Roadmap for MultiLingual Information Access in the European Library

Maristella Agosti[1], Martin Braschler[2], Nicola Ferro[1], Carol Peters[3], and Sjoerd Siebinga[4]

[1] University of Padua, Italy
{maristella.agosti, nicola.ferro}@unipd.it
[2] Zurich University of Applied Sciences Winterthur, Switzerland
martin.braschler@zhwin.ch
[3] ISTI-CNR, Area di Ricerca – 56124 Pisa, Italy
carol.peters@isti.cnr.it
[4] The European Library, The Netherlands
Sjoerd.Siebinga@KB.nl

Abstract. The paper studies the problem of implementing MultiLingual Information Access (MLIA) functionality in The European Library (TEL). The issues that must be considered are described in detail and the results of a preliminary feasibility study are presented. The paper concludes by discussing the difficulties inherent in attempting to provide a realistic full-scale MLIA solution and proposes a roadmap aimed at determining whether this is in fact possible.

1 Introduction

This paper reports on a collaboration [1,4,5] conducted between DELOS[1], the European Network of Excellence on Digital Libraries funded by the EU Sixth 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 ultimate goal of MultiLingual Information Access (MLIA) in TEL is to enable users of TEL to access and search the library in their own (or preferred) language, retrieve documents in other languages and have the results presented in an interpretable fashion (e.g. possibly with a summary of the contents in their chosen language). The problem is complex and many factors are involved. These include: the number of languages involved, the current heterogeneous setup of TEL, the lexical tools and resources needed.

[1] http://www.delos.info/
[2] http://www.theeuropeanlibrary.org/
[3] http://www.cenl.org/

L. Kovács, N. Fuhr, and C. Meghini (Eds.): ECDL 2007, LNCS 4675, pp. 136–147, 2007.

- *Number of Languages.* The number of different languages represented in TEL constitutes a major hurdle for MLIA, as ideally it should be possible to launch a query in any one of the national languages of the TEL collections and retrieve relevant material in any one of the collections. Possible approaches to the problem might be the use of multilingual ontologies, metadata and subject authority data, statistical translation resources, or some kind of interlingua.
- *Heterogeneous set-up.* A serious problem is represented by the heterogeneous set up of TEL, as there are severe limitations on how the existing infrastructure is able to process a cross-language query result.
- *Resources Needed.* Any cross-language strategy implies the acquisition and development of appropriate lexical tools and linguistic resources such as stemmers, morphologies, bilingual dictionaries, etc. As more languages are involved, not only does the number of resources increase, but the type of resources needed becomes more complex and more difficult to acquire.

The implementation of MLIA in TEL is thus an ambitious task and must be considered a medium/long-term goal, to be achieved through a series of intermediate steps. In this paper, we will try to determine the scope of implementation, attempt to identify the main obstacles, and devise a road-map which could help us to determine whether the full implementation of free-text MLIA in TEL is in fact practicable.

This document is organised as follows: section 2 describes the current TEL architecture; in section 3 we discuss the underlying motivations for our study and the main goals; Section 4 presents solutions studied so far; finally Section 5 proposes further experimentation and outlines a Roadmap for future investigations.

2 TEL Architecture Overview

Figure 1 shows the architecture of the TEL system. The TEL project aims at providing a "low barrier of entry" for the national libraries that should be able to join TEL with only minimal changes to their systems [7]. This ease 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: provides users with the TEL portal;
- a central index: harvests catalogue records from national libraries which support the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[5] and provides integrated access to them via SRU;

[4] `http://www.loc.gov/standards/sru/`
[5] `http://www.openarchives.org/OAI/openarchivesprotocol.html`

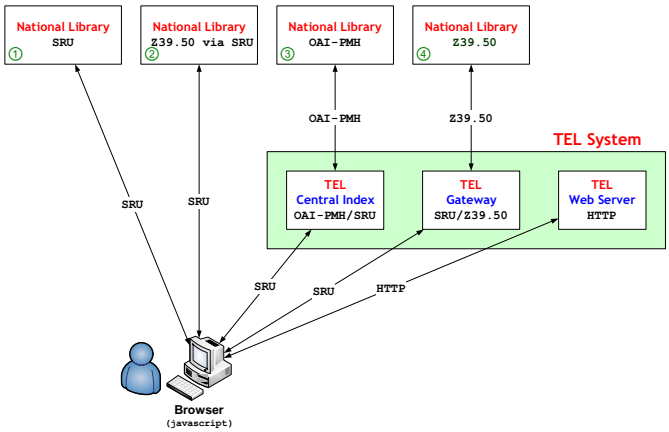

Fig. 1. Architecture of the TEL system

– a gateway between SRU and Z39.50: allows national libraries which support only Z39.50[6] to be accessible via SRU.

This light architecture allows TEL to support and integrate as follows:

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 were a native SRU library;
3. a national library which supports only Z39.50 can rely on the central SRU/Z39.50 gateway offered by the TEL system in order to be searched by clients;
4. 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.

Figure 2 illustrates an example of interaction with the TEL system using the sequence diagram notation of Unified Modeling Language (UML)[7]. The example considers the case in which a user wants to query, simultaneously, a national library which exported its records to the TEL central index, a Z39.50 national library, and a native SRU national library.

– the user asks the browser to 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 using Javascript.

[6] `http://www.loc.gov/z3950/agency/`
[7] `http://www.omg.org/technology/documents/formal/uml.htm`

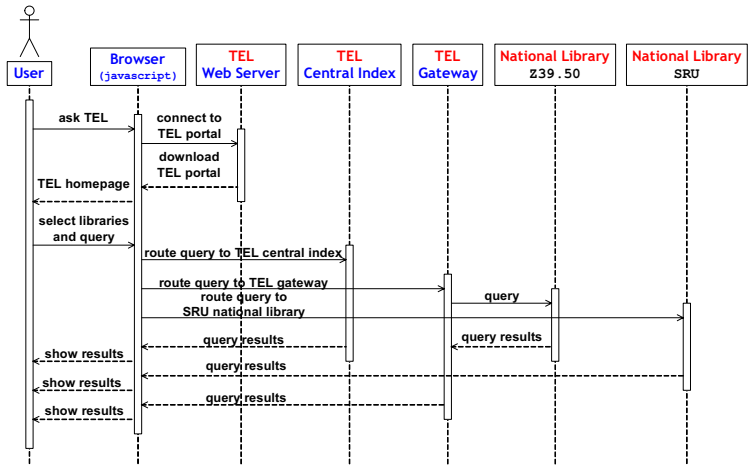

Fig. 2. Sequence diagram of the functioning TEL system

If the user decides to send a query to the national libraries mentioned above:

- the browser, using SRU, routes 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 to the native SRU national library, and waits for the results to come back;
- the browser receives the query results back from each system and displays them to the user.

3 Motivations and Goals

True multilingual access is more than just being able to search in more than one language. It means that the intended result is retrieved in each target collection regardless of language, character-encoding, metadata-schema, or normalisation rules. TEL is a heterogeneous federated search service for national libraries in Europe. This heterogeneous set-up poses a wide variety of problems for the implementation of MLIA functionality.

MLIA in TEL can be roughly divided into three areas: 1. Multilingual user interface; 2. Multilingual mapping/linking of controlled vocabulary; 3. Multilingual search on free-text.

1. The TEL Portal interface and help texts are currently available in the 20 languages of the full partners.[8] Localization is the responsibility of the individual libraries and translation files are updated with each new release.

[8] Languages of full-partners (20): Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Serbian, Slovakian, Slovenian. Languages to be added in the near future (8): Bulgarian, Icelandic, Irish, Norwegian, Romanian, Russian (?), Spanish, Swedish.

2. Under the EDLProject[9] a study is currently underway to determine how existing controlled vocabulary initiatives can be integrated into the TEL service. This study builds on previous work done under the TEL-ME-MOR project[10]. The initiatives being examined include:

- Subject Headings: Multilingual ACcess to Subjects (MACS), Multilingual Subject Access to Catalogues (MSAC), CRISSCROSS, etc.
- Authority Files: Linking and Exploring Authority Files (LEAF), Virtual International Authority File (VIAF).
- Classification Schemata: feasibility of linking various translations of classification schemata like Universal Decimal Classification (UDC), Dewey Decimal Classification (DDC), and so on.

3. Studies in the MLIA domain have mostly focused on the remaining area: multilingual search on free-text. The main reason is that it is very difficult to use controlled vocabularies for MLIA unless a multilingual version is available covering all the languages in the collection and all documents have been classified using this. We thus decided to make free-text MLIA the main focus of this paper.

3.1 Issues to Be Considered

The implementation of free-text MLIA in TEL is an ambitious aim and success is not guaranteed. Although great advances have been made in recent years in the field of multilingual information access[11], it remains difficult to implement this functionality outside a controlled setting with unified full-text data-resources. In TEL, MLIA must be applied to hybrid resources that contain only snippets of free-text (often only keywords or ungrammatical sentences) in rigidly structured bibliographic files.

We list here the main questions that must be asked when considering free-text MLIA-implementation in TEL.

1. How can we implement MLIA with the TEL "low barrier of entry" approach.
2. How can we implement a multilingual component for multiple federated targets that is scalable both with respect to content and languages?
3. Is it actually possible to have a one to many and many to one cross-language access to 20-31 languages and still get good results?
4. Are the necessary language processing tools and resources available for all target languages?
5. Is it possible to use pivot-languages to reduce the amount of linguistic resources needed?
6. How should we deal with languages that have small collections and a relatively small number of native speakers?

[9] `http://www.edlproject.eu`
[10] `http://www.telmemor.net/`
[11] See, for example, the results published by the Cross Language Evaluation Forum: `http://www.clef-campaign.org`

7. Which metadata fields are relevant for free-text multilingual search: 1) title; 2) description; 3) keywords; 4) type; 5) abstract?
8. How do we solve the problem of limited context? The content of these fields is generally very short and often ungrammatical.
9. Can response times be sufficiently fast in an operational web environment?

3.2 TEL User and System Requirements

User interaction and empowerment have always been key principles of The European Library. The same should apply with respect to MLIA. MLIA functionality must be integrated in the portal in a non-obtrusive and intuitive manner. From a system design perspective the following key requirements for MLIA can be identified: 1. similar functionality on local and federated targets; 2. reduction of query complexity; 3. scalability; 4. speed and reliability; 5. full Unicode compliance; and finally 6. focus on open-source software and linguistic resources. These requirements are examined in more detail in the rest of this section.

1. TEL aims to provide a unified integrated access to the resources of European National Libraries. It is therefore important that any MLIA solutions proposed provide the user the same functionality regardless of the type of targets being queried, i.e. the local TEL central index or federated targets. Another important constraint is that due to the TEL 'low barrier of entry' principle, the implementation of MLIA must be on the portal-side, i.e. no data manipulation on the partner side.

2. The question of how to query tens - sometimes hundreds - of collections/targets with a large number of translated query terms, has no easy answer. Bag-of-words or concatenated OR queries are often very complex and could lead to serious retrieval degradation. In order to reduce query complexity and improve response time, the option of creating target-specific queries could be explored.

3. Any MLIA solution should be flexible and scalable, in order to deal with the explosive growth of The European Library both in the number of targets and languages. Currently, TEL offers content from 23 National Libraries in 242 collections. In 2007, the numbers are expected to rise to 32 National Libraries and over 350 collections. Other relevant future expansions will be the addition of much more full-text in addition to the bibliographic records and integration with non-library Cultural Heritage institutions, like museums and archives.

4. An important aspect of scalability is reliable and consistent retrieval performance. During the prototyping phase, performance benchmarking should be done against very large amounts of data to ensure reliability. Because TEL is a fully operational service, response time is an important consideration. From an operational viewpoint, what would be an acceptable upper threshold for MLIA transactions in TEL (5-10 seconds)? This is not a easy question when search engines, like Google, give subsecond results. If multilingual retrieval is too slow, maybe returning a monolingual result first and presenting multilingual results later via pop-up or URL-link, would be a good intermediate solution.

5. All aspects of the MLIA implementation should be fully Unicode compliant, in order to properly handle the profusion of special diacritics and character encodings found in Europe[12]. Even though most targets support Unicode, the problem of composed vs. decomposed Unicode characters still gives incomplete results. For example, ¨a and `a"` are usually displayed in the same way to the user but, when processed, give back different results. Special attention should also be paid, to how target-side normalisation affects retrieval. For example, some of the database search interfaces of our targets replace `ž` with a `*` wildcard. This leads to large numbers of unwanted results.

6. It is necessary to focus on open-source software and linguistic resources to support the community and reduce cost of the working system. The European Library is funded by the national libraries themselves and therefore does not have the capital to buy expensive licensing for translation software and linguistic resources.

4 Implementing MLIA Functionality: A Feasibility Study

The architecture and functioning of the TEL system as described in the previous sections pose some problems when planning to introduce MLIA.

TEL 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 functionality into the TEL system would have no effect on the national library systems. Thus, in order to achieve full MLIA functionality, not only the TEL system but also all the national library systems would have to be modified. This is an unviable option as it would require a very big effort and disregards the "low barrier of entry" guideline adopted when designing the TEL system.

A two-step solution is suggested and two complementary approaches are proposed: *isolated query translation* and *pseudo-translation* [1,4,5]. The first provides a basic cross-language search functionality for the entire TEL system; the second operates on the TEL central index.

Figure 3 shows the architecture of the TEL system with the two new components: the first performs the "isolated query translation", while the second 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 represents an extension of the TEL central index, which would be enhanced with MLIA functionalities. These two approaches are outlined below. They have both been well tested: the former via a set of mock-up implementations; the latter via a comparative evaluation setup. A full description of these studies is given in the literature cited above.

[12] Under the TEL-ME-MOR project, an extensive survey was done on Unicode support and several recommendations were made to TEL partners.

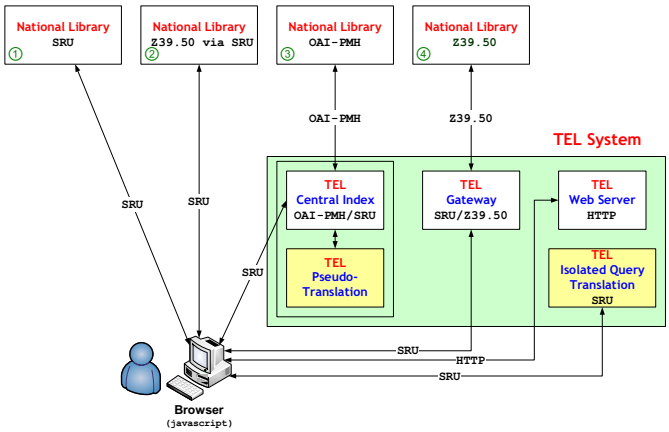

Fig. 3. Architecture of the TEL system with new MLIA functionalities

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, the user asks the browser to translate it. The browser sends the query via SRU to the "isolated query translation" component which translates it and can also apply query expansion techniques to reduce the problem of missing translations [3]. 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 is executed using both the translated query and the original one.

"Isolated query translation" requires some user interaction, because the users 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 their needs.

The main advantage of this solution is its ease of implementation and its compliance with the "low barrier of entry" approach of TEL. No changes to the national library systems are required and this new functionality can be transparently added to them, even if it is actually performed in the TEL system.

The main drawback is that, as the translation is 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; translation components are needed for each pair of source/destination language the system is going to support.

4.2 Pseudo-translation

The aim of the "pseudo-translation" approach was to tackle two problems that can arise when applying MLIA strategies developed for information retrieval on collections of lengthy full-text documents to library records.

In fact, a preliminary analysis of the records accessible in TEL originating from the Bibilothèque Nationale de France and the British Library confirmed that there is little full text in the TEL documents that can be used for retrieval. The characteristics of a sample of 10,000 random French and English records were studied. While all records contain a (short, basically one-sentence) title, only approximately 13% of all records in the 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).

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

This problem was addressed by applying methods originally developed for query expansion to the records, adding additional terms that may be used for subsequent retrieval. Using this strategy, the key concepts expressed in the limited text fields of the record were strengthened, and the probability that these concepts "survive" translation was increased.

The feasibility study simulated an environment in which as large a sample as possible of the bibliographic records - namely 151,700 - was enriched by expansion terms. Each record was run as a query against all other records of the sample, selecting those terms from expansion with highest weight that did 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 formed no sentences. This was deemed to be unproblematic for the following translation stage, as the nature of the existing text in the records does not also lend itself specifically to machine translation (short, often ungrammatical text). For this reason, any translation resource that covers an extensive vocabulary should be suitable . The same expansion idea can be applied to the query in an analogous way.

This second component of the feasibility study was evaluated carefully and the results were very encouraging. We tested the method by performing cross-language retrieval with German queries on the pseudo-translated English records.

When applying overlap analysis, 55% of queries analyzed showed evidence of good retrieval results, and 83% of queries showed evidence that they did not suffer significantly from the cross-language setup when compared to the monolingual baseline (note that for some of these queries there simply will be no relevant records in the collection!). The latter number is encouraging, being in line with what has been reported as state-of-the-art for Cross Language

Information Retrieval (CLIR) in the Cross-Language Evaluation Forum (CLEF) campaign on lengthy documents. A full description of the evaluation is given in [4]. Please note, however, that the numbers have to be treated with care, owing to the limitations described above. This approach should actually benefit in terms of effectiveness when scaling up to larger collections, which would occur when implementing the approach in the actual TEL system.

Combining Both Approaches. It is important to note that these two approaches can be implemented in conjunction in order to improve the MLIA functionality offered to TEL users. The implementation is facilitated as they share common components at the architectural level. For example, the translation engine or the translation resources, whether machine translation, machine readable dictionaries, or a combination of methods, can be shared by both approaches in order to reduce the development effort.

5 Towards Full MLIA

Although the results of the feasibility study were encouraging, they are a long way from solving the problem of implementing true MLIA functionality in TEL. Both solutions were proposed only in an language-to-language context (i.e. with queries in one language against target collections in a second language). This is very far from the one to many problem represented by TEL and mentioned in Section 3.

To a large extent, the isolated query translation approach may be the only feasible solution for cross-language querying on all the TEL collections with the existing TEL architecture, and it has the advantage that it offers the possibility for user interaction, giving the user the chance to check and modify the translation proposals offered. However, this approach also has significant limitations. In particular, it is impossible to perform additional query or preliminary results refinement on the basis of the contents of the target collections as these are held by national libraries and are not available for further processing in the TEL system. Furthermore, once we begin to talk about one-to-many querying with both collections and queries in more than 20 languages, problems clearly arise. The number of translation resources needed to cover all the possible language pairs would be enormous and, for many pairs of languages, probably non-existent. It seems clear that this solution is not viable to meet TEL's ambitious goal of enabling its users to search all target collections in their own language.

The pseudo-translation method appears to have more potential than the isolated query strategy but can only be applied to collections present in the TEL central index and lacks any kind of user interaction. In this method, the key concepts expressed in the limited text field of the bibliographic records are strengthened via an expansion process; the expanded record is then translated into the language of the query (German in the example cited) and monolingual retrieval is performed. These procedures (both document expansion and pseudo-translation) can be performed off-line with regular refreshings as the collections in the central index expand. However, again, once we begin to talk about queries and collections in a large number of languages, the problems are all too evident. The need

to pseudo-translate each collection of expanded records into more than 20 languages would mean that the TEL archives and indexes would become enormous and, as already stated, the number of translation resources needed would be very large.

In our opinion, the only possibility for true multilingual retrieval when we are faced with such a large number of languages is to use an interlingua or pivot language of some sort. The obvious candidates are English or French, as these are the languages for which bilingual dictionaries and machine translation resources are most easily available. Although the adoption of an interlingua involves multiple translation steps and thus considerably increases translation errors, it becomes a feasible option when faced with a potentially large number of query and target languages. A number of studies have attempted to evaluate the performance loss that can be expected with a pivot language and strategies to reduce this have been proposed [2,6]. Therefore, our proposal for a future feasibility study is to experiment again with both approaches in a truly multilingual context, introducing an interlingua and employing machine translation and/or bilingual dictionary sources which translate between English and the other languages involved in the experiments, ideally the same languages as those listed in point 5 of the roadmap below.

With the pseudo-translation approach on the central index, the idea would be to pseudo-translate a large set of the expanded documents from their source language (whatever it is) to English. A set of queries in a number of languages will then also be translated into English. The results will be evaluated and compared with the results that would have been obtained from a monolingual search.

In order to test the isolated translation approach using an interlingua, two alternatives can be explored. The first option would be to convince national libraries to also provide an English translation of their main metadata fields, e.g. title, keywords, and abstract. In this way, it would be possible to test this method, translating queries formulated in a number of languages into English and then sending them to the local collections selected. The second option, perhaps preferable as it does not require action from the national libraries, would be to perform multiple translations: instead of doing a language-to-language translation, we should perform a query language-to-interlingua and an interlingua-to-target language translation. Again, in both cases, evaluation would be done by comparing the results obtained against a monolingual search of the same collections.

Roadmap. Here below we propose a Roadmap. The main purpose of the Roadmap is to investigate whether full-scale MLIA in TEL is actually possible. In order to determine this, we propose the following steps:

1. Set up a survey to determine the availability of linguistic resources for the 20 languages of TEL-full-partners, paying special attention to languages with relatively small number of native speakers (e.g. Estonian and Slovenian).
2. Simultaneously, identify an exhaustive list of TEL user requirements, with sets of sample queries and possible use-cases. The sample queries should also contain queries which should give problems with partner-side normalisation and Unicode character-encoding;

3. Perform a feasibility study on how the inter-lingua approach can be used to meet the TEL user requirements.
4. TEL must design a component-based prototype which allows for easy scalability with new language components and integrates well in a federated architecture.
5. Test and benchmark the prototype's retrieval performance with a realistic number of representative languages from TEL-partners, e.g. Germanic (English, German), Romance (French, Portuguese), Slavic (Polish, Czech), Greek, Baltic (Latvian) and Finno-Ugric (Finnish).
6. Determine if retrieval performance (speed and accuracy) of the prototype is reliable and scalable enough to take into production.

Acknowledgements

The work reported in this paper has been 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).

References

1. Agosti, M., Braschler, M., Ferro, N.: A Study on how to Enhance TEL with Multilingual Information Access. DELOS Research Activities 2006. ISTI-CNR at Gruppo ALI, Pisa, Italy, pp. 115–116 (August 2006)
2. Ballestreros, L.A.: Cross-Language Retrieval via Transitive Translation. In: Advances in Information Retrieval: Recent Research from the CIIR, pp. 203–234. Kluwer Academic Publishers, Dordrecht (2000)
3. Ballesteros, L., Croft, W.B.: Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In: Proc. 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997), pp. 84–91. ACM Press, New York (1997)
4. Braschler, M., Ferro, N.: Adding MultiLingual Information Access to The European Library TEL. In: DELOS Conference 2007 Working Notes. ISTI-CNR, Gruppo ALI, Pisa, Italy, pp. 39–49 (February 2007)
5. Braschler, M., Ferro, N., Verleyen, J.: Implementing MLIA in an existing DL system. In: Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006) [last visited 2006, October 2], pp. 73–76 (2006), http://ucdata.berkeley.edu/sigir2006-mlia.htm
6. Lehtokangas, R., Airio, E.: Translation via a Pivot Language Challenges Direct translation in CLIR. In: Proc. SIGIR 2002, pp. 73–76. ACM Press, New York (2002)
7. van Veen, T., Oldroyd, B.: Search and Retrieval in The European Library. A New Approach. D-Lib Magazine 10(2) (February 2004)