INFO-H-509 : Technologies XML
TP 4 – XQuery

Professor : Stijn Vansummeren
Teaching Assistant : Michaël Waumans
http://cs.ulb.ac.be/public/teaching/infoh509

2015 - 2016

DBLP is an online bibliographical database for computer science containing around 1 million references. Its content is publically available in XML format. Since this content is more than 400 MB, a small excerpt of this data will be used for this exercice.

The DBLP collection follows the BibTex format and contains the following types of references : article, improceedings, proceedings, book, incollection, phdthesis, masters-thesis and www.

The fields describing the above types of references are the following : author, editor, title, booktitle, pages, year, address, journal, volum, number, month, url, ee, cdrom, cite, publisher, note, crossref, isbn, series, school and chapter. Notice that not all fields are allowed in all reference types ; please refer to the DTD file for this information.

Define in XQuery the following queries. For each query, the output format is specified.

Exercise 1.1

Give the types of publications in the file.

```
<publications>
  <type>article</type>
  <type>book</type>
  <type>incollection</type>
  <type>inproceedings</type>
  <type>mastersthesis</type>
  <type>phdthesis</type>
  <type>proceedings</type>
</publications>
```

Exercise 1.2

Give the number of publications of each type.

```
<publications>
  <reference type="article">
    <number>263</number>
  </reference>
  <reference type="book">
    <number>9</number>
  </reference>
```

```
  ...
</publications>
```

Exercise 1.3

Give the list of author's names.

```
<authors>
  <name>A. A. Majid</name>
  <name>A. B. M. Shawkat Ali</name>
  <name>A. Bernoussi</name>
  ...
</authors>
```

Exercise 1.4

Give the number of authors.

```
<nb_authors>1478</nb_authors>
```

Exercise 1.5

Give the names of authors who are also editors.

```
<authors_editors>
  <name>Ana Paiva</name>
  <name>Chris George</name>
  <name>Du Zhang</name>
  <name>Ed Dawson</name>
  ...
</authors_editors>
```

Exercise 1.6

Give the number of publications by author.

```
<author>
  <name>Ahmed Sameh</name>
  <number>3</number>
</author>
...
```

Exercise 1.7

Give the authors ordered by the number of publications, in descending order.

```
<author>
  <name>Christoph Meinel</name>
  <number>47</number>
</author>
<author>
  <name>Dieter Baum</name>
  <number>21</number>
</author>
...
```

Exercise 1.8

Give the author(s) having the highest number of publications.

```
<prolific_authors>
  <author>
  <name>Morshed U. Chowdhury</name>
  <number>5</number>
  </author>
</prolific_authors>
```

Exercise 1.9

Give for each author the number of publication types.

```
<author>
  <name>Mazeyar E. Makoui</name>
  <pubtypes>1</pubtypes>
</author>
...
```

Exercise 1.10

Give for each author the total number of publications and the number of publications by type.

```
<author>
  <name>A. B. M. Shawkat Ali</name>
  <total>3</total>
  <pubtype type="inproceedings">
    <number>3</number>
  </pubtype>
</author>

<author>
  <name>A. Bernoussi</name>
  <total>1</total>
  <pubtype type="article">
    <number>1</number>
  </pubtype>
</author>
...
```

Exercise 1.11

Give the list of proceedings that have at least one editor that is also author of at least one article in the proceedings.

```
<proc_editor_author>
  <title>Proceedings of the International (...)</title>
  <title>Advanced Data Mining and Applications,  (...)</title>
  <title>Proceedings of the 5th International  (...)</title>
</proc_editor_author>
```

Exercise 1.12

Define a function that pretty-prints the references of type article in HTML format.

```
<html>
  <head>
    <meta http-equiv="Content-Type" content="text/html; charset=utf-8"/>
    <title>DBLP Articles</title>
  </head>
  <body>
    <h2>DBLP Articles</h2>
    <p>P. Berthon, C. B. Williams. Stages of e-democracy: (...)</p>
    ...
  </body>
</html>
```