Comparaison des formats EPUB et XML pour la production de livres adaptés

Alex Bernier
9 septembre 2014

1) Introduction

Le nombre de fichiers au format EPUB déposés par les éditeurs en réponse aux demandes de titres formulées sur Platon par les organismes agréés augmente. EPUB est souvent confondu avec XML : cette note a pour objectif d'expliciter les différences entre ces deux formats dans le contexte de la production de livres adaptés et d'exposer les arguments qui conduisent les organismes adaptateurs à préférer XML.

2) Précision sur les versions d'EPUB et de XML

Actuellement, la plupart des fichiers dits « XML » déposés sur Platon utilisent la DTD "littérature générale" (LG) et la totalité des fichiers dits « EPUB » utilisent la version 2 de la spécification de ce format. Ainsi, nous comparons dans cette note le format XML LG avec le format EPUB 2.0 : la plupart des arguments exposés restent toutefois valables avec EPUB 3.

3) Distinction entre formats de distribution et formats d'édition

EPUB est un format de distribution : cela signifie qu'il est destiné à être traité par un logiciel dont le but est de restituer le document en vue de sa consultation par l'utilisateur. Dans sa conception, un format de distribution diffère des formats d'archivage et des formats d'édition destinés à être utilisés dans les chaînes de productions de documents.

Aujourd'hui, les organismes produisant le plus grand nombre de documents adaptés utilisent les formats XML DTBook ou XML ZedAI spécifiés par le consortium DAISY comme formats d'édition. Ces formats utilisent des grammaires riches qui permettent de la complexité des documents destinés à être lus en braille, gros caractères et/ou à être écoutés. Les fichiers « XML LG » déposés par les éditeurs sont ainsi convertis en XML DTBook ou ZedAI et enrichis.

4) EPUB 2.0 est moins riche et moins normalisé que XML LG

Le format EPUB est d'un point de vue sémantique moins riche que le format XML LG. À titre d'exemple, XML LG définit des balises spécifiques pour rendre compte de notes de bas de pages, de la numération des pages et des dialogues : ces éléments n'existent pas dans le format EPUB 2.0 car celui-ci est basé sur HTML et que ce langage ne définit pas de manière standard de restituer ces types de contenu. La manière de produire ces éléments diffère ainsi en fonction du logiciel utilisé pour générer le document au format EPUB. Lorsqu'ils reçoivent un document EPUB, les organismes adaptateurs ne peuvent ainsi pas se reposer sur une grammaire bien définie alors qu'ils le peuvent lorsqu'ils reçoivent le même document au format XML LG.

5) EPUB 2.0 ne permet pas de reconstituer la structure du document d'origine

Un document XML LG contient l'intégralité du texte du livre dans un fichier unique. Un document EPUB 2.0 est la plupart du temps composé d'un nombre important de fichier HTML contenant le texte du livre. Dans un document XML LG, la hiérarchie du livre est clairement exposée (par exemple par l'usage des balises "par", "chap", "schap"). Dans un document EPUB 2.0, la richesse moindre

du format HTML et la fragmentation du texte dans un nombre important de fichiers rendent difficiles voir impossibles de reconstituer la hiérarchie du livre.
Par exemple, un document EPUB 2.0 contient les fichiers suivants :
- x1.html avec le titre "Première partie"
- x2.html : avec le titre "Chapitre 1" puis le contenu du chapitre
- x3.html : avec le titre "Chapitre 2" puis le contenu du chapitre
- x4.html avec le titre "Deuxième partie"
- Etc.

EPUB 2.0 n'offre pas de moyens permettant de retrouver automatiquement les liens hiérarchiques entre les fichiers (dans l'exemple, x1.html et x4.html sont au même niveau; x2.html et x3.html sont de sous-niveaux de x1.html).

6) Absence de logiciel permettant d'éditer des documents EPUB

La conséquence des éléments exposés ci-dessus est que les logiciels permettant de générer des documents EPUB ne l'utilise pas comme un format de production : ils utilisent un format interne (InDesign, DocBook, DTBook, etc.) pour éditer le document et convertissent ce format interne en EPUB. Ainsi, il n'est actuellement pas possible (et pas souhaitable, pour les raisons exposées ci-dessus) d'éditer directement un document au format EPUB 2.0. Les organismes adaptateurs ne disposent donc pas d'outils d'éditions pour traiter l'EPUB.

7) Conclusion

Deux manières différentes pouvaient être envisagées pour produire un document adapté à partir d'un fichier EPUB 2.0 :
- Le convertir dans un format XML DTBook ou ZedAI pour l'intégrer dans la chaîne de production
- L'éditer directement avec un logiciel permettant de l'enrichir

Les arguments exposés ci-dessus mettent en évidence que ces deux pistes ne sont pas exploitables en pratique :
- EPUB 2.0 étant moins riche et moins fiable que XML LG, la conversion vers XML DTBook ou XML ZedAI occasionne des pertes d'informations et n'est plus automatique
- L'édition directe d'un document EPUB 2.0 est rendue très difficile (impossible en pratique) de par la conception même d'EPUB comme un format de distribution

En conséquence, la livraison sur Platon de documents au format XML LG (ou de tout autre format XML "riche") doit être privilégiée car ce sont ces formats qui permettent d'automatiser les premières étapes de l'adaptation tout en garantissant la conservation d'informations sémantiques et structurelles du livre.