Name

herold — HTML to DocBook converter

Synopsis

`herold` [OPTIONS]

Description

The reuse of HTML content in presentation-neutral form is a frequent problem. One possible solution is to convert HTML to DocBook XML, because DocBook is a semantic markup language for documentation, which enables its users to create document content that captures the logical structure of the content. The command line tool herold can be used to convert HTML to DocBook. Because HTML elements are often used not as intended, the possibilities for such a transformation are somewhat limited. herold is part of the dbdoclet suite of tools. For more information visit http://www.dbdoclet.org.

Options

--docbook-add-index, -x
Automatically add an index element at the end of the document.

--docbook-decompose-tables, -T
Decomposes the tables from the HTML code into single paragraphs. This can be useful, if a document contains a lot of tables for formatting reasons.

--docbook-encoding, -d
Specifies the encoding of the generated DocBook XML files.

--docbook-root-element, -r
The root element of the document. Possible values are: book, article and reference. The default value for this option is 'article'

--docbook-title, -t
The title for the resulting document.

--in, -i
Specifies the HTML input file.

--help, -h
Prints a help page on the console.

--html-encoding, -s
Specifies the encoding of the HTML source files, such as ISO-8859-1.

--out, -o
Specifies the DocBook XML destination file.

--profile, -p
A profile file with predefined settings.

--verbose, v
Enables the verbosity for the console output.

--version, -V
Displays the version of herold.

Configuration

The details of a transformation can controlled by a profile file. A profile offer more possibilities to influence the transformation than the command line arguments. The following example shows a typical profile.file.

```
 1: transformation html2docbook;
 2:
 3:
 4: input {
 5: }
 6:
 7: output {
 8: }
 9:
10: section HTML {
11:     encoding = "windows-1252";
12:     exclude = [ "//p[starts-with(@class, 'MsoToc')]", "" ];
13:     section-numbering-pattern = "(((\d\.)+)?\d?\.?\p{Z}*).*";
14: }
15:
16: section DocBook {
17:     add-index = true;
18:     create-xref-label = false;
19:     decompose-tables = false;
20:     document-element = "book";
21:     encoding = "UTF-8";
22:     image-data-formats = [ "gif", "base64" ];
23:     title = "Tutorial";
```

```
24:     use-absolute-image-path = false;
25: }
```

Syntax

A profile file consists mainly of sections. Sections are used to group parameters which share the same context. Every section must start with the keyword `section` followed by the name of the section. After the name comes the block of parameters, which is surrounded by curly braces. Parameters can be of type String, Number, Boolean or Array. Strings must be framed with double quotes, Arrays with square brackets. Inside an array, the elements must be comma separated. Every assignment must be finished by a semicolon. Multi line comments have the form `/* my comment */` , single line comments look like `// my comment\n`.

Mandatory Elements

A profile for herold must start with the line `transformation html2docbook;`. After this line the two mandatory sections, `input` and `output`, are following. These sections can be used to define fixed input and out files. Use the param file to define a path inside these sections, eg file = "./index.html";. Normally input and output files are set via command line arguments.

Section HTML

The section HTML defines parameters, which control the loading and parsing of the HTML input data.

encoding
: The character set used to read the input stream.

exclude
: Defines an array of xpath expressions. All matches are removed from the HTML DOM tree before transformation.

section-numbering-pattern
: Normally you want to get rid of the section numbering that comes with the HTML data, because it becomes part of the title text in DocBook.

The section numbers will the appear twice in your target media. One from HTML and one from the DocBook XSL processing. The parameter section-numbering-pattern defines a regular expression, which is matched against the beginning of every section title. If it matches, this part is removed.

Section DocBook

add-index
If set to true, an index element is inserted at the end of the DocBook XML.

create-xref-label
if set to false, anchor elements doesn't get a xreflabel attribute.

decompose-tables
If set to true, tables structures will be ignored. The content of the table cells will be inserted into the DocBook XML as a sequence of paragraphs. This parameter can be useful if your HTML contains tables for formatting purposes. Normally you want to get rid of them, because they tamper the logical structure.

document-element
The document element you want to use. Must be one of article, book, part or reference.

encoding
The character set which will be used for writing the output file.

image-data-formats
An array of image formats. These formats will be inserted as imageobject elements, additionally to the format found in the src attribute of the corresponding img element. The original format is inserted twice with the roles "html" and "fo". The other formats are inserted as "html-<FORMAT>" and "fo-<FORMAT>".

title
The title of the resulting document. If this parameter is undefined, herold tries to dected the title from the head section of the HTML data.

use-absolute-image-path

If you want absolute image paths in the fileref attribute of the imagedata element, set this parameter to true.

Copyright