Getting ebooks on Wikisource: a first step to a Semantic Digital Library

Thomas Pellissier Tanon   Thibaut Horel

Wikimania

July 13, 2012

Outline

Introduction

- Current state
  - 4th project of the Wikimedia Foundation (number of edits)
  - 17,000 books
  - $\simeq$ 1 million proofread pages

- Problems
  - books not easily accessible
  - no real bibliographic database

Introduction

- Current state
  - 4th project of the Wikimedia Foundation (number of edits)
  - 17,000 books
  - $\simeq$ 1 million proofread pages

- Problems
  - books not easily accessible
  - no real bibliographic database

Outline

WSexport

What can it do?

- export Wikisource books to EPUB
- extract books metadata

Key principle: a specific microformat

How does it work?
a set of html classes added to the author and book templates

Book template

| title = The Original Fables of La Fontaine
| author = Jean de La Fontaine

Specification:
`http://wikisource.org/wiki/Wikisource:Microformat`

Key principle: a specific microformat

How does it work?
a set of html classes added to the author and book templates

Generated HTML

```
<span id="ws-title">The Original Fables of La Fontaine </span>
<span id="ws-author">Jean de La Fontaine </span>
```

Specification:
`http://wikisource.org/wiki/Wikisource:Microformat`

Key principle: a specific microformat

How does it work?

a set of html classes added to the author and book templates

Exported metadata

<dc:title>The Original Fables of La Fontaine </dc:title>

<dc:creator>Jean de La Fontaine </dc:creator>

Specification:

`http://wikisource.org/wiki/Wikisource:Microformat`

Collection

Isn't the Collection extension already enough?

- no support for Wikisource specific features (advanced transclusion,...)
- no support for custom metadata
- no need to create a page for each book: it already exists

Examples

- `http://wsexport.fr.nf`
- "Citer ce livre" *(quote this book)* on the French Wikisource
- BookMaker gadget: Collection + Wsexport

Outline

Need of an internal system for bibliographic metadata

- organizing and navigating the library
- interacting with other libraries

Index pages

- metadata about books (title, authors, . . .)
- one index page = one physical book

But: no easy way to process this data

Standard

- internal set of properties for a book
- map properties to labels (language-dependent)
- map properties to bibliographic standards (dublin core, ...)

What it made possible

New API to support the OAI-PMH protocol

Libraries identifiers for bidirectional links:
- import libraries metadata into Wikisource
- links to Wikisource in the libraries records

Conclusion

- minimal effort in terms of standardization and coordination
- huge gain in terms of interoperability and accessibility