

INFO-H509: XML and Web Technologies

Assignment 2: XSLT

Introduction

DBLP¹ is an online bibliographical database for computer science publications containing around 1.6 million references. Its content is publicly available in XML format². 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).

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 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/dblp.xml.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 website³.

Concretely, for each distinct person name P found in a `author` or `editor` field of some record in the `dblp.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 the attached `Maier.David.html` for an illustrating example⁴

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/>

³Use the `xsl:result-document` command to generate the multiple output files.

⁴Note 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.

Important! Please ensure that all links that you generate correspond correctly to the generated file’s name, *including capitalization*. For example, if the generated file name is **Maier.David.html** then having a link `` will only work if your filesystem is case insensitive. Your solutions will be tested on a *case-sensitive* platform. Hence, please be particularly attentive to all to your paths in your solutions. Any incoherence in the case of your paths that leads to an error on our testing environment will cost you points.

Submission

Write a *single* XSLT 2.0 stylesheet **generate-author-pages.xslt** that generates the required html files. The required source file (**dblp-extract.xml**), the files with background information (**dblp.dtd** and **dblp.xml.pdf**), and an XSLT tool may be found on UV.

This assignment contributes 5/20 to the overall grade.

This assignment is to be made in groups of two persons. You are asked to form the groups via the activity “Groups for Assignment 2” on UV.

You are asked to submit, per group, a short report (in English) containing all the hypotheses that you have made during your design, as well as the XSLT stylesheet.

This report and all the required documents have to be uploaded as a zip file to “Assignment 2” on UV by **April 30, 2021**.