CLIR and Digital Libraries

RAFFAELLA BERNARDI

UNIVERSITÀ DEGLI STUDI DI TRENTO

P.ZZA VENEZIA, ROOM: 2.05, E-MAIL: BERNARDI@DISI.UNITN.IT

Contents

1. Multilingual services

- Datawords `http://www.datawords.it/index.php`

- Lion Bridge `http://it-it.lionbridge.com/`

[1]   cerca altro

2. Multilingual Libraries

[Jiangping Chen 2009] "Information Access across Languages on the Web: From Search Engines to Digital Libraries". DL with multilingual access:

Library Name	URL	Languages
Meeting of Frontiers	http://frontiers.loc.gov/intldl/mtfhtml/mfsplash.html	English/Russian
in	http://international.loc.gov/intldl/fiahtml/fiahome.html	English/French
Parallel Histories	http://international.loc.gov/intldl/eshtml/	English/Spanish
International Children's Digital Library	http://www.icdlbooks.org/	Digital Objects in 11 languages. Users can do the keyword search in 51 languages.
The Perseus Digital Library	http://www.perseus.tufts.edu	Greek, English, Latin

2.1. Multi-lingual but not Cross-language

The above digital libraries share the following characteristics:

- They have been funded by various funding agencies, especially from the federal government;

- They are the products of collaboration. People from different countries work together to produce the bilingual or multilingual collections;

- They serve a broader or global user community in which users speak different languages;

- They do not employ cross-language search, cross-language information retrieval techniques or machine translation.

2.2. Wishes in 2009

Digital library (DL) developers should:

1. Collaborate with researchers in CLIR and MT As digital objects are more organized than Web pages crawled by search engines, it is possible that better performance of machine translation could be achieved through the construction of a customized knowledge base for machine translation software.

2. Collaborate with DL developers in other countries to increase the languages that the users can access.

3. Collaborate with the users. In the current digital age, even monolingual digital libraries are accessed by people who dont know the language (Sorid, 2009). Social computing has been widely used on the Internet, and it can play a big role in involving users in the multilingual information access services: users may even volunteer to translate digital objects into another language.

4. Take a step-by-step approach. DL developers can first implement a multilingual interface, then the metadata, and then the whole collection.

2.3. Cross-Language DL

- http://www.zbw.eu/e_catalogues.htm

2.4.   iCLEF

iCLEF (interactive Cross Language Evaluation Forum): http://nlp.uned.es/iCLEF/

> A central research question is how best to assist users when searching information written in unknown languages, rather than how best an algorithm can find information written in languages different from the query language.

Since iCLEF 2008: focus on the shared analysis of a medium/large search log from a single search interface provided by the iCLEF organizers (i.e. our current focus is on log analysis, rather than on system design).

FlickLing has been the search interface provided by iCLEF organizers consisting of a basic cross-language retrieval system for the Flickr image database. We presented it as an online game with a clear goal:

> the user is given a target image, and she must find it again without any a priori knowledge of the language(s) in which the image is annotated.

2.5. CLIR in DL in US

☑   altro?

2.6. CLIR in DL in EU

3   aggiungi Europeana

2.7. CLIR in DL in IT

AquaBrowser:

- Camera di Commercio di Roma Promoroma `http://www.aquabrowser.it/PROMOROM`

- Istituto e Museo di Storia della Scienza `http://biblioteca.imss.fi.it/indice.html`

- Consiglio Regionale Puglia – Teca del Mediterraneo `http://www.aquabrowser.it/CRP2/`

- Consiglio Regionale della Toscana – Rete CoBiRe `http://www.aquabrowser.it/COBIRE3/`

- ARPA Lazio – Biblioteca Ambientale "Paolo Colli" `http://www.aquabrowser.it/ARPA_LAZIO/`
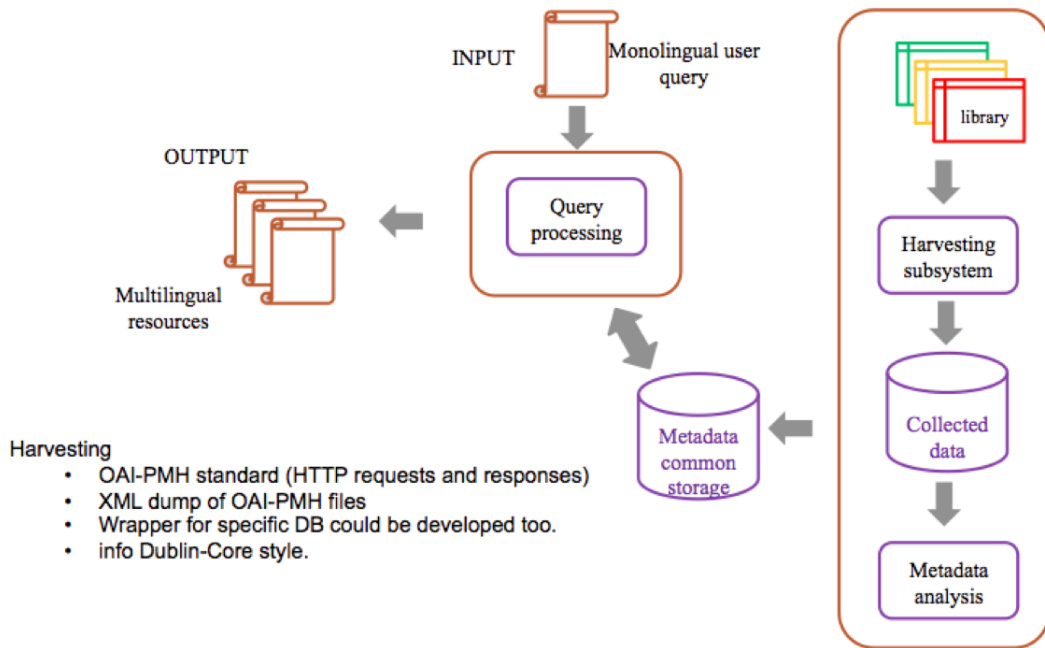
- Pontificia Università Lateranense `http://aquabrowser.pul.it/LAT`

2.8. EU project: CACAO – CLS

Cross language Access to CAtalogue and Online libraries (CACAO). (2008-2010)
CLS (http://www.cross-library.com/) now in Trento.

- LT CELI modules for metadata and query analysis. On test FBK modules

- Lucene search engine: indexing and ranking; currently using boolean search

- Query translation (English, Italian, French, German and other languages)

2.9. CACAO architecture

INPUT — Monolingual user query

OUTPUT — Multilingual resources

Query processing

library

Harvesting subsystem

Collected data

Metadata analysis

Metadata common storage

Harvesting
- OAI-PMH standard (HTTP requests and responses)
- XML dump of OAI-PMH files
- Wrapper for specific DB could be developed too.
- info Dublin-Core style.

2.10. CACAO query analysis