

Spécialité Informatique - 1^{re} année

Rapport de mini-projet

Le format



Tristan Lamontagne
Martin Théault

Table des matières

Introduction.....	4
1 Découverte du format ePub.....	6
1.1 Qu'est-ce que ePub ?.....	6
1.2 Qui est à l'origine de ce format ?.....	6
1.3 Pourquoi ePub?.....	6
1.4 Avantages et limites.....	7
2 Les spécificités techniques d'un fichier ePub.....	8
2.1 Les standards	8
2.1.1 OCF.....	8
2.1.2 OPS.....	8
2.1.3 OPF.....	8
2.2 L'arborescence.....	8
2.3 Description des fichiers.....	9
2.4 Création manuelle d'un fichier ePub.....	12
3 Outils	13
3.1 Aide à la création	13
3.2 Validation.....	13
3.3 Conversion de format.....	13
3.4 Visualisation	14
Conclusion.....	15
Références bibliographiques.....	16

Une version eBook de ce rapport au format ePub est disponible à l'adresse suivante :

http://www.martintheault.fr/projet_epub/rapport.epub

Introduction

C'est en 1971 que le premier eBook apparut via le projet Gutenberg qui avait pour but de numériser des livres. Il s'agissait de la « Déclaration d'indépendance des États-Unis » [1].

D'un point de vue technique, on appelle eBook, ou encore « livre électronique » ou « livrel », tout fichier électronique contenant un texte numérisé.

Ces livres sont destinés à être lus sur des appareils appelés des **liseuses**. Le marché est très récent et les eBooks commencent à occuper une part de plus en plus importante sur le marché des livres. On peut d'ailleurs rencontrer de temps en temps des gens équipés de liseuses dans les transports en commun.

Les livres électroniques ont manqué un rendez-vous dans les années 2000 [2]. Cet échec provenait en partie de la piètre qualité des écrans – et donc du mauvais confort de lecture et de l'importante fatigue des yeux en résultant – mais également du prix extrêmement élevé des liseuses (le « Cybook » de « Cytale » par exemple, était vendu presque 900€).

Mais aujourd'hui, les eBooks signent un retour remarqué : une technologie d'écran idéale pour la lecture et des prix tout à fait abordables (on en trouve d'excellents à partir de 250€).

Le leader du marché des liseuses est « Amazon » qui commercialise le « Kindle » (cf Figure 1).

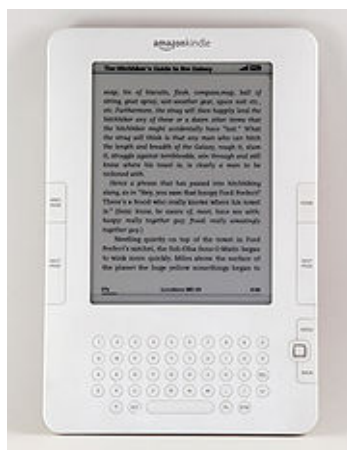


Figure 1: Le Kindle d'Amazon, disponible en France depuis 2009

On peut également citer des marques tel « Sony » avec son « Sony Reader » ou encore « iRex » qui propose « iLiad ». La fourchette de prix des liseuses est relativement large, il faut compter aux alentours de 100€ pour la moins chère et le haut de gamme peut atteindre 500€.

Il est également possible de lire les eBooks sur certains téléphones portables ou des ordinateurs à l'aide d'un logiciel. Ce fut le cas pour nous lors de ce projet ; nous avons utilisé le logiciel « Calibre » [3] sur Linux afin de visualiser le fichier ePub que nous avons créé.

Il existe plusieurs formats de données pour les eBooks : PDF, txt, Moby Pocket, etc.

Or, il n'existe pas d'harmonisation du côté des constructeurs. Il y a donc un risque qu'un eBook avec un format donné ne soit pas reconnu par certaines liseuses. ePub a été

inventé pour palier ce problème, le but étant de fournir un format ouvert et libre qui pourra être lu sur le plus grand nombre de liseuses possible.

Ce rapport est divisé en trois parties : la première concerne les généralités sur le format ePub, la seconde traite des aspects techniques du format, enfin la troisième présente des outils intéressants liés au format.

1 Découverte du format ePub

1.1 Qu'est-ce que ePub ?

ePub (acronyme de « electronic publication » ou « publication électronique » en français) est un format ouvert standardisé pour les livres électroniques. L'extension associée à ce type de fichier est « **.epub** ».

Il s'agit un format compressé (que l'on peut ouvrir avec un gestionnaire d'archive) qui contient entre autres des pages HTML, XML, des feuilles de style CSS et des images. Nous reviendrons en détail sur le rôle de ces différents fichiers dans la seconde partie de ce rapport.

La norme ePub est libre et ouverte : tous les systèmes d'exploitation disposent d'un logiciel lisant ePub.

Le principal but de ePub est de devenir un standard dans le monde des eBooks.

1.2 Qui est à l'origine de ce format ?

C'est en septembre 2007 que l'IDPF [4] (International Digital Publishing Forum, cf Figure 2) a proposé le format ePub. Basée à Toronto, cette organisation a pour but de développer l'édition électronique. Ainsi, l'IDPF a mené une étude sur des formats et standards qui seraient adaptés à une lecture sur écran d'ordinateur ainsi que sur liseuse. Le format ePub est l'aboutissement de cette réflexion.

Le travail de l'IDPF ne s'arrête pas là. En effet l'organisation fournit également des statistiques sur le marché américain des eBooks et organise des conférences et des forums de discussion sur ces technologies.

Il est possible pour des entreprises ou des associations de devenir membre de l'IDPF moyennant des cotisations allant de 650\$ à 3000\$ par an.

Aujourd'hui, la plupart des membres de cette organisation sont américains.



Figure 2: Le logo de l'IDPF, groupe de réflexion à l'origine d'ePub

1.3 Pourquoi ePub?

Il existe aujourd'hui beaucoup de formats différents d'eBooks. Généralement, chaque constructeur possède son propre format de données. ePub a donc vu le jour pour offrir un format compatible avec toutes les liseuses du marché.

Le format ePub est en train de s'imposer sur le marché comme le nouveau standard de l'industrie. Les éditeurs l'adoptent de plus en plus largement. Des eBooks avec un tel format seront donc compatibles avec le plus grand nombre de plates-formes différentes.

S'il est de plus en plus utilisé par les éditeurs, c'est parce que ePub est vu comme le futur du marché.

1.4 Avantages et limites

Le format ePub présente de nombreux avantages par rapport à ses concurrents.

Tout d'abord, il peut par exemple être remis en forme à la volée, en fonction de l'écran qui l'affiche. Il gère également les notes de bas de page et les tables des matières.

Ensuite, le format ePub est gratuit, libre et ouvert, ce qui permet à tout le monde d'avoir accès aux informations techniques de ce format.

Il est possible de lire des eBooks au format ePub sur de nombreux supports :

- directement sur une liseuse telle que le Reader de Sony ;
- sur l'iPhone avec l'application Stanza [5] ;
- sur un PC avec un logiciel approprié tel Calibre que nous avons utilisé ;
- ou encore en ligne avec une application telle que Bookworm [6].

Enfin, les fichiers ePub ont l'atout de pouvoir être convertis en plusieurs autres formats (le logiciel Calibre est capable de réaliser cette opération). Citons par exemple le format MobiPocket pour la liseuse Kindle.

Toutefois, le format ePub possède quelques limites. Par exemple, les annotations ne sont pas gérées par le format.

En outre, il existe un problème du côté des terminaux mobiles : les petits écrans ne peuvent tout simplement pas afficher la page entière.

Finalement, il n'existe pas à l'heure actuelle de vidéos, d'animations, de Flash ni d'interactivité.

2 Les spécificités techniques d'un fichier ePub

En réalité, un fichier avec l'extension .epub n'est rien d'autre qu'une archive au format .zip. On retrouve dans une telle archive plusieurs composantes ayant chacun un rôle bien précis. Nous nous sommes aidés des informations disponibles sur les sites [7] et [8] en annexe.

2.1 Les standards

Le standard ePub est fondé autour de trois spécifications qui ont chacune un rôle différent. En effet, il est constitué de l'OCF (Open Container Format) pour le conteneur, de l'OPS (Open Publication Structure) pour le contenu et l'OPF (Open Packaging Format) pour les métadonnées.

2.1.1 OCF

L'OCF a pour objectif de spécifier l'organisation des fichiers à l'intérieur de l'archive, c'est à dire l'arborescence. Les deux fichiers utilisés pour respecter cette spécification sont « container.xml » et « mimetype » que l'on décrira par la suite.

2.1.2 OPS

L'OPS indique quels types de fichiers peuvent être utilisés dans un eBook. Concernant le format ePub, il peut y avoir des fichiers HTML, CSS ainsi que des images.

2.1.3 OPF

L'OPF doit décrire les métadonnées requises et optionnelles ainsi que la table des matières. Les fichiers relatifs à cette spécification auront l'extension .opf ou .ncx. Toutes les informations concernant l'emplacement physique des fichiers ainsi que leur ordre sont spécifiés grâce à l'OPF.

2.2 L'arborescence

Concernant l'arborescence, le standard ePub impose qu'il y ait un fichier mimetype ainsi qu'un dossier META-INF à la racine. Dans ce dernier dossier, le fichier container.xml est lui aussi obligatoire. La première chose que fera une liseuse est de lire ces fichiers afin de savoir où trouver les informations utiles pour afficher l'eBook correctement. Le créateur du livre électronique est libre de créer les dossiers qu'il veut ensuite à condition qu'il existe bien un fichier .opf ainsi qu'un fichier .ncx afin de respecter les spécifications décrites précédemment. Par convention, ces fichiers seront présents dans le dossier OEBPS pour Open eBook Publication Structure. Dans ce dossier seront situés le fichier .ncx, le fichier .opf ainsi que le réel contenu du livre (fichiers .html, .css ainsi que les images au format .jpeg, .gif, .png

ou .svg). Pour que l'archive soit bien organisée, on peut mettre ces fichiers dans différents dossiers afin de retrouver rapidement ce que l'on cherche. Le nom de ces dossiers ainsi que celui des fichiers n'est pas imposé car il sera renseigné dans les fichiers .opf et .ncx et le nom de ces derniers est renseigné dans le fichier container.xml.

Voici donc une possibilité d'arborescence (* représente un nombre quelconque de fichiers) :

```
mimetype
META-INF/
    container.xml
OEBPS/
    content.opf
    stylesheet.css
    toc.ncx
    style/
        *.css
    contenu/
        *.html
    images/
        *.png, *.jpeg, *.gif, *.svg
```

2.3 Description des fichiers

Dans cette partie, nous allons décrire un à un les différents fichiers présents dans un eBook au format ePub. Voici donc la liste des fichiers :

- mimetype : ce fichier doit obligatoirement être présent à la racine et en 1ère position dans l'arborescence. Le nom ne doit pas être modifié, sinon le fichier ne sera pas validé par le standard ePub. Ce fichier est seulement constitué de la ligne suivante :

```
application/epub+zip
```

Ce fichier a d'autres restrictions : il ne doit en aucun cas posséder un retour à la ligne à la fin de son nom ou tout autre caractère. Il permet uniquement de décrire le format des données; ici il donne l'information que le fichier est au format ePub;

- container.xml : ce fichier au format XML doit obligatoirement être présent dans le dossier META-INF lui-même situé à la racine de l'arborescence. Ce nom de dossier est obligatoire car les liseuses commencent par lire les informations présentes dans ce dossier. Le fichier a pour but d'indiquer l'endroit où est situé le fichier .opf, c'est pourquoi le nom de ce dernier fichier peut-être choisi par le créateur du livre. Voici un exemple de contenu du fichier :

```
<?xml version="1.0"?>
<container version="1.0"
xmlns="urn:oasis:names:tc:opendocument:xmlns:container">
  <rootfiles>
    <rootfile full-path="OEBPS/content.opf"
      media-type="application/oebps-package+xml" />
```

```
</rootfiles>
</container>
```

Dans ce fichier, seule l'information « OEBPS/content.opf » peut être modifiée, le reste doit être identique quel que soit le livre au format ePub;

- content.opf : ce fichier, présent dans le fichier OEBPS par convention peut posséder le nom que l'on désire et être situé dans le dossier que l'on souhaite, tout comme les fichiers que nous allons citer ci-dessous. Ce fichier sert à indiquer le chemin physique de toutes les données du livre (fichiers HTML, CSS et images) ainsi que celui du fichier .ncx. Voici un exemple de contenu :

```
<?xml version="1.0"?>
<package version="2.0" xmlns="http://www.idpf.org/2007/opf" unique-
identifiant="BookId">
  <metadata xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:opf="http://www.idpf.org/2007/opf">
    <dc:title>Rapport projet</dc:title>
    <dc:language>fr</dc:language>
    <dc:identifiant id="BookId"
opf:scheme="ISBN">theaultlamontagne</dc:identifiant>
  </metadata>
  <manifest>
    <item id="ncx" href="toc.ncx" media-type="text/xml"/>
    <item id="cover" href="titre.html" media-type="application/xhtmll
+xml"/>
    <item id="content" href="contenu.html" media-type="application/xhtmll
+xml"/>
    <item id="cover-image" href="images/couverture.png" media-
type="image/png"/>
    <item id="css" href="stylesheet.css" media-type="text/css"/>
  </manifest>
  <spine toc="ncx">
    <itemref idref="cover"/>
    <itemref idref="content"/>
  </spine>
  <guide>
    <reference href="titre.html" type="cover" title="Cover"/>
  </guide>
</package>
```

Ce fichier est constitué de quatre parties distinctes entourées par les balises suivantes :

- **<metadata>** : cette partie sert à donner quelques informations sur le document dont deux sont obligatoires : le titre du document (<dc:title>) ainsi que son identifiant (<dc:identifiant>) qui à l'obligation d'être unique. D'autres éléments sont facultatifs comme le nom des créateurs (<dc:creator>), la langue utilisée (<dc:language>), la date de publication

(<dc:date>), le copyright (<dc:rights>) ainsi que le nom de la personne qui publie le document (<dc:publisher>). Ces métadonnées se conforment au standard de Dublin Core [9];

- **<manifest>** : cette partie déclare l'ensemble des ressources présentes dans le livre. Tous les fichiers utiles à la représentation du document doivent être indiqués dans cette partie (fichiers HTML, CSS, images). Nous devons également renseigner le type de média pour chaque fichier. Cette partie contient également le lien vers le fichier .ncx;

- **<spin>** : les fichiers listés dans le manifest sont dans le désordre. Spin permet de préciser l'ordre de lecture. Chaque référence présente dans cette partie doit posséder un numéro d'id présent dans la liste des fichiers faite dans le manifest.

- **<guide>** : cette dernière partie est optionnelle mais conseillée ; en effet, elle sert à donner plus d'informations concernant chaque partie. La partie manifest sert à donner le lien physique vers chaque fichier, la partie spin à définir un ordre et la partie guide à donner l'utilité de chaque fichier. Les différentes utilités possibles sont à renseigner dans « type » et peuvent prendre les valeurs suivantes : cover, title-page, toc, index, glossary, acknowledgements, bibliography, colophon, copyright-page, dedication, epigraph, foreword, loi, lot, notes, preface ou text (éléments présents dans le standard OPF 2.0).

- toc.ncx : ce fichier est le dernier concernant les métadonnées. Il reprend beaucoup d'informations contenues dans le content.opf mais ne fait pas partie des spécifications de l'ePub ; en effet, il suit les spécifications du DAISY Consortium. Il a été créé pour fournir une bonne expérience de lecture à ceux qui ne peuvent pas utiliser de livres traditionnels. Voici un exemple de contenu de ce fichier :

```
<?xml version='1.0' encoding='utf-8'?>
<!DOCTYPE ncx PUBLIC "-//NISO//DTD ncx 2005-1//EN"
"http://www.daisy.org/z3986/2005/ncx-2005-1.dtd">
<ncx xmlns="http://www.daisy.org/z3986/2005/ncx/" version="2005-1">
  <head>
    <meta name="dtb:uid" content="theautlamontagne"/>
    <meta name="dtb:depth" content="1"/>
    <meta name="dtb:totalPageCount" content="0"/>
    <meta name="dtb:maxPageNumber" content="0"/>
  </head>
  <docTitle>
    <text>Rapport projet</text>
  </docTitle>
  <navMap>
    <navPoint id="navpoint-1" playOrder="1">
      <navLabel>
        <text>Page de couverture</text>
      </navLabel>
      <content src="titre.html"/>
    </navPoint>
    <navPoint id="navpoint-2" playOrder="2">
      <navLabel>
        <text>Contenus</text>
      </navLabel>
      <content src="contenu.html"/>
    </navPoint>
  </navMap>
</ncx>
```

Ce fichier est constitué de trois parties délimitées par les balises suivantes :

- **<head>** : cette partie contient les informations principales du livre comme l'identifiant (dtb:uid), le nom de niveau différents dans le livre (dtb:depth). Les deux dernières informations (totalPageCount et maxPageNumber) sont importants seulement pour les livres papiers donc ils peuvent être laissés à 0 dans notre cas.

- **<doctitle>** : cette partie contient uniquement le titre du livre;

- **<navMap>** : cette partie est la plus importante car elle permet, comme dans le fichier content.opf de donner la liste des fichiers utiles pour reconstituer le livre (avec leur lien physique), l'ordre dans lequel ils doivent être utilisés mais également une description pour chaque partie. Pour chaque élément dans cette partie, on doit renseigner le lien physique vers le fichier (content), l'ordre dans lequel doit être traité chaque élément (playOrder) ainsi qu'une description du fichier (navLabel/text).

- les fichiers de données : les fichiers restants constituent le contenu réel du livre. Le format utilisé est le XHTML. Ces fichiers doivent être valides et respecter les standards du WEB. Les fichiers au format XHTML servent à donner le contenu texte du livre et peuvent contenir des images dans l'un des format suivant : .png, .jpeg, .gif ou .svg. Pour mettre en forme le livre, il faudra utiliser des feuilles de style au format .css. Tous ces fichiers peuvent être situés dans les dossiers que l'on veut mais il est plus pratique de créer un dossier par type d'éléments (contenu, images, style).

2.4 Création manuelle d'un fichier ePub

Comme nous l'avons vu, un fichier ePub est en fait un ensemble de fichiers compressés au format zip. Pour que l'ePub soit créé correctement, il est indispensable que le fichier mimetype soit le premier fichier de l'arborescence et il ne doit pas être compressé. Pour faire cela, il suffit de taper cette commande sous linux :

```
$ zip -0Xq my-book.epub mimetype
$ zip -Xr9Dq my-book.epub *
```

Nous en avons fini avec la description technique des fichiers contenus dans un livre électronique au format ePub. Dans la partie 3 de ce rapport, nous nous intéresserons aux différents outils disponibles liés à ce standard.

3 Outils

Il existe aujourd'hui de nombreux outils très utiles pour manipuler les fichiers ePub dans quatre domaines distincts que sont la création, la validation, la conversion et la visualisation. Un large panel d'applications est disponible sur le site ebouquin.fr [10].

3.1 Aide à la création

Dans ce projet, nous avons créé manuellement notre document ePub. Chaque fichier a été écrit séparément puis nous avons compressé l'ensemble pour arriver à construire l'eBook. Ce procédé se révèle long et il n'est pas facile d'accès pour un utilisateur ordinaire. C'est pourquoi plusieurs logiciels existent pour créer de manière graphique un document ePub.

Nous avons testé le logiciel « Sigil » [11]. Il s'agit d'une application WYSIWYG (What You See Is What You Get), compatible sur tous les systèmes d'exploitation permettant d'éditer des documents ePub. Il se présente comme un traitement de texte qui est facile d'utilisation et intuitif.

3.2 Validation

Comme pour d'autres standards tel que le HTML ou le CSS, il existe un outil permettant de valider un document ePub [12]. La création d'un fichier valide nous assure que l'expérience de lecture sera la meilleure possible quel que soit le support utilisé.

Notre rapport au format ePub disponible en ligne a été testé sur le validateur officiel et respecte donc bien le standard. Toutefois, nous avons rencontré un problème lors de nos essais : le fichier « mimetype » ne faisait pas une taille correcte. Cela était dû au fait que nous avons compressé ce fichier.

3.3 Conversion de format

Il peut être extrêmement utile d'effectuer des conversions « format quelconque vers ePub » et « ePub vers format quelconque ».

Le logiciel « Calibre » nous semble être le meilleur convertisseur. En effet, il est capable d'effectuer des conversions dans les 2 sens avec un nombre très important de formats :

- Formats d'entrée : CBZ, CBR, CBC, CHM, EPUB, FB2, HTML, LIT, LRF, MOBI, ODT, PDF, PRC, PDB, PML, RB, RTF, TCR, TXT
- Formats de sortie : EPUB, FB2, OEB, LIT, LRF, MOBI, PDB, PML, RB, PDF, TCR, TXT

On remarque ainsi que le ODT est supporté en entrée par Calibre, il nous a donc été possible de convertir ce rapport en ePub directement. Ce procédé peut être très bénéfique en terme de temps pour un utilisateur souhaitant créer un fichier ePub.

3.4 Visualisation

Il existe de nombreuses plate-formes capables de lire des eBooks au format ePub : liseuse, iPhone, logiciel sur PC.

« Calibre » ou « Sigil » présentés précédemment sont capables d'afficher des eBooks. Mais nous avons choisi de vous présenter un outil extrêmement utile : Bookworm. Il s'agit d'une plate-forme libre capable de lire en ligne des fichiers ePub. Nous vous proposons cet outil afin que vous n'ayez rien à télécharger et installer pour lire notre rapport, tout se passe en ligne.

La démarche est simple, rendez-vous à l'adresse suivante : <http://bookworm.oreilly.com/>. Une fois sur le site, cliquez sur « Sign in ». Un login et un mot

de passe vous sont demandés. Nous avons créé un compte qui a pour login et mot de passe le mot : « ensicaen ». Connectez-vous avec ce compte. Vous découvrirez alors une petite liste des livres que nous avons ajouté à la bibliothèque. Sélectionnez notre rapport pour le visualiser.

Conclusion

Ce travail était principalement un projet de découverte. D'un côté, le but était de découvrir les spécificités techniques du standard ePub et de comprendre l'intérêt et les enjeux de l'arrivée de ce format. Il nous a en outre permis d'apprendre à créer un document ePub qui respecte les normes du standard.

Et maintenant, quel avenir pour ePub ?

De nombreux spécialistes s'accordent pour dire que ePub est en passe de devenir le nouveau standard pour les livres électroniques. Certains lui prédisent le même avenir que le très populaire format « mp3 » pour les fichiers audio.

Quelques améliorations sont encore à apporter au format. Dans le futur, nous pouvons attendre des liens entre des fichiers ePub différents, la gestion des annotations ou encore la présence de la signature numérique.

Références bibliographiques

- [1] <http://www.gutenberg.org/etext/16780> : La déclaration d'indépendance des États-Unis, premier livre numérisé.
- [2] SVM "le magazine du numérique" n°287 décembre 2009. "prêts pour la révolution eBook."
- [3] <http://www.ebooksgratuits.com/logiciels.php> : page recensant les différents formats, les logiciels de lecture ainsi que de création d'eBooks dont "Calibre".
- [4] <http://www.idpf.org/> : site officiel de l'IDPF, organisation étant à l'origine du format ePub.
- [5] <http://stanza.softonic.fr/iphone> : application iphone permettant de lire des eBooks.
- [6] <http://bookworm.oreilly.com/> : plateforme libre permettant de lire en ligne des fichiers au format ePub.
- [7] http://www.hxa.name/articles/content/epub-guide_hxa7241_2007.html ; par Harrison Ainsworth : un guide pour la construction d'un document ePub.
- [8] <http://www.ibm.com/developerworks/xml/tutorials/x-epubtut/index.html> : permet d'apprendre à créer un livre au format ePub.
- [9] <http://dublincore.org/> : Site officiel de Dublin Core : schéma de métadonnées générique qui permet de décrire des ressources numériques ou physiques.
- [10] <http://www.ebouquin.fr/2010/02/04/comment-creer-un-fichier-epub/> : site présentant un large choix d'outils liés à epub.
- [11] <http://code.google.com/p/sigil/> : Le site du projet Sigil, une application WYSIWYG compatible Windows, Linux, Mac capable d'éditer des fichiers au format ePub.
- [12] <http://threepress.org/document/epub-validate/> : Le validateur de documents ePub.