CHAPITRE 3

Au cœur d'un fichier ePub

Peu de gens écrivent intégralement les fichiers ePub à partir de zéro, car les ordinateurs sont beaucoup mieux adaptés à cette tâche, quoiqu'ils ne puissent pas tout faire eux-mêmes. La meilleure méthode, jusqu'à ce qu'il existe une solution robuste et complète pour la création de fichiers ePub, consiste à employer un outil pour générer le contenu initial, comme InDesign ou Word, puis un éditeur de texte pour le peaufiner.

Voici les thèmes abordés dans ce chapitre :

- désarchiver un fichier ePub généré par InDesign ou autre, et examiner son contenu ;
- reconnaître et créer les fichiers individuels qui composent un document ePub, y compris ceux du texte et des images du livre numérique, ceux qui servent de base à la table des matières interactive et celui qui recense le contenu de l'ouvrage ;
- préciser le fichier d'image qui servira de couverture au livre numérique ;
- archiver correctement les fichiers individuels dans un document ePub ;
- vérifier la validité du fichier ePub à l'aide d'EpubCheck.

Désarchiver un fichier ePub

Si vous avez créé votre fichier ePub avec InDesign ou si vous souhaitez savoir comment une autre personne a créé son document ePub, vous devez le désarchiver afin d'obtenir les images et les fichiers XHTML et CSS qui le composent.

1. Créez un répertoire dans lequel vous placez le fichier ePub (voir Figure 3.1). Lorsque vous allez le désarchiver, vous allez obtenir un grand nombre de fichiers et il est préférable qu'ils ne soient pas dispersés sur votre bureau.

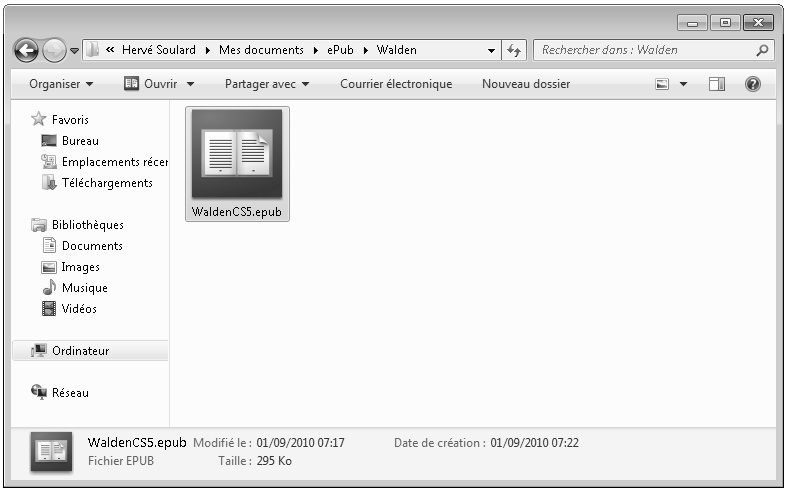

Figure 3.1

Le fichier WaldenCS5.epub *a été placé dans le répertoire* Walden*, qui se trouve dans le répertoire* ePub *du dossier* Mes documents*.*

2. Pour extraire le contenu du fichier ePub, vous avez besoin d'un logiciel de décompression. Pour un ordinateur Windows, vous avez le choix entre de nombreux outils ; nous avons opté pour 7-Zip (**http://7zip.fr/**). Pour un Mac, il suffit d'ouvrir Terminal et de lancer l'utilitaire unzip.
3. Cliquez du bouton droit sur le fichier ePub. Dans le menu qui s'affiche, sélectionnez 7-Zip > Extract Here (voir Figure 3.2). Si vous utilisez un Mac, allez dans le répertoire qui contient le fichier ePub, puis désarchivez le fichier à l'aide de l'utilitaire unzip (la commande est `unzip WaldenCS5.epub`).

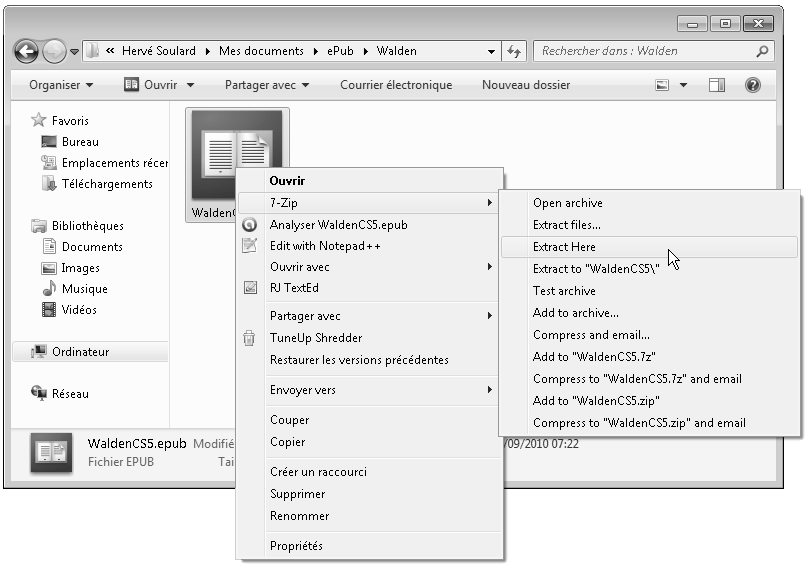

Figure 3.2

Extrayez le contenu du fichier ePub dans le dossier courant.

4. Le contenu du document ePub, c'est-à-dire les fichiers XHTML, CSS, XML et images, est placé dans des sous-dossiers du répertoire courant (voir Figure 3.3).

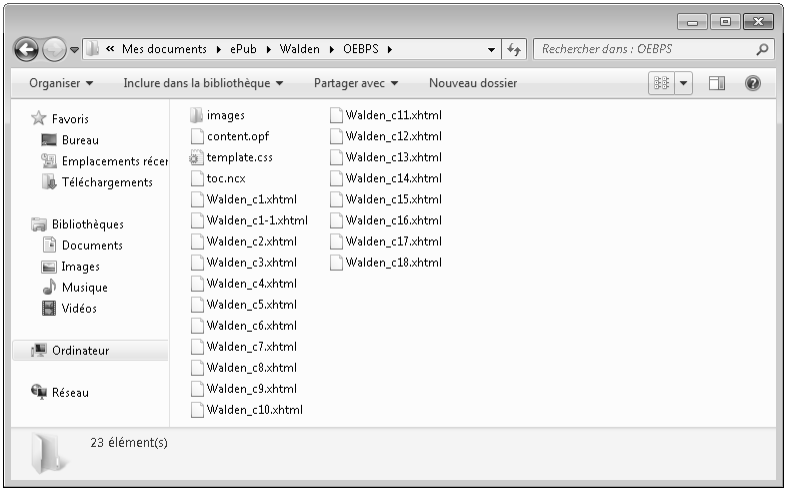

Figure 3.3

Le contenu principal du livre au format ePub se trouve dans le dossier OEBPS.

Astuce

Je préfère placer les trois composants d'un fichier ePub – les dossiers OEBPS *et* META-INF*, ainsi que le fichier* mimetype *– dans un dossier distinct. Ainsi, je peux aller dans ce dossier, apporter les modifications nécessaires, créer une nouvelle archive ePub et procéder à des tests, tout cela sans toucher au document* WaldenCS5.epub *d'origine produit par InDesign.*

Vous pouvez également utiliser l'outil 7-Zip pour examiner l'intérieur du fichier ePub et modifier le contenu des fichiers XHTML et CSS sans extraire l'intégralité de l'archive.

En remplaçant l'extension du fichier ePub par .zip*, vous pouvez extraire son contenu à l'aide des outils intégrés à Windows. Si vous utilisez un Mac, il vous suffit alors de double-cliquer sur le fichier. En revanche, vous aurez toujours besoin d'un outil d'archivage pour reconstituer le fichier ePub.*

Les fichiers d'un document ePub

Si vous utilisez InDesign pour générer le fichier ePub, vous pouvez extraire son contenu pour ensuite examiner ou modifier les fichiers qu'il contient (voir la section précédente). Si vous utilisez Word ou si vous écrivez à la main le contenu du livre numérique, vous devez organiser manuellement le contenu du document ePub. Dans tous les cas, il est préférable de connaître le rôle et l'emplacement de chaque fichier.

Dans l'exemple, nous avons créé un livre numérique pour l'ouvrage *Walden* de Henry David Thoreau. Sa structure est illustrée à la Figure 3.4.

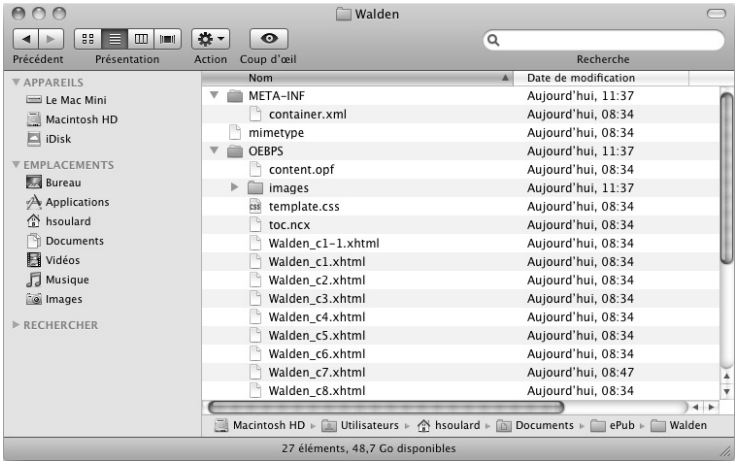

Figure 3.4

La structure d'un document ePub vue sur un Mac.

Pour modifier les fichiers d'un document ePub, vous avez besoin d'un éditeur de texte qui dispose de préférence d'une fonction de recherche fondée sur les expressions régulières. Cela vous permettra de gagner beaucoup de temps lors de l'adaptation du contenu du fichier ePub. Vous pouvez vous servir de l'éditeur fourni avec votre ordinateur, par exemple TextEdit sur le Mac ou Notepad sous Windows, mais il existe des outils beaucoup plus performants, notamment pour la manipulation des fichiers XHTML, comme BBEdit (**http://www.barebones.com**) ou TextMate (**http://macromates.com**) pour le Mac, et Notepad++ (**http://notepad-plus-plus.org**) ou RJ TextEd (**http://www.rj-texted.se**) pour Windows. Surtout, n'utilisez pas Microsoft Word car il mettra la pagaille dans votre code HTML.

Le fichier *mimetype*

Le fichier *mimetype* est un simple fichier texte, dont l'unique ligne contient `application/epub+zip`, sans retour chariot ni passage à la ligne. Puisqu'il ne varie pas d'un document ePub à l'autre, vous pouvez simplement le copier et le coller dans le dossier de chaque livre numérique. Vous pouvez utiliser celui fourni avec les exemples téléchargeables depuis le site web consacré à cet ouvrage.

Le dossier *META-INF*

Le dossier *META-INF* comprend un seul fichier, *container.xml*, dont voici le contenu :

```
<?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>
```

La seule valeur que vous pourriez avoir à modifier est celle de l'attribut `full-path` de l'élément `rootfile`. Elle doit désigner le fichier *content.opf* du document ePub. Si vous placez toujours ce fichier dans le dossier *OEBPS*, comme le préconisent les conventions et comme nous le faisons dans cet ouvrage, vous pouvez garder le même fichier *container.xml* pour chaque document ePub.

Si vous incorporez des polices dans le fichier ePub, InDesign ajoute le fichier *encryption.xml* dans le dossier *META-INF*. Il est également utilisé lorsque vous appliquez des DRM au livre numérique. Pour de plus amples informations concernant le chiffrement, consultez les spécifications OPS à l'adresse **http://www.idpf.org/ocf/ocf1.0/download/ocf10.htm#_Ref8795282**. Cet ouvrage ne s'intéresse pas à la mise en œuvre des DRM sur les livres numériques.

Le dossier *OEBPS*

C'est dans le dossier *OEBPS* que se trouve le contenu du livre numérique ; OEBPS est l'acronyme de *Open eBook Publication Structure*. Bien que vous puissiez le nommer simplement « livre », à condition de corriger l'attribut `full-path` dans le fichier *container.xml*, les conventions préfèrent *OEBPS*.

Quoi qu'il en soit, le dossier *OEBPS* comprend tous les fichiers XHTML dans lesquels se trouve le texte de votre ouvrage, un fichier CSS pour les instructions de mise en forme, un dossier pour les images, un dossier pour les polices incorporées et deux fichiers XML spéciaux, *toc.ncx* et *content.opf*. Le premier, *toc.ncx*, est utilisé par les liseuses pour générer la table des matières interactive. Le second, *content.opf*, contient les métadonnées sur l'ouvrage ainsi qu'une liste des fichiers de contenu.

Fichiers XHTML et CSS

Les fichiers de contenu qui se trouvent dans le dossier *OEBPS* peuvent apparaître dans n'importe quel ordre, mais il est probablement plus simple de commencer par le contenu de l'ouvrage rédigé dans le langage XHTML 1.1 et mis en forme avec des règles CSS. XHTML est une variante plus stricte de HTML, le langage des pages web. CSS est le langage utilisé pour mettre en forme le contenu d'un document XHTML.

Les documents XHTML d'un ouvrage au format ePub doivent commencer par des déclarations XML et DOCTYPE appropriées et être enregistrés selon un encodage UTF-8 :

```
<?xml version="1.0" encoding="utf-8"?>
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN"
  "http://www.w3.org/TR/xhtml11/DTD/xhtml11.dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
```

Après ces informations initiales sur le document XHTML, vous trouvez la section d'en-tête qui contient toutes les métadonnées à propos du contenu lui-même, notamment le titre du document et l'emplacement de la feuille de style :

```
<head>
  <title>Walden_c1.xhtml</title>
  <link href="template.css" rel="stylesheet" type="text/css" />
</head>
```

Le titre d'un fichier XHTML *ne détermine pas* le titre de l'ouvrage tel que présenté dans les liseuses ou en ligne (dans cet exemple, le titre a été généré par InDesign et correspond au nom du fichier XHTML). Ce rôle est assuré par le document *content.opf*. L'élément `link` précise l'emplacement du fichier de la feuille de style CSS qui servira à la mise en forme du document (dans cet exemple, *template.css*).

Après la section d'en-tête vient le corps du document. En voici un exemple, tiré du premier chapitre de *Walden* :

```
<body>
  <div id="walden-c1">
    <div class="generated-style">
      <h1 class="titre" id="toc-anchor">Économie</h1>
      <p class="premier-paragraphe"><span class="lettrine">Q</span>
        <span class="petites-capitales">uand </span>j'ai écrit les pages
       suivantes, ou plutôt la plupart d'entre elles, je vivais seul...</p>
      <div class="group">
        <div class="generated-style">
       <img src="images/Thoreau_Maison_fmt.jpeg" alt="Thoreau_Maison.jpg"/>
        </div>
        <div class="generated-style-2">
         <a id="la-maison-de-thoreau-au-bord-de-l-tang-de-walden-anchor"/>
          <p class="legende" xml:lang="fr">La maison de Thoreau au bord de
          l'étang de Walden.</p>
        </div>
      </div>
      <p class="corps">Je n'imposerais pas tant mes affaires...</p>...
      <p class="citation">Inde genus durum sumus, experiensque...</p>...
    </div>
  </div>
</body>
</html>
```

Voici quelques remarques à propos de ce document :

- Il existe quatre classes d'éléments `p` (`premier-paragraphe`, `legende`, `corps` et `citation`) afin que les paragraphes correspondants puissent avoir une mise en forme différente.
- Le fichier se nomme *walden_c1.xhtml* et se trouve dans le dossier *OEBPS*.
- Cette version raccourcie de *walden_c1.xhtml* correspond au premier chapitre de l'ouvrage. Lorsque chaque chapitre du livre est placé dans un fichier séparé, la plupart des liseuses les commencent sur une nouvelle page. Il s'agit de la meilleure manière de créer un saut de page.
- Vous pouvez également créer des documents XHTML séparés pour la couverture (voir section « Créer la couverture » plus loin dans ce chapitre), pour les pages liminaires (pages de titre, table des matières, préface, etc.) et pour les parties annexes (index et publicités, par exemple).

Les techniques de mise en forme des ouvrages numériques sont détaillées au Chapitre 4.

Le fichier *toc.ncx* de la table des matières interactive

Le fichier *toc.ncx* est l'un des deux fichiers dans lesquels les liseuses trouvent les informations concernant le livre numérique. Il s'agit d'un simple fichier texte encodé au format UTF-8 et placé dans le dossier *OEBPS*. Son nom n'est pas important, pourvu qu'il ait l'extension *.ncx* (elle signifie *Navigation Control file for XML*). Les liseuses se servent de ce fichier pour créer la table des matières interactive de l'ouvrage (voir Figure 3.5) ; je distingue cette table des matières de celle qu'il est possible de trouver dans le contenu principal de l'ouvrage et dont l'utilité est sans doute moindre.

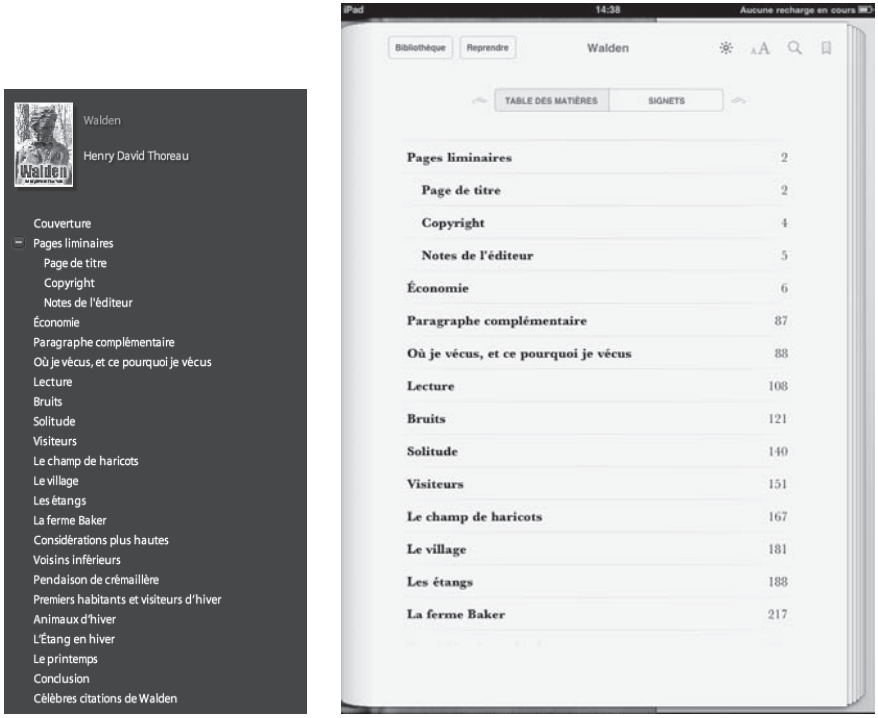

Figure 3.5

La table des matières interactive de l'ouvrage, avec deux niveaux hiérarchiques (à gauche dans Adobe Digital Editions, à droite sur l'iPad).

Si vous utilisez InDesign pour créer le fichier ePub, vous pouvez lui demander de générer automatiquement une table des matières interactive – un fichier *toc.ncx* – à partir des titres de votre document. Pour de plus amples informations, consultez la section « Créer une table des matières interactive » au Chapitre 2. En général, vous

n'apporterez aucune modification à ce fichier, mais, si vous souhaitez ajouter une entrée à la table des matières, suivez les instructions données ci-après. Si le contenu a été créé à la main ou avec Word, vous pouvez également construire ce fichier manuellement.

Le fichier *toc.ncx* doit être un document XML valide et commencer par le balisage suivant :

```
<?xml version="1.0"?>
<!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">
```

Vient ensuite la section `head`, qui doit comprendre quatre éléments `meta` nommés `dtb:uid`, `dtb:depth`, `dtb:totalPageCount` et `dtb:maxPageNumber`, même si seuls les deux premiers ont un intérêt avec les livres numériques.

La valeur de `dtb:uid` correspond à l'identifiant d'utilisateur et doit être une chaîne de caractères ou un code unique au livre numérique. Par exemple, les éditeurs peuvent employer l'ISBN de l'ouvrage. Si un numéro ou un code n'est pas déjà associé au document ePub, générez votre propre identifiant unique universel (*uuid*) et utilisez-le ; rechercher « générateur uuid » dans un moteur de recherche pour trouver l'outil approprié. Les conventions veulent que l'uuid soit identifié en le préfixant par `urn:uuid:`. Si InDesign a produit le document ePub, il a généré automatiquement cet identifiant à votre place.

La valeur de `utb:depth` fait référence au nombre de niveaux et de sous-niveaux de la table des matières. L'exemple se fonde sur deux niveaux ; le niveau 1, réservé aux pages liminaires, aux annexes et aux titres des chapitres, et le niveau 2, aux sous-sections de certains chapitres (généralement les pages liminaires et les annexes).

Les deux derniers éléments `dtb` concernent uniquement les ouvrages papier et leur valeur peut être fixée à zéro, mais ils doivent être présents.

```
<head>
  <meta name="dtb:uid" content="3bf2928d-962f-9990-2463-feab747875af"/>
  <meta name="dtb:depth" content="2"/>
  <meta name="dtb:totalPageCount" content="0"/>
  <meta name="dtb:maxPageNumber" content="0"/>
</head>
```

La deuxième partie importante du fichier *toc.ncx* réside dans l'élément `docTitle` et son élément `text` imbriqué. Bien qu'il doive préciser le titre du livre, j'ai pu constater qu'aucune liseuse ne s'en servait. Il est toutefois préférable de lui donner la valeur appropriée :

```
<docTitle>
  <text>Walden</text>
</docTitle>
```