Creating Dictionaries

Contents

Introduction to Dictionaries

A dictionary is a Kindle eBook (MOBI file) with extra tags added to support search and lookup functionality. Dictionary eBooks:

- Contain a primary index: a list of words or sentences that are sorted in alphabetical order. Readers can search quickly in this list by typing the beginning of the word and selecting the desired entry.
- Are marked as dictionaries. The input and output languages of the dictionary must be defined properly so that Kindle devices can use the dictionary for in-book lookup.

For example, an English (monolingual) dictionary lists English as both the input and output language. A French-English dictionary lists French as the input language and English as the output language. To build a bidirectional bilingual dictionary (example: Spanish-French and French-Spanish), you must create two separate eBooks: one for Spanish-French and one for French-Spanish.

A Kindle dictionary should have all the same components as a normal Kindle eBook. There should be an OPF file and HTML files with CSS. Every dictionary should have:

- A cover image
- A copyright page
- Any relevant front or back matter (explanations of symbols, appendices, etc.)
- Definitions of words (this is the bulk of the file)

This format doesn't currently support Enhanced Typesetting.

Metadata Guidelines

The OPF file of a dictionary is similar to that of other Kindle books, except that it contains specialized metadata tags in the <x-metadata> section. These extra tags in the OPF file set the source language and the target language for the dictionary. If the dictionary has multiple indices, the OPF file also specifies the name of the primary lookup index.

- The <DictionaryInLanguage> element contains the ISO 639-1 language code for the language of the books this dictionary is designed to be used on. For a Spanish-French dictionary, the input language is Spanish.
- The <DictionaryOutLanguage> element contains the ISO 639-1 language code for the language of the definitions returned by the dictionary. For a Spanish-French dictionary, the output language is French.
- The <DefaultLookupIndex> element indicates the index that will open first when the dictionary is used for lookup from another eBook. The default index must be specified if the dictionary has more than one index. The index name that is wrapped in the <DefaultLookupIndex> tags in the OPF file also should appear as the value of the name attribute in the <idx:entry> elements in the content of the dictionary (see details on HTML).

As an example, for a Spanish-French dictionary, the input language code would be es; the output language code would be fr, and the primary index might be named Spanish. See a list of country codes.

Example: (Bilingual Dictionary Metadata)

```
<x-metadata>

<DictionaryInLanguage>es</DictionaryInLanguage>

<DictionaryOutLanguage>fr</DictionaryOutLanguage>

<DefaultLookupIndex>Spanish</DefaultLookupIndex>

...

</x-metadata>
```

For a monolingual dictionary, the same language code must appear twice: once to identify the input language, and again to identify the same language as the output language. To identify a regional variant for the source and/or target languages, a regional suffix may be appended to the ISO 639-1 code. For example, en-gb indicates British English, while en-us indicates US English.

Example: (Monolingual Dictionary Metadata, Regional Variant)

```
<x-metadata>

<DictionaryInLanguage>en-us</DictionaryInLanguage>

<DictionaryOutLanguage>en-us</DictionaryOutLanguage>

<DefaultLookupIndex>headword</DefaultLookupIndex>
```

```
...

</x-metadata>
```

Text Guidelines

A simple, clean format works best for in-book lookup. Amazon recommends these dictionary content and formatting features for a high-quality user experience:

- The headword (word being defined) should come first in the entry, and should be distinguished from surrounding content (on its own line, flush left, in bold).
- Every dictionary entry should contain a definition (or translation, for bilingual dictionaries).
- Horizontal rules should appear between each entry.
- Each alphabet letter section should begin on a new page.
- Images should be avoided (see image constraints).
- Tables should not be used (see table constraints).
- Font color, size, and typeface should not be forced (see text guidelines).

Basic Dictionary HTML

Format

Dictionaries for Kindle must be in MOBI 7 format, not in KF8. For this reason, the dictionary layout should use a single-column format. Multiple columns and sidebars are not supported in MOBI 7 format.

Frameset Element

All dictionaries must have an <mbp:frameset> element as the first child of the <body> element. This frameset element contains all of the <idx:entry> elements of the dictionary.

The namespace for this <mbp:frameset> element is

xmlns:mbp="https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf" and it must be declared in the root <html> element of the XHTML document.

Example:

```
<html xmlns:math="http://exslt.org/math" xmlns:svg="http://www.w3.org/2000/svg"
xmlns:tl="https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf"

xmlns:saxon="http://saxon.sf.net/" xmlns:xs="http://www.w3.org/2001/XMLSchema"
```

```
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"

xmlns:cx="https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf"
xmlns:dc="http://purl.org/dc/elements/1.1/"

xmlns:mbp="https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf"
xmlns:mmc="https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf"
xmlns:idx="https://kindlegen.s3.amazonaws.com/AmazonKindlePublishingGuidelines.pdf">

<head><meta http-equiv="Content-Type" content="text/html; charset=utf-8"></head>

<body>

<mbp:frameset>

<idx:entry name="english" scriptable="yes" spell="yes">

<idx:short><a id="1"></a>

<idx:orth value="aardvark"><b>aard•vark</b>

<idx:infl>

<idx:iform value="aardvarks"></idx:iform>

<idx:iform value="aardvark's"></idx:iform> <idx:iform value="aardvarks'"></idx:iform>

</idx:infl>

</idx:orth>

<p> A nocturnal burrowing mammal native to sub-Saharan Africa that feeds exclusively on ants and
termites. </p>

</idx:short>

</idx:entry>

[...other entries…]

</mbp:frameset>

</body>

</html>
```

Headword Index

In order to make an alphabetical index of headwords, it is necessary to use special tags that are not standard HTML. The source still will be valid XHTML with these added <idx> mark-ups.

<idx:entry>..</idx:entry>

The <idx:entry> tag marks the scope of each entry to be indexed. In a dictionary, each headword with its definition(s) should be placed between <idx:entry> and </idx:entry>. Any type of HTML may be placed within this tag.

The <idx:entry> tag can carry the name, scriptable, and spell attributes. The name attribute indicates the index to which the headword belongs. The value of the name attribute should be the same as the default lookup index name listed in the OPF. The scriptable attribute makes the entry accessible from the index. The only possible value for the scriptable attribute is "yes". The spell attribute enables wildcard search and spell correction during word lookup. The only possible value for the spell attribute is "yes".

Example:

```
<idx:entry name="english" scriptable="yes" spell="yes">
```

The <idx:entry> tag also may carry an id attribute with the sequential id number of the entry. This number should match the value of the id attribute in an anchor tag used for cross-reference linking:

Example:

```
<idx:entry name="japanese" scriptable="yes" spell="yes" id="12345">

<a id="12345"></a>
```

The entry id number is not used for in-book lookup; instead, the wordform entity to be indexed for lookup must be contained in the <idx:orth> element as described in the following sections.

<idx:orth>..</idx:orth>

The <idx:orth> tag is used to delimit the label that will appear in the index list and that will be searchable as a lookup headword. This is the text that users can enter in the search box to find an entry.

Example:

```
<idx:orth>Label of entry in Index</idx:orth>
```

Here is an example of an extremely simple entry that could be part of an English dictionary. From this example code, the word "chair" would appear in the index list and would be searchable by users.

Example:

```
<idx:entry>
```

```
<idx:orth>chair</idx:orth>

A seat for one person, which has a back, usually four legs, and sometimes two arms.

<idx:entry>
```

The value attribute can be used on the <idx:orth> tag to include a hidden label in the entry. This attribute maintains lookup functionality in the presence of the special formatting that commonly appears on headwords in dictionaries.

Example:

```
<idx:orth value="Hidden Label of entry in Index">Display format</orth>
```

If the headword should be displayed in the dictionary with a superscripted number to indicate homographs, with a registered trademark symbol, with middle dots to separate syllables, or with any other added symbols, this special formatting should appear on the text between the <idx:orth> tags, but not on the text in the value attribute. The text in the value attribute should match exactly the form to be used for lookup. If a value attribute is not supplied, then the entity between the <idx:orth> tags will be indexed for lookup. If middle dots, superscripted numbers, or any other symbols are included in the text between the <idx:orth> tags, then in-book lookup will fail unless a hidden label with the lookup form is supplied in the value attribute.

Example:

```
<idx:orth value="Amazon">A•ma•zon®<sup>3</sup></orth>
```

If the dictionary uses more than one orthographic script, then the format attribute on the <orth> tag can be used to identify each script for building the index.

Example:

```
<idx:orth format="script name">
```

Along with this primary index of headwords for all entries in the dictionary, in-book lookup also requires a supplementary index of inflected forms for each headword. To build the hidden inflection index, additional data should be nested within the <idx:orth> tag as follows.

Inflections for Dictionaries

Dictionaries should be built so that multiple inflected forms of a single root word all access the same entry. A complete list of inflected wordforms should be provided for every headword. If an entry uses multiple orthographies, then separate inflections must be provided for each orthography.

Inflection Index

To construct the hidden inflection index, the inflected wordform data should be wrapped within <idx:infl> and <idx:iform /> tags nested inside the <idx:orth> element. This index will not be directly searchable by the user, but instead will be used for in-book lookup.

<idx:infl>..</idx:infl>

The <idx:infl> element may contain multiple <idx:iform /> elements. The <idx:iform /> elements are always empty elements, and are used only to carry attributes, not visible content. The value attribute indicates the inflected forms that make up the inflection index.

Example:

```
<idx:orth>record
<idx:infl>
<idx:iform value="records" />
<idx:iform value="recording" />
<idx:iform value="recorded" />
</idx:infl>
</idx:orth>
```

The <idx:infl> tag, the <idx:iform /> tag, and the value attribute are mandatory. The <idx:infl> element also may carry an optional inflgrp attribute to denote part of speech, and the <idx:iform /> element may carry an optional name attribute to indicate the inflection paradigm category. For languages that use extensive inflection, including these optional categories will expand the size of the inflection index and may result in slower performance during word lookup.

Example:

```
<idx:orth>record
<idx:infl inflgrp="noun">
<idx:iform name="plural" value="records" />
</idx:infl>
<idx:infl inflgrp="verb">
<idx:iform name="present participle" value="recording" />
```

```
<idx:iform name="past participle" value="recorded" />
<idx:iform name="present 3ps" value="records" />
</idx:infl>
</idx:orth>
```

The values listed as attributes of the <idx:iform /> tag will be invisible to the user, but rather will provide the information needed to redirect from inflected forms to the associated headwords during inbook lookup. To inform the user about parts of speech or inflection paradigms, additional text should be included in the body of the entry (i.e., alongside the definition and examples).

<idx:key>..</idx:key> (DEPRECATED)

Like the <idx:infl> tag, the <idx:key> tag is designed to enable search for an entry in the index by means of an alternative lookup wordform. However, the presence of <idx:key> tags in a Kindle dictionary can create instability in the lookup functionality and can interfere with the operation of the exact-match parameter. For these reasons, the use of <idx:key> tags in Kindle dictionaries is deprecated. Instead, <idx:infl> and <idx:iform /> tags should be used to wrap the alternative lookup forms.

Exact-match Parameter

By default, the Kindle device uses a fuzzy algorithm for matching diacritics during word lookup. Languages that use contrastive diacritics to distinguish between distinct word forms should use the exact="yes" attribute in the <idx:iform /> tag to force exact match of diacritics during lookup.

Example:

```
<idx:entry name="spanish" scriptable="yes" spell="yes">
<a id="12345"></a>
<idx:orth value="uña"><b>uña</b>
<idx:infl>
<idx:iform value="uñas" exact="yes" />
</idx:infl>
</idx:orth>
```

Setting the exact parameter to "yes" forces the device to match uñas to the headword uña ("fingernail") and prohibits a match to una ("one").

Building a Dictionary with KindleGen

When building a dictionary with KindleGen via the command line, use the following syntax:

```
kindlegen.exe [filename.opf] -c2 –verbose -dont_append_source
```

If the dictionary entries are contained in a single, very large XHTML file, then KindleGen may not be able to build the dictionary. If the dictionary fails to build, this problem may be resolved by splitting the dictionary content into two or more XHTML files.

See guidance on using KindleGen.

Testing Kindle Dictionaries

Format Testing

Amazon recommends verifying that the converted dictionary is properly formatted to provide a good visual experience for the user:

- Check the formatting of the definitions by paging through the dictionary and reading several definitions. (The format of the dictionary may be checked using Kindle Previewer or any Kindle device; however, lookup testing requires the use of an E-reader device.)
- Check words for unsupported characters, broken or joined words, proper display of accented characters, symbols, pronunciation guide, etc.
- Check that there are no typos.
- Check that links (if present) are working correctly. (Links will be disabled in the in-book lookup window, but links should function inside the dictionary itself.)
- If any images are used, check that these images are clear and readable.
- Check that the font color and typeface are not forced.

Lookup Testing

Amazon recommends verifying that definitions return correctly when the dictionary is used to look up words in other books. This component of testing can be done only with E-reader devices (not including Previewer), because only E-reader devices allow the user to set the default dictionary for lookup.

- Sideload the dictionary onto the E-reader device. To do this, connect the Kindle to your computer with a USB-to-mini-USB cord. Your computer should detect the device. In the window that pops up, you should see a folder called Documents. Put the dictionary file into this folder and then eject your Kindle from the computer.

- Set the test dictionary as the default dictionary for lookup:
    - Kindle Paperwhite: Go to Home > Menu > Settings > Device Options > Language and Dictionaries > Dictionaries > [Source Language]
- Look up a variety of words to see what definition is returned. Open a title other than the dictionary, select a word, and note the definition returned in the lookup window. If lookup fails entirely, check for errors in the HTML tagging. Suggestions of words to look up include:
    - Conjugations of regular and irregular verbs
        - Example: walk, walks, walked, walking; go, goes, went, gone, going
    - Nouns, adjectives, adverbs and their conjugations/declensions
        - Example: desk, desks; wolf, wolves; hot, hotter, hottest
    - Grammatical and punctuation conventions commonly used in the language
        - Example: contractions, elisions, verbs with clitic pronouns
- Check the index view of the dictionary. To do this, open the dictionary and start typing a word in the Search box. An alphabetized list of headwords should appear and should update dynamically based on which letters are typed. Selecting a headword from the index list should redirect the user to the dictionary entry for that headword.