UNIVERSITE LIBRE DE BRUXELLES 19 Avril 2012
Faculté des Sciences Appliquées

INFO-H-509 : Technologies XML
Project : XSLT

Introduction

DBLP[1] is an online bibliographical database for computer science publications containing around 1.6 million references. Its content is publically available in XML format[2]. Since this content is more than 800 MB a small excerpt of this data will be used for this project (see the course's webpage for this excerpt).

In essence, the `dblp.xml` file describes a collection of DBLP records:

```
<?xml version="1.0" encoding="ISO-8859-1">
<!DOCTYPE dblp SYSTEM "dblp.dtd">
<dblp>
  record 1
  record 2
  ...
  record n
</dblp>
```

Each record in the collection describes a bibliographic reference of a publication. These publications can be of 8 types: `article`, `inproceedings`, `proceedings`, `book`, `incollection`, `phdthesis`, `mastersthesis`, and `www` following the ad-hoc BibTeX format.

Depending on the publication type, records can mention the following fields: `author`, `editor`, `title`, `booktitle`, `pages`, `year`, `address`, `journal`, `volume`, `number`, `month`, `url`, `ee`, `cdrom`, `cite`, `publisher`, `note`, `crossref`, `isbn`, `series`, `school`, and `chapter`. Notice that not all fields are allowed in all publication types; please refer to the DBLP DTD file `dblp.dtd` and the DBLP description paper at `http://dblp.uni-trier.de/xml/docu/dblpxml.pdf` for detailed information on the different publication types and their allowed fields.

Assignment

The goal of this project is to write a single XSLT 2.0 stylesheet that generates, from the `dblp-excerpt.xml` file, a number of HTML files that together collectively emulates part of the DBLP website[3].

Concretely, for each distinct person name P found in a `author` or `editor` field of some record in the `dbpl.xml` file the stylesheet should generate a file

`a-tree/`*`first-letter-of-lastname`*`/`*`lastname.firstname`*`.html`

For example, for the author "David Maier", the file `a-tree/m/Maier.David.html` should be created, whereas for "Michael Ley", the file `a-tree/l/Ley.Michael.html` should be created. Blanks should be mapped to underscores; all other characters which are not alphanumeric should be mapped to '='. This avoids illegal filenames/URLs.

The contents of the HTML file for person P should consist of the following parts. See `Maier.David.html` on the course's webpage for an illustrating example[4]

1. The person's name in a `h1` tag.
2. Followed by a table of all of the person's publications, grouped per year (sorted descending on year and subsequently on publication title). Apart from the rows that indicate the beginning of a new year, each row in this table should be of the form

   publication number, *link to online version*, *publication reference*

   See the `Maier.David.html` example for more details (some useful explanations are embedded as comments in the HTML source).

[1] `http://www.informatik.uni-trier.de/~ley/db/`
[2] `http://dblp.uni-trier.de/xml/`
[3] Use the `xsl:result-document` command to generate the multiple output files.
[4] Note that "David Maier" is not present in the DBLP excerpt.

3. Followed by "Co-author index" in a `h2` tag.

4. Followed by a table listing all other persons with whom P has jointly published. For each such *co-author*, there should be a row describing the pair (*co-author-name*, *list-of-references-to-joint-publications*). Each reference in this list should link to the corresponding publication in the publication table. The table should be sorted by co-author lastname. See the `Maier.David.html` example HTML source code for more details.

Modalities

Write a *single* XSLT 2.0 stylesheet `generate-author-pages.xslt` that generates the required html files. The required source file (`dblp-extract.xml`) as well as the files with background information (`dblp.dtd`, and `dblpxml.pdf`) may be found on the course's website.

As with the the first assignment, this second assignment contributes 2/20 to the overall grade (there is one more assignments to follow, for another 2/20 points). The written exam contributes the remaining 14/20 points.

This assignment *must* be made in groups of two persons. You are asked to send, per group, the names of the group members to Mr. Julien Roland (`juroland@ulb.ac.be`) by April 24 at the latest. If you cannot find a partner, please indicate so by sending an email to Mr. Julien Roland, who will hook you up with a partner.

You are asked to submit, per group, a small report containing all the hypotheses that you have made during your design, as well as the XSLT stylesheet. The report has to be sent to Mr. Julien Roland in paper (2N3.211C, or in his rack). Furthermore, the stylesheet itself has to be sent by e-mail to `juroland@ulb.ac.be` **no later than Tuesday, May 8 2012**.