Computer Science 400W

Update this Assignment

View 29 submitted assignments

Build a system to create a mirror of an Open Archive and automatically create metadata for records based on PDF source data files.

This assignment requires that you implement the following:

- Harvest data from an OAI-PMH repository. Harvesting must be either scheduled (e.g., using cron) or an explicit operation (e.g., running an application). Consecutive harvesting operations must be incremental.
- Extract metadata for each record harvested, based on the digital object linked to it, and store the new richer record in a local repository.
- Publish the local repository as an OAI data provider.

Implementation notes:

- You may use any publicly-available tools as the basis for the first and third parts of the assignment (see www.openarchives.org). The extraction of metadata must be done using original code in conjunction with PDF tools to extract text and metadata (e.g., pdftotext).
- Your system must have the ability to easily use a different baseURL - this will be used for testing.
- Use the baseURL http://simba.cs.uct.ac.za/~hussein/cgi-bin/OAI-XMLFile-2.2/XMLFile/dl05/oai.pl for sample data. The hidden test suite will be an archive in the same format.
- Ignore multiple metadata formats - only use Dublin Core.
- Ignore sets - assume the source and mirror archives are completely "flat".

You may work in groups of up to THREE students. You must submit a file to Moodle containing your code/package. External documentation must be in the form of a README file included with the code. Internal documentation must be in the form of source code comments and file headers. No marks will be awarded for documentation, but marks WILL be subtracted for missing documentation (up to 30%).

The assignment will be marked as follows:

- Correctness of harvesting implementation (20%)
- Correctness of data provider implementation (20%)
- Basic metadata extraction - from pdftotext's HTML conversion (20%)
- High quality metadata extraction - using heuristics (15%)
- Reasonable results with hidden data set (15%)
- Packaging of all code for rapid installation/deployment (10%)

This will be assessed during a demonstration and interview with each group of students. You will be asked to:
a) (if possible) install the code from a package on either a FreeBSD or Windows system (of your choosing) with a clean non-root/non-Administrator account
b) harvest data from the sample baseURL to show that harvesting works
c) use the Repository Explorer's automatic tests (re.cs.uct.ac.za) to show that the OAI

Due date: Friday, 8 April 2005, 07:00 PM

Upload a file (Max size: 10MB)

Browse...

Upload this file

You are logged in as Hussein Suleman (Logout)

CSC400W