Implementing MLIA in an existing DL system

Martin Braschler
Zurich University of Applied Sciences Winterthur
Switzerland
martin.braschler@zhwin.ch

Nicola Ferro
University of Padova
Italy
nicola.ferro@dei.unipd.it

Julie Verleyen
The European Library
The Netherlands
Julie.Verleyen@kb.nl

ABSTRACT

We discuss the initial results of a feasability study aimed at highlighting the problems involved in implementing *MultiLingual Information Access (MLIA)* functionalities in an existing federated digital library system. We suggest some possible solutions which we are currently investigating and which keep into account the constraints given by the architecture and functioning of the existing system and try to overcome them in order to offer as effective as possible MLIA features.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.7 [**Information Storage and Retrieval**]: Digital Libraries

General Terms

Design

Keywords

multilingual information access, digital libraries

1. INTRODUCTION

This paper reports on work in progress conducted in collaboration between DELOS[1], the European Network of Excellence on Digital Libraries funded by the EUs 6^{th} Framework Programme, and *The European Library (TEL)*[2], a service fully funded by the participant national libraries members of the *Conference of European National Librarians (CENL)*[3], which aims at providing a co-operative framework for integrated access to the major collections of the European national libraries. The paper therefore mainly describes the methods we have identified to this date as promising for our work, plus initial findings when applying them to the data. Final result presentation will only be possible upon completion of the study.

The aim of the collaboration is to conduct a feasability study which identifies the main issues involved in implementing full *MultiLingual Information Access (MLIA)* functionalities in TEL. By full MLIA we mean the possibility for users of TEL to access and search the federated libraries in their own (or preferred) language, retrieve documents in other languages, have the results presented in an interpretable fashion (e.g. possibly with a summary of the contents in their chosen language).

In particular, we discuss the initial results of a feasability study which highlights the problems we encounter when we try to apply methods and techniques developed in the MLIA research field to a running system, which was not designed from the outset to provide advanced MLIA functionalities. Then, we present the solutions we are examining in order to overcome the highlighted problems and we discuss the effectiveness of the proposed solutions.

The paper is structured as follows: Section 2 provides an overview of the TEL system, its architecture, and its functioning; Section 3 introduces the problems in adding MLIA functionalities to the TEL system; Section 4 presents the possible solutions and discusses their pros and cons; finally, Section 5 draws some conclusions.

2. TEL OVERVIEW

Figure 1 shows the architecture of the TEL system. As discussed in [4], the TEL project aims at providing a "low barrier of entry" in the TEL system to the national libraries which want to join it. This easiness of integration is achieved by extensively using the *Search/Retrieve via URL (SRU)*[4] protocol in order to search and retrieve documents from national libraries. In this way, the user client can be a simple browser, which exploits SRU as a means for uniformly accessing national libraries.

With this objective in mind, TEL is constituted by three components:

- a Web server: it provides users with the TEL portal;
- a central index: it harvests catalogue records from national libraries which support *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)* [2] and provides integrated access to them via SRU;

[1] `http://www.delos.info/`
[2] `http://www.theeuropeanlibrary.org/`
[3] `http://www.cenl.org/`
[4] `http://www.loc.gov/standards/sru/`

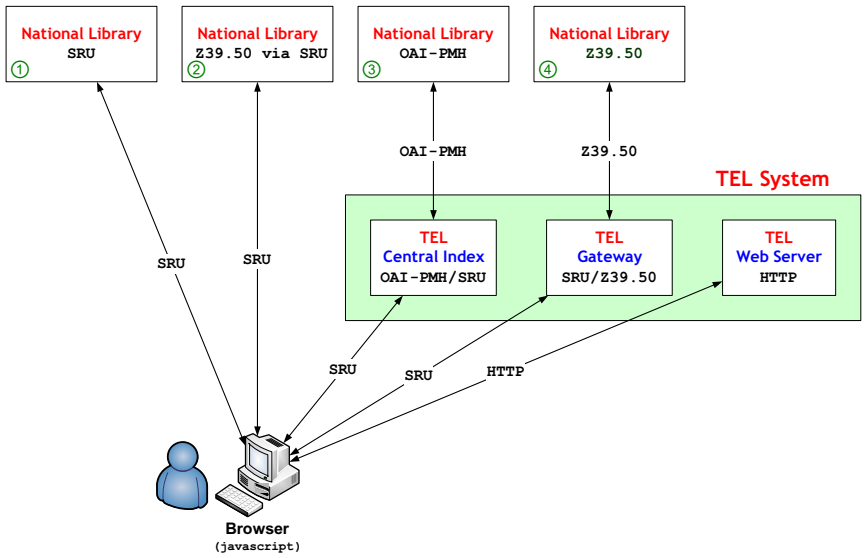

Figure 1: Architecture of the TEL system.

- a gateway between SRU and Z39.50: it allows national libraries which support only Z39.50[5] to be accessible via SRU.

This light architecture allows TEL to support and integrate the following cases:

1. a national library which natively uses SRU can be directly searched by the client;
2. a national library can have a local gateway between Z39.50 and SRU, so that the client can access it as if it was a native SRU library;
3. a national library able to share metadata records by using OAI-PMH can be searched via the TEL central index, which harvests those records and makes them accessible to the client via SRU;
4. a national library which supports only z39.50 can rely on the SRU/Z39.50 gateway offered by the TEL system in order to be searched by clients.

We have now to examine how this architecture is actually used and how the interaction between the client and the different involved systems happens, because all these factors influence how MLIA can be integrated into TEL.

Figure 2 illustrates an example of interaction with the TEL system by using the sequence diagram notation of *Unified Modeling Language (UML)* [3]. The example considers the case in which a user wants to query, at the same time, a national library which exported its records in the TEL central index, a Z39.50 national library, and a native SRU national library.

- the user asks to the browser the connect to the *Uniform Resource Locator (URL)* of the TEL portal;
- the browser connects to the TEL Web server, which downloads all the TEL portal on the client. From now on, there is no more interaction with the TEL Web server, but all the computation and interaction with the user is managed by the browser by using Javascript;
- suppose the user decides to send a query to the national libraries mentioned above;
- the browser, by using SRU, takes care of routing the user's query to, respectively: the TEL central index for the national library which exported its record via OAI-PMH, the TEL gateway for the Z39.50 national library, and directly the native SRU national library. Then, it waits for the results to come back;
- the TEL gateway is in charge of actually querying the Z39.50 national library and gather its results;
- as soon as the queried systems respond, the browser receives the query results back from each systems and shows them to the user.

3. PROBLEMS

The architecture and functioning of the TEL system pose some problems when planning to introduce MLIA.

The TEL system has no control on queries sent to the national libraries, since the client browser directly manages the interaction with national library systems via SRU. As a consequence, introducing MLIA functionalities into the TEL system would have no effect on the national library systems. Thus, in order to achieve full MLIA functionalities, not only the TEL system but also all the national library systems should be modified and this is an unviable option that would require a very big effort and disregards the "low barrier of entry" guideline adopted in designing the TEL system.

In order to avoid the problem discussed above, still offering some MLIA functionalities, we plan to introduce an *isolated query translation* step in the query processing, as discussed in Section 4.1.

On the other hand, the TEL central index harvests catalogue records from national libraries, which beside catalogue

[5] `http://www.loc.gov/z3950/agency/`

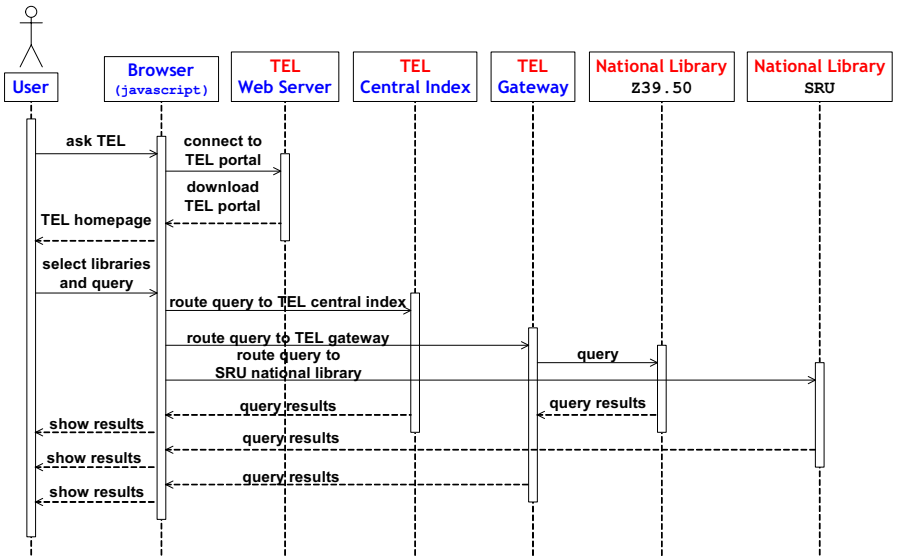

Figure 2: Sequence diagram of the functioning TEL system.

metadata may contain other information useful for applying MLIA techniques, such as an abstract. Since the central index is completely under the control of the TEL system, we plan to extend its functionalities by adding a component able translate the catalogue records in order to perform MLIA on them. We call this approach *pseudo-translation* and it is discussed in Section 4.2.

4. SOLUTIONS BEING EXAMINED

Figure 3 shows the architecture of the TEL system with two new components: the first one is in charge to perform the "isolated query translation", while the second one is responsible for the "pseudo-translation".

Note that the "isolated query translation" component can be directly accessed by the client browser by using the SRU protocol and thus the interaction with this new component is explicit. On the other hand, the "pseudo-translation" component is not directly accessed by the client browser but it represents an extension of the TEL central index, which would be enhanced with MLIA functionalities.

4.1 Isolated Query Translation

"Isolated query translation" can be considered as a sort of pre-processing step where the translation problem is treated as completely separate from the retrieval.

Before actually submitting the query, as it happens in figure 2, the user asks the browser to translate the query. Then, the browser sends via SRU the query to the "isolated query translation" component which takes care of translating the query and, if necessary, of applying query expansion techniques to reduce the problem of missing translations. At this point, the user can interactively select the translation which best matches his needs or can change some query term to refine the translation. In this latter case, the translation process may be iterated. Once the desired translation of the query is obtained, the retrieval process happens as in the case of figure 2, using both the translated query and the original one.

The main advantage of this solution is its easiness of implementation and its compliance with the "low barrier of entry" approach of TEL. Indeed, the national library systems do not require any modification and this new functionality can be transparently added to them, even if it is actually performed in the TEL system.

"Isolated query translation" requires some user interaction, because the user may need to choose among multiple translations of the same term in order to disambiguate them or may need to modify the original query if the translated query does not match his needs.

The main drawback of this approach is that, being the translation separated from the retrieval process, relevant documents may be missing in the result set and thus the performance may be low. Moreover, huge linguistic resources, such as dictionaries, are needed since the vocabulary used in queries is expected to be very large; this has to be repeated for each pair of source/destination language the system is going to support.

4.2 Pseudo-translation

With the idea of "pseudo-translation" we want to tackle two problems that we expect to arise if applying MLIA approaches that were developed for information retrieval on collections of lengthy full-text documents to library records instead.

Preliminary analysis of the records accessible in TEL that originate from the Bibilothèque Nationale de France and the British Library confirm that there is little full text that can be used for retrieval. We selected a sample of 10,000 random French and English records each to study their characteristics. While all records contain a (short, basically one-sentence) title, only approximately 13% of all records in our French data sample contained additional data suited for retrieval. In the English sample, approximately 88% of records contain subject keywords that may prove to be suitable for retrieval. Other fields interesting for retrieval are only contained in a small number of records (11% contain an alternative title, 6% a listing of the table of contents, and 1% an abstract).

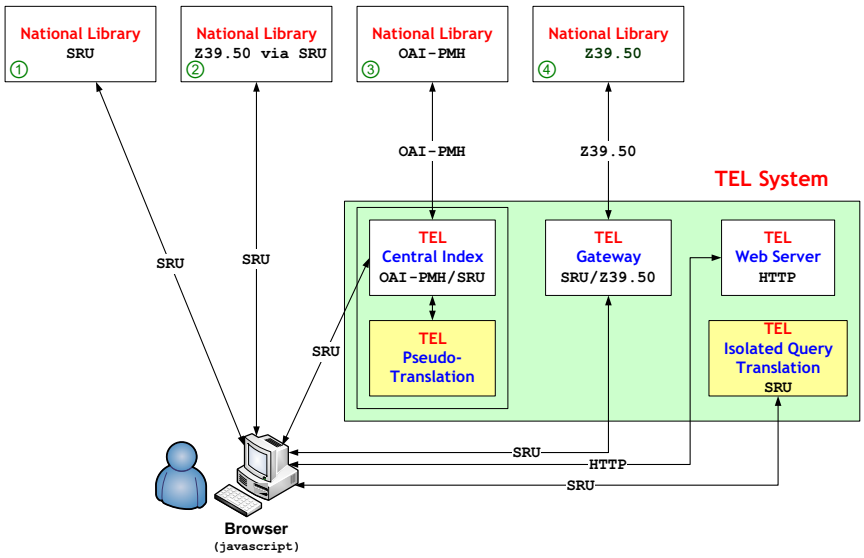

Figure 3: Architecture of the TEL system with new MLIA functionalities.

Having only a small number of content-bearing words to work with means that we expect translation failures (out-of-vocubulary words) to have serious consequences. If several key words go untranslated, a record can easily "disappear", i.e. it becomes impossible to retrieve the record.

We plan to counter this problem by applying methods originally developed for query expansion to the records, adding additional terms that may be used for subsequent retrieval. Using this strategy, we hope to strengthen the key concepts expressed in the limited text fields of the record, and therefore increase the probability that these concepts "survive" translation.

To this end, we try to simulate in our feasability study an environment in which as large a sample as possible of the record data is enriched by expansion terms. Each record is run as a query against all other records of the sample, selecting those terms from expansion with highest weight that do not originally appear in the record. To simulate this process for analysis, any retrieval system that allows query expansion can potentially be used.

The resulting additional terms form no sentences. This was deemed to be unproblematic for the following translation stage, as the nature of the existing text in the records also lends itself not specifically to machine translation (short, often ungrammatical text). We expect any translation resource that covers an extensive vocabulary to be suitable at this stage to get an impression of the workability of the overall method.

The same expansion idea can be applied to the query in an analogous way. In the course of the feasability study we aim to determine whether the expansion minimizes retrieval problems due to out-of-vocabulary query terms (see also [1]).

5. CONCLUSIONS

This paper reports the initial results of a feasibility study carried out to add MultiLingual Information Access functionalities to an existing federation digital libraries, *The European Library (TEL)* system.

We proposed two different approaches for introducing MLIA functionalities in the TEL system: the first one, called "isolated query translation", performs a pre-processing step to translate the query and then routes the translated query to the national library systems. The second one, called "pseudo-translation", involves only queries sent to the TEL central index but merges the translation process with the retrieval one in order to offer more effective MLIA functionalities.

Acknowledgments

We thank Thomas Arni for his work on the pseudo translation approach.

This work was partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

6. REFERENCES

[1] L. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In N. J. Belkin, A. D. Narasimhalu, P. Willett, W. Hersh, F. Can, and E. Voorhees, editors, *Proc. 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997)*, pages 84–91. ACM Press, New York, USA, 1997.

[2] OAI. The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0. `http://www.openarchives.org/OAI/openarchivesprotocol.html` [last visited 2006, February 28], October 2004.

[3] OMG. Unified Modeling Language (UML), Version 2.0, formal/05-07-04. `http://www.omg.org/technology/documents/formal/uml.htm` [last visited 2006, February 28], 2004.

[4] T. van Veen and B. Oldroyd. Search and Retrieval in The European Library. A New Approach. *D-Lib Magazine*, 10(2), February 2004.