TRANSFORMING XML AND HTML DOCUMENTS TO PDF USING CSS

PART 4 – ADVANCED FUNCTIONALITIES

Julien Lacour, Syncro Soft

julien_lacour@oxygenxml.com

Oxygen Webinar, September 15th 2021

syncro soft

XML Editor

Overview

- Hyphenation
- PDF Bookmarks
- PDF Metadata
- Text Extraction
- Dynamic Links
- Change Bars

Poll

What is your level of CSS knowledge?

Poll

What document type are you using as PDF input?

Poll

Are you familiar with XPath?

Before we start...

- "Le Tour du monde en quatre-vingts jours" (Around the World in Eighty Days) by Jules Verne

- Text and images are from Project Gutenberg

  https://www.gutenberg.org/ebooks/46541

- CSS has been extracted into separate style-sheets

What is hyphenation?

- Can be manual or automatic
    - Use of the HYPHEN or SHY characters
    - Use of a hyphenation dictionary

- Uses TeX typesetting system

- Relies on hyphenation dictionaries

How to use Hyphenation

- Use CSS 'hyphens' property

- Use the 'auto' value (dictionaries)

```
body {
    hyphens: auto;
}
```

https://developer.mozilla.org/en-US/docs/Web/CSS/hyphens

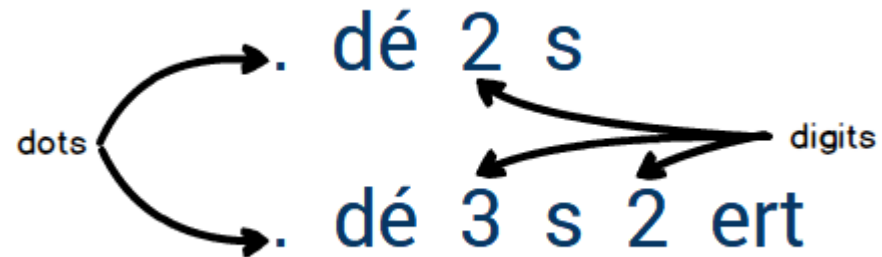# How does a hyphenation dictionary work?

- Uses @lang or @xml:lang attribute

- Uses .hyp files (compiled dictionaries)

- Defines patterns containing
  - Some characters
  - One or more digits (0-9)
  - (Optional) One or two dots

- Creates hyphenation points for wrapping words

How does a hyphenation dictionary work?

- Pattern:
    - Sets of characters are matched against the words
    - Dots are used to mark the starting/ending of a word

- Match (for each word):
    - Substring is assigned a digit 0-9 between the characters (hyphenation point)
    - Pattern conflict: the highest value is assigned

- Digits:
    - Odd value: hyphenation is allowed
    - Even value: no hyphenation is allowed

Example of Hyphenation

- The french hyphenation dictionary

- The 'désert' word

What is a PDF bookmark?

- Additional way of navigating (similar to TOC)

- Automatically created for HTML documents

- Can be customized from CSS

How to customize bookmarks

- Use of the dedicated properties
  - bookmark-level
  - bookmark-label
  - bookmark-state
  - bookmark-target

- Use of CSS content(), counter(), attr() and other functions

  https://www.w3.org/TR/css-gcpm-3/#bookmarks

What is PDF metadata?

- Additional information about the document
  - Dynamic: collected from the document
  - Static: specified from the CSS directly

- Automatically created from HTML <meta> tags

- External tools for filtering, indexing, or navigating

- Accessible by search engines

How to define PDF metadata

- Use Oxygen's custom properties:
    - -oxy-pdf-meta-author
    - -oxy-pdf-meta-custom
    - -oxy-pdf-meta-description
    - -oxy-pdf-meta-keyword
    - -oxy-pdf-meta-keywords
    - -oxy-pdf-meta-title

- Can be used with content() and attr() CSS functions

How to extract text from the content

- Use Oxygen's oxy_xpath() function

- Supports all XPath 2.0 functions

- Works on both @page CSS at-rule and classic selectors

- Can be tested using the XPath Builder

  Warning: Must be optimized: can cause performance issues!

How to define PDF metadata

- From <meta> tag:

```
<meta name="keywords" content="Phileas Fogg, Passepartout, Londres, Inde" />
```

- From oxy_xpath() function:


```
oxy_xpath("normalize-space(substring-before(substring-after(text(), 'of'), ',')))")

<title>The Project Gutenberg eBook of Le tour du monde en quatre-vingts jours, by Jules Verne</title>
```

Le tour du monde en quatre-vingts jours

How to generate dynamic links

- Use CSS target-counter() function

- Gets a counter value for a given URL

- URL target must be inside the document

Example of dynamic links

How to display modifications

- Use Oxygen's custom ::changebar pseudo-element

- Use Oxygen's custom properties:
    - -oxy-changebar-offset
    - -oxy-changebar-placement
    - -oxy-changebar-style
    - -oxy-changebar-color
    - -oxy-changebar-width

```
ins::changebar {
    -oxy-changebar-placement: start;
    -oxy-changebar-offset: 0.1in;
    -oxy-changebar-style: solid;
    -oxy-changebar-color: silver;
    -oxy-changebar-width: 1.5pt;
}
```

Additional Links

XPath documentation:

https://developer.mozilla.org/en-US/docs/Web/XPath/Functions/normalize-space

https://developer.mozilla.org/en-US/docs/Web/XPath/Functions/substring-before

https://developer.mozilla.org/en-US/docs/Web/XPath/Functions/substring-after

Oxygen PDF Chemistry User-Guide

https://www.oxygenxml.com/doc/versions/23.1/ug-chemistry/topics/ch_advanced_styling_hyphenation.html

https://www.oxygenxml.com/doc/versions/23.1/ug-chemistry/topics/ch_pdf_output.html#ch_bookmarks

https://www.oxygenxml.com/doc/versions/23.1/ug-chemistry/topics/ch_pdf_output.html#ch_metadata

https://www.oxygenxml.com/doc/versions/23.1/ug-chemistry/topics/ch_advanced_styling_using_xpath_in_css.html

https://www.oxygenxml.com/doc/versions/23.1/ug-chemistry/topics/ch_cross-references.html#ch_internal_links

https://www.oxygenxml.com/doc/versions/23.1/ug-chemistry/topics/ch_change_bars.html

THANK YOU!

Any questions?

Julien Lacour

julien_lacour@oxygenxml.com