rclis in vision and reality

Thomas Krichel

2005-10-22

rclis

- rclis stands for Research in Computing and Library and Information Science.
- It is pronounced as “reckless”.
- rclis is a clone of the RePEc project.

what is RePEc?

- RePEc is a aggregator of about 500 different archives. It predates OAI-PMH and XML.
- The archives provide information about documents, including URLs for the full text.
- Some archives describe working papers (usually free) some others published papers.

RePEc is a library

- RePEc does not have a user service.
- Instead the data that it collects is freely available to third parties to build user services.
- There are a range of such services, we can’t go through this here.

RePEc is relational

- RePEc builds a relational database between four types of entities
  - documents
    - working papers
    - articles
  - document collections
    - journals
    - archives
  - people
  - institutions

institution registration

- This is done in RePEc through one single person, Christian Zimmermann.
- His initial idea was to compile a list of all the departments that had a web presence.

author registration

- Done through the RePEc author service.
- Authors contact the service to say what papers they have written. This type of service was pioneered by RePEc.
- Author registration is crucial for feedback of performance statistics to the author.
- With support from OSI, a generic software is being written. This software is known as ACIS.

technical innovation

- RePEc is built on attribute: value templates.
- rclis is built on a purpose built format called the Academic Metadata Format.
- I set up this format. It is tailor-made to suit the needs of rclis and RePEc.
- There is some usage of AMF in RePEc
  - – RePEc OAI interface
  - – ernad, the software feeding NEP, a RePEc service.

using already existing resources

- There is already a very large computer science bibliography called DBLP, see http://dblp.uni-trier.de
- The data has no abstracts. It has some full-text links, mainly to toll-gated sites.
- I have done work to convert parts of it to AMF.
- I am now searching if free full text versions of the papers exist anywhere on the Web. This is the Konz project.

the Konz project

- Current state
  - – I use Google API to search of titles.
  - – I examine responses and download pages.
  - – I scan the pages for PDF and Word files.
  - – I examine the text in the file to find the title.
- Limitations
  - – pdf and word full text
  - – conference paper data still being processed
  - – significant hardware and disk problems.

DoCIS

- Konz currently finds 35k papers with free versions out of the paper out of a 140k searched. Not particularly exiting.
- This data is integrated with DBLP AMF data and the result forms a new service called DoCIS.
- DoCIS lives at http://wotan.liu.edu/docis

DoCIS service

- DoCIS is implemented in mod_perl with swish++ and therefore very fast.
- The web pages are written by XSLT scripts directly from the AMF data.
- The service is available to copy from the web, I am more than happy to run it on other sites.
- But the most interesting thing are the service principles.

construction transparency

- DoCIS is an open digital library service because it allows users to inspect exactly how the service runs
  - DoCIS is built using open source software.
  - There is a special interface http://wotan.liu.edu/strip/docis/ that allows to see almost all internal file. Non visible files are specially documented.
- The hope is that it may be used for teaching purposes.

transportability

- Everything in DoCIS is built is such a way that it should be easy to move the service somewhere else and establish copies.
- The ideas may not make a lot of technical sense but it should increase to non-proprietary nature of the system.
- Note that this has not been tested.

usage transparency

- All usage is logged and the logs are made public.
- This it is hoped that it could be used for digital library research.
- Ways will be found to aggregate usage on different physical installations.

open digital service

- DoCIS is an example for a new type of service where the source code of the library is openly.
- It is an open library service.
- This contrasts favorably with the black box approach of the commercial search engines.

E-LIS in rclis

- E-LIS should export its data to rclis.
- In fact Zeno and I plan to work on this, but it may take a little while.
- If it is in rclis, it will also be in DoCIS.
- Deduplication will not be done at this stage.

Julio’s data

- Julio has data that he has collected on a lot of LIS publications.
- The data itself has some technical problems.
- But it still is our best hope to get a good coverage of LIS metadata.
- It currently has its own user service, DoIS.

to do list

- finish a version of konz that recognizes HTML full text
- make Julio and E-LIS data available
- open institutional registration for rclis
  - some work already done by Tom Wilson for LIS departments
  - work on computer science departments will not be so easy.
- open author registration for rclis

Am I crazy?

- Money does not make the world go round. Ideas do.
- When I started to work on RePEc a totally free and improved A&I dataset in 1993, nobody gave it a high probability to succeed.
- There is no reason we can not do it again.

collaboration is welcome!

http://openlib.org/home/krichel

Thank you for your attention!