EDLproject: challenges of multilingual access to multilingual European content

Maja Žumer
University of Ljubljana
Slovenia

1. Background

EDLproject is a Targeted Project funded by the European Commission under the eContentplus Programme, within the area of Cultural content and scientific/scholarly content. It is coordinated by the German National Library. EDLproject builds on the existing The European Library, a service funded by CENL, the Conference of European National Librarians, providing unified access to the electronic resources of the main European National Libraries as well as to other library services. The project is also a continuation of the TEL-ME-MOR project, which has supported The European Library with the inclusion in the service of the ten New Member States National Libraries. EDLproject integrates the bibliographic catalogues and digital collections of the National Libraries of Belgium, Greece, Iceland, Ireland, Liechtenstein, Luxembourg, Norway, Spain and Sweden, into The European Library: by the end of 2007 ALL EU countries will be members of the European Library service. EDLproject further enhances access to the European Library portal, by continuing to develop its multi-lingual capacity. EDLproject leverages the influence and resources of CENL as a key player and stakeholder in the content field to work towards a consensual resolution of certain issues raised by the Communication "i2010: Digital Libraries", such as potential availability of digital content from national libraries and the scope for collaboration between The European Library and other content providers funded by eContentplus.

EDLproject has a total budget of € 2.114 million EUR, of which € 1 million contribution by the eContentplus programme. The project started in September 2006 and will last for 18 months.

2. Multilinguality

The research in this area started in the TEL-ME-MOR project. From the end-user viewpoint there are several layers of multilinguality:

- Language of the interface
- Language of bibliographic records/cataloguing language/language of metadata
- Language of the resources

Ideally, each user should be able to:

- Use his/her own language when communicating with the system
- Use his/her own language to formulate a query and, as the result, retrieve all relevant (digital) objects in any language

The first part, the interface in all languages, has been solved. The portal interface has been translated into all partner languages, updates are being translated and appropriate tools are available. Therefore this only remains as an organisational issue.

The second part presents a big challenge. There are (too) many language combinations and an all-to-all mapping of free-text may not be feasible.

Therefore some smaller steps were investigated in TEL-ME-MOR, particularly mapping of subject access tools (subject headings, classification). The final recommendations are incorporated into the EDLproject workplan:

- Recommendations for improving subject access interoperability
    - Test a selection of cross-language approaches to subject data, including MACS, MSAC and CrissCross
    - Investigate the feasibility of loading into the MACS system the Luxembourg subject headings (Laval English / French) and the Spanish subject headings (Spanish – English), and if feasible, load the data
    - Provide an updated overview of recent European projects and initiatives that may have relevance to cross-language access to various types of collections.

  - Test cross-language searching in the European Library using the 70’000 RAMEAU/LCSH links against data from British Library and Bibliothèque nationale de France
  - Test large-scale linking (Sports and Theater – 1000 terms)
    - Load MSAC data in to the European Library portal
    - Incorporate more languages into the EL cross-language interface
  - Use the MACS Link Management system (replicated as required) central source (clearinghouse) of mapping results between subject headings used in European national libraries (classifications and subject headings)
- Testing automated mapping
  - Investigate HILT, TermSciences and WebDewey as tools to speed up link creation
  - [In TELplus, plans to test Stitch]
  - Contact OCLC to investigate their project exploring new web-based services for dynamic mapping between subject authority lists http://www.oclc.org/research/projects/termservices/
  - Contact the MULIR project:
    - “Our current Multi-Lingual Information Retrieval (MULIR) entry vocabulary index maps from English Library of Congress Subject Headings (LCSH) to words and phrases in over 100 languages and vice versa. This prototype was created from over ten million records of the University of California MELVYL online library catalog.”
- Recommendations for improving interoperability across authorities
  - Extending the theoretical analysis of D3.5 of TEL-ME-MOR and building on projects such as LEAF, VIAF and ONESAC, design and test the feasibility of a name authority control tool, implemented by (automatically) clustering existing variant headings, thus enabling searching on any of possible names for a person, persona, corporate body or geographic name. Result: A prototype of common (name) authority tool.

In addition to tasks proposed as recommendations of TEL-ME-MOR, the plan in EDLproject is to investigate free-text searching. For that an overview of existing tools would be necessary: language recognition tools, stemmers, automated translation tools etc. At this point we are not aiming at a fully functional cross-language IR for all languages, but at least a prototype of a limited application. Therefore cooperation with any relevant research is needed.