UNIVERSITE LIBRE DE BRUXELLES 7 Mars 2014
Faculté des Sciences Appliquées

INFO-H-509 : Technologies XML
Projet 2 : XSLT

Introduction

DBLP [1] est une base de données bibliographique en ligne pour les publications relevant des Sciences Informatiques. Celle-ci contient 1.6 millions de références. Son contenu est disponible au format XML [2]. Vu que sa taille dépasse les 800 Mb, seule une petite portion de ces données seront utilisées pour ce projet. (Voir page web du cours).

Le fichier `dblp.xml` décrit une collection d'entrées DBLP :

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

Chaque entrée de la collection décrit une référence bibliographique d'une publication. Ces publications peuvent être de 8 types : `article`, `inproceedings`, `proceedings`, `book`, `incollection`, `phdthesis`, `mastersthesis`, et `www` suivant le format ad-hoc BibTex.

En fonction du type de publication, les entrées peuvent mentionner les champs suivants : `author`, `editor`, `title`, `booktitle`, `pages`, `year`, `address`, `journal`, `volume`, `number`, `month`, `url`, `ee`, `cdrom`, `cite`, `publisher`, `note`, `crossref`, `isbn`, `series`, `school`, et `chapter`. Il est important de noter que tous les champs ne sont pas autorisés pour chaque type de publication ; veuillez vous référer au fichier DTD de DBLP `dblp.dtd` et l'article décrivant DBLP (`http://dblp.uni-trier.de/xml/docu/dblpxml.pdf`) pour des informations détaillées sur les différents types de publication et les champs associés autorisés.

Projet

Le but de ce projet est d'écrire une stylesheet XSLT 2.0 qui génère, à partir du fichier `dblp-excerpt.xml`, un certain nombre de fichiers HTML qui, ensemble, émulent une partie du site web DBLP. [3].

Concrètement, pour chaque nom de personne distinct P trouvé dans un champ `author` ou `editor` des entrées du fichier DBLP `dbpl.xml`, la stylesheet devra générer un fichier :

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

Par exemple, pour l'auteur "David Maier", le fichier `a-tree/m/Maier.David.html` devra être créé, là où, pour "Michael Ley", le fichier `a-tree/l/Ley.Michael.html` devra être créé. Les espaces " " devront être convertis en underscores "_" ; tous les autres caractères non alphanumériques devront être remplacés par "=". Ceci permettra d'éviter des noms de fichiers ou URLs illégaux.

Le contenu du fichier HTML pour la personne P contiendra les informations suivantes. Consultez le fichier `Maier.David.html` sur la page du cours pour avoir un exemple. [4]

1. Le nom de la personne dans un tag `h1`.
2. Suivi d'une table de toutes les publications de cette personne groupées par année (et triées par ordre décroissant sur l'année et ensuite par ordre croissant sur le titre). Mis à part les lignes indiquant le début d'une nouvelle année, chaque ligne de la table devra être de la forme suivante :

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

1. `http://www.informatik.uni-trier.de/~ley/db/`
2. `http://dblp.uni-trier.de/xml/`
3. Utilisez la commande `xsl:result-document` pour générer plusieurs fichiers.
4. Notez que "David Maier" n'est pas présent dans le fichier DBLP fourni.

Regardez le fichier d'exemple `Maier.David.html` pour plus de détails. (Ce fichier contient quelques explications utiles sous la forme de commentaires.

3. Suivi par "Co-author index" dans un tag `h2`.
4. Suivi par une table listant les autres personnes avec qui P a écrit une publication. Pour chacun de ces *co-auteurs*, une ligne devra décrire la paire (*co-author-name*, *list-of-references-to-joint-publications*). Chaque référence dans la liste devra être liée à la publication correspondante dans la table des publications. La table devra aussi être triée par nom de famille de co-auteur. Regardez le fichier d'exemple `Maier.David.html` pour plus de détails.

Modalités

On vous demande d'écrire *une stylesheet XSLT 2.0* `generate-author-pages.xslt` qui génère les fichiers HTML requis. Le fichier source (`dblp-extract.xml`) ainsi que les autres fichiers mentionnés (`dblp.dtd`, et `dblpxml.pdf`) peuvent être trouvés sur la page du cours.

Tout comme lors des deux premiers projets, celui-ci contribuera à 2 points sur 20 de la note finale, l'examen écrit vaudra par conséquent 14 points sur 20.

Ce projet doit être réalisé par groupe de deux personnes. Il vous est demandé d'envoyer, par groupe, les noms des membres de votre groupe à Mr. Michaël Waumans (`mwaumans@ulb.ac.be`) pour le **21 Mars** au plus tard. Si vous ne parvenez pas à trouver un partenaire de travail, veuillez contacter Mr. Michaël Waumans par e-mail afin qu'il puisse vous indiquer un coéquipier.

Il vous est demandé de rendre, par groupe, un petit rapport (en Français ou en Anglais), contenant l'ensemble de vos hypothèses ainsi que la XSLT stylesheet. Chaque fichier devra être clairement documenté.
Le rapport ainsi que tous les documents requis doit être rendu à M. Michäel Waumans par email à l'adresse `mwaumans@ulb.ac.be` **au plus tard pour Lundi, 12 Mai, 2014.** (Veuillez mentionner comme sujet de votre e-mail : "INFO-H-509 - Projet 2")