Převod webových stránek s publikacemi do formátu BibTeXML

Petr Zemek

Aneb jak jsem řešil projekt do předmětu ISJ...

25.4.2008

Zpracování vstupu

- získání vstupních dat ze zadaného URL (`urllib`)
- překódování do unicode (utf-8)
- uložení autora z `<title>` a `<h1>`
- „vykousnutí“ publikací ze stránky
  - BibTEX
  - HTML
    - `<p>`, `<td>`, `<li>`
    - `<tag1>` ... autor ... rok ... `<tag2>`
  - Text bez HTML značek - oddělovač `\n\n`
- výstupem je seznam řetězců obsahujících (zřejmě) publikace

Parsing publikací

- „rozkouskování“ řetězce podle obvyklých oddělovačů (`,;:"'`) a HTML značek (`<b>D., T., Cabrero</b>`)
- převod HTML entit na dané znaky (`BeautifulSoup`)
- postupné procházení seznamu částí rozděleného řetězce
  - pořadí údajů (autor, titulek, místo publikace, ...)
  - formát (isbn, issn, počet stránek, rok, ...)
  - klíčová slova (autor, seznam zemí, měsíců, ...)
- oddělení autorů pomocí `and`
- kontrola „legality“ a určení typu publikace
- výstupem pro každý řetězec je slovník s klíči dle BibTeXu

Výstup do formátu BibTeXML

- `xml.dom.minidom` ... klasika
- přiřadit každé publikaci unikátní ID
- každý slovník s publikací → `entry` element

```
<?xml version="1.0" encoding="utf-8"?>
<file xmlns="http://bibtexml.sf.net/">
    <entry id="695115626">
            <inproceedings>
                <author>Miguel A. Alonso</author>
                ...
            </inproceedings>
    </entry>
</file>
```

Díky za pozornost!