le fonds en descendant

[22]EAD en bibliothèque, Guide des bonnes pratiques, http://bonnespratiques-ead.net/guide/communication-utilisation/docs_en_relation/bibliographie.
[23] TEI Guidelines, http://www.tei-c.org/release/doc/tei-p5-doc/fr/html/ref-sourceDesc.html.

niveau par niveau dans sa structure. En outre, cette balise englobant tout l'ensemble documentaire encodé décrit son contenu, son contexte et son importance matérielle[24]. Dans le cadre d'un fonds de bibliothèque et non d'archives, la description matérielle est plus importante pour la description de chaque pièce que pour celle de tout le fonds.

3.2. Notice du manuscrit

La balise <archdesc> permet également de passer à des niveaux inférieurs, qui sont encodés dans la balise <dsc>. Après la description générale de la collection, on descend donc à <dsc>, au sein duquel il existe plusieurs <c> (composants[25]). Chaque composant correspond en l'occurrence à la notice du manuscrit et représente ainsi les données converties du fichier TEI de la balise <msDesc>: <dsc type="combined"> >> <c level="otherlevel" otherlevel="notice">.

Il ne sera pas inutile de mettre en question le choix des attributs pour les balises du fonds (<archdesc> et <c>). Ces attributs sont obligatoires dans le format EAD, qui vise à refléter la structure du fonds. Étant donné qu'ils sont absents dans la TEI, il est nécessaire de les ajouter lors de la conversion vers l'EAD. En même temps, la logique dans l'interaction des niveaux en TEI doit être gardée en EAD: la répartition des attributs correspondant à la notice du manuscrit au sein d'un catalogue.

La redistribution des attributs et de leur contenu paraît être une question importante pour l'EAD, car elle incite à une réflexion sur la nature du document numérique produit. Quant à la description archivistique du fonds (dans le cadre de laquelle l'EAD est destiné à fonctionner), le niveau le plus bas, et donc le plus précis sur cette échelle, est celui de la pièce (*item*). D'après le schéma des niveaux de description du fonds d'archives, on peut encoder cette hiérarchie des niveaux en référence aux instruments de recherche qui leur correspondent (»fonds«, »sous-fonds«, »série«, »sous-série«, »dossier«, »pièce«[26]). Deux autres options au niveau de l'attribut @level, »collection« et »otherlevel«, sont aussi possibles.

Premièrement, il faut souligner l'importance primordiale de la différence entre la notion de »fonds« et celle de »collection«. Elle s'explique par le fait que, du point de vue de l'EAD, le fonds

---

[24] Encoded Archival Description, Tag Library (voir n. 16), Description Archivistique Encodée, Dictionnaire des balises, p. 42.
[25] Encoded Archival Description, Tag Library (voir n. 16), p. 59.
[26] Ce schéma traite les niveaux suivants: fonds, sous-fonds, série, sous-série, dossier, pièce. Cf. Bruno GALLAND, Christine NOUGARET, Les instruments de recherche dans les archives, Paris 1999, p. 119.

représente un ensemble organique de documents, tandis que la collection est une réunion artificielle de documents en fonction de critères communs sans considération de leur provenance[27]. C'est pourquoi, dans le fichier EAD pour le fonds manuscrit de la BASO, la notion »collection« est proposée. La mention de ce niveau est placée dans la balise <archdesc>, qui caractérise la nature de tout l'ensemble documentaire: le catalogue contient les manuscrits réunis au sein de la bibliothèque municipale, acquis pour leur plupart pendant la Révolution française lors de la confiscation des biens du clergé.

Deuxièmement, une question se pose sur l'attribut @level="otherlevel", qui correspond à une réunion factice de pièces qui peut être faite autour d'un thème fédérateur[28]. La décision finale a été prise de ne pas indexer la notice de chaque manuscrit de cette collection comme *item*, <c06 level="item"> (pièce correspondant à la plus petite unité documentaire, d'un point de vue matériel et intellectuel[29]). De fait, cette décision respecte les particularités de la source plutôt que les exigences hiérarchiques de sa description. Si on encode le manuscrit en tant que pièce (*item*), un tel encodage ne permet pas de descendre plus bas dans la hiérarchie. Or, un grand nombre de manuscrits sont composites, ce qui signifie que leur structure est complexe, contenant plusieurs entités textuelles ainsi que des unités codicologiques séparées. Cette structure serait négligée dans ce cas. Le choix est donc fait en faveur de la nature de la source, ce qui cause évidemment un souci dans l'attribution des niveaux de description. Ainsi, on est obligé de passer de l'attribut *@level* vers celui de *@otherlevel* malgré le fait que cela évoque plutôt la description de la description et non la description d'un objet (de la notice de manuscrit).

La solution proposée pour chaque notice du manuscrit en EAD est donc celle-ci: <c level="otherlevel" otherlevel="notice">, ce qui est précisé par le contenu du deuxième attribut *@otherlevel* (cf. Annexe 3). Différentes parties codicologiques du manuscrit composite (<msPart> en TEI), quant à elles, peuvent être exprimées en EAD dans un autre sous-élément <c> avec le contenu »sous-notice« de son attribut @otherlevel: <c level="otherlevel" otherlevel="sous-notice">.

L'avantage de cette décision réside dans le fait qu'elle permet de garder la hiérarchie de la source ainsi que de son encodage en TEI, un format très attentif à la nature du document.

---

[27] EAD en bibliothèque, Guide des bonnes pratiques, cf. http://bonnespratiques-ead.net/guide/organisation-description/level.
[28] Ibid.
[29] Ibid.

L'inconvénient consiste en la définition du niveau »notice« pour le manuscrit entier et »sous-notice« pour ses parties intérieures. Ces notions sont pourtant comprises en tant que valeurs conventionnelles par l'EAD, ce qui crée une certaine confusion pour son fonctionnement standardisé[30]. La conversion effectuée permet de réfléchir sur les capacités des formats à gérer les informations des sources historiques.

3.3. Description matérielle et l'histoire du manuscrit

Pour la description matérielle de la source, les deux formats, TEI et EAD, contiennent un grand nombre de balises. Le »Manuel de catalogage des manuscrits médiévaux« de la BnF propose de créer de multiples <physfacet> à l'intérieur de la <physdesc>[31]. Chaque <physfacet> a l'attribut @type dont le contenu correspond parfaitement aux étapes de description codicologique au sein de la balise <physDesc> du TEI. L'Annexe 4 présente la correspondance entre les balises dans les deux formats. Comme leur niveau d'indexation est complètement différent (en TEI, le balisage est extrêmement fin, tandis qu'en EAD, les paragraphes entiers de texte sont un phénomène assez fréquent), il faut souvent utiliser la formule <xsl:text></xsl:text> pour formater les données dans le fichier de sortie. Le travail de conversion à cette étape doit être particulièrement minutieux car il est nécessaire de bien vérifier la syntaxe et la correspondance des données des balises au texte qui les décrit.

Quant à l'histoire de la conservation du manuscrit (y compris les personnes morales et les personnes physiques qui l'ont possédé), elle est contenue en EAD dans la balise <custodhist>[32] et en TEI dans la balise <history>. La structure élaborée du document TEI dans <history> suppose que chaque manuscrit peut avoir plusieurs provenances. Ces provenances sont regroupées par les possesseurs. La solution proposée (cf. Annexe 5) est élaborée en portant attention à la structuration du fichier en TEI. De plus, elle permet de faciliter la conversion par les règles du XSLT. Au sein d'une grande <custodhist>, plusieurs <custodhist> sont créés dont chacun correspond à la <provenance> de la TEI.

---

[30] Ibid.

[31] Manuel de catalogage des manuscrits médiévaux, cf. http://guideducatalogueur.bnf.fr/abn/GPC.nsf/gpc_page?openform&type_page=fiche&unid=5BBE42CB5B73612FC1257C6200486D2E.

[32] EAD en bibliothèque, Guide des bonnes pratiques, cf. http://bonnespratiques-ead.net/guide/communication-utilisation/acquisition_evaluation/historique_garde.

En même temps, on peut observer la transformation complète du paragraphe <p> contenant les balises <locus>, <date> et <idno> en TEI vers un paragraphe <p> contenant seulement les données textuelles en EAD. Du point de vue de la conversion, le retour des données indexées et donc interopérables vers le texte diminue l'efficacité de leur traitement numérique. C'est exactement cette caractéristique de l'EAD qui cause des doutes chez les chercheurs sur pertinence de son usage pour le contenu des sources historiques.

4. Conclusion et perspectives

Le travail de conversion démontre donc qu'au niveau de la syntaxe et de la structure, les deux formats ont tant des ressemblances que des différences au sein du même langage XML. En ce qui concerne leurs ressemblances, il faut mentionner que la hiérarchie du contenu du manuscrit dans la diversité de ses unités intellectuelles et codicologiques est conservée dans le fichier de sortie en EAD. Les caractéristiques principales de la notice du manuscrit sont aussi bien présentées dans le format TEI que dans l'EAD: l'identifiant du manuscrit, son contenu, sa description physique, ses données de provenance et de datation.

Les différences sont pourtant aussi nombreuses. On peut les répartir en trois groupes. Les premières concernent la réorganisation structurelle des informations (ainsi, avec la bibliographie). Les transformations de ce type ne concernent que les balises et leur position au sein de différents formats. À ce niveau, il est intéressant de s'interroger sur les différents besoins des formats et les formes d'encodage dont ils disposent.

Deuxièmement, des différences portent sur le changement des métadonnées liées à la structure qui possède l'édition électronique (c'est-à-dire les données concernant l'institution éditrice). En ayant effectué la conversion complète en fichier EAD, il ne faudra pas oublier leur actualisation. De plus, il faut tenir compte de la modification technique, c'est-à-dire faire attention à la syntaxe ainsi qu'aux changements de formulations (ici, il n'est pas inutile de tenir compte de la prépondérance du texte dans l'EAD, etc.).

Troisièmement, la TEI vise à une indexation très précise des données en utilisant plusieurs attributs et toute une liste de leurs contenus possibles. Or, pour le catalogue en EAD utilisé par CCFr, cette sorte de données semble trop détaillée car, comme il a été mentionné plus haut, le CCFr

préfère retenir les données d'une manière qui facilite le lancement de requête. Cette divergence témoigne des différents objectifs de ces deux formats dans l'encodage des informations.

La multiplicité des options de conversion permises par le nombre des balises existantes dans la TEI et l'EAD suscite un vaste champ d'hypothèses et de réflexions sur les particularités des formats et pousse les institutions présentant le produit définitif à limiter leur choix en fonction des utilisateurs et des enjeux qu'ils représentent. Le travail de conversion aide également à adapter le format à partir duquel on effectue cette conversion, en l'occurrence la TEI, car le document doit être bien structuré et encodé de façon logique, ce qui se répercutera sur une extraction rapide des données.

De même, le traitement de chaque information visant à conserver celles qui sont importantes reste une des principales questions. Comment mener au mieux le dialogue entre le créateur, l'utilisateur et le destinataire? Certaines solutions de conversion présentées ici visent à respecter la structure du document de sortie en EAD et à garder la logique de la notice très enrichie en TEI.

Annexes

1. Métadonnées

<teiHeader> <fileDesc> <titleStmt/> <title/> <author /> <respStmt/> <publicationStmt/> <sourceDesc/> </fileDesc> <encodingDesc/> **</teiHeader>**	**<eadheader>** <filedesc> <titlestmt/> <editionstmt/> <publicationstmt /> <notestmt> <list/> </notestmt> </filedesc> **</eadheader>**

2. Changement de structure

• <msDesc>
  • <msIdentifier>
  • <settlement/>
  • <repository/>
  • <idno/>
  • <altIdentifier/>
  • </msIdentifier>
  • <msContents>
  • <summary/>
  • <textLang/>
  • </msContents>
• </msDesc>
	• <archdesc level="collection" type="catalogue">
	  • <did>
	  • <repository/>
	  • <unitid/>
	  • <unittitle/>
	  • <unitdate/>
	  • <langmaterial/>
	  • </did>
	  • <accessrestrict/>
	  • <userrestrict/>
	  • **<bibliography/>**

3. *Hiérarchie de l'EAD*

<archdesc level="collection“ type=“catalogue”>

<did> </did>

<dsc type="combined“>

<c01 level=“otherlevel“ otherlevel=“notice”>

<c02 level=“otherlevel“ otherlevel=“sous-notice”>

<…>

</archdesc>

4. Description matérielle

<physDesc> <objectDesc> <supportDesc> <extent> <dimensions> <collation> </supportDesc> <layoutDesc/> </objectDesc> <handDesc> <scriptDesc/> <decoDesc/> <bindDesc/> </physDesc>	<physdesc> <physfacet type="écriture"/> <physfacet type="mains"/> <physfacet type="décoration/> <physfacet type =« codicologie"/> <physfacet type="support"/> <physfacet type="réglure"/ > <physfacet type="reliure"/> <extent/> <dimensions/> </physdesc>

5. Histoire du manuscrit

<history> <origin> </origin> <provenance> *(plusieurs)* *(plusieurs)* <locus/> <date/> <q/> <note/> </provenance> </history>	<c> <did> <unitdate era="ce" calendar="Gregorian" normal="1375/1400">Fin du XIVe siècle</unitdate> </did> <custodhist> *(plusieurs)* TEXTE </custodhist>