Project no. 507618
DELOS
A Network of Excellence on Digital Libraries

Instrument: Network of Excellence

Thematic Priority: IST-2002-2.3.1.12
Technology-enhanced Learning and Access to Cultural Heritage

Deliverable D8.3.1 – Feasibility study on Multilinguality

Due Date of Deliverable: 31st December 2006
Actual Submission Date: xx XXXX 2006

Start Date of Project: 01 January 2004 **Duration: 48 Months**

Organisation Name of Lead Contractor for this Deliverable
CNR-ISTI (2)

Version: final

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)

Dissemination Level: PU (Public)

MultiLingual Information Access in TEL

Nicola Ferro[1], Martin Braschler[2], Thomas Arni[2], Carol Peters[3]

[1] Department of Information Engineering (DEI)
University of Padova
Padova, Italy

[2] Institute of Applied Information Technology
Zurich University of Applied Sciences
Winterthur, Switzerland

[3] Institute of Information Science and Technologies (ISTI)
Italian National Research Council (CNR)
Pisa, Italy

Table of Contents

Executive Summary

The present report describes a feasibility study on how to implement multilingual information access functionality in *The European Library* (TEL)[1] system. This work has been conducted as part of a collaboration between DELOS, the European Network of Excellence on Digital Libraries, funded by the EC Sixth Framework Programme, and the TEL service, fully funded by the participant national libraries, members of the *Conference of European National Librarians* (CENL), which aims at providing a co-operative framework for integrated access to the major collections of the European national libraries.

This report analyzes the viability of two approaches, potentially to be used in conjunction:

1. using an "isolated query translation feature, which, as far as possible, isolates the problem of multilingual searching from the remainder of the system, and

2. using expansion techniques and pseudo translation on bibliographical records, to alleviate problems of translation errors and vocabulary coverage often encountered in systems for multilingual searching

After a thorough analysis of the current architecture of the TEL system and of its functioning, the following results have been achieved:

- Investigation of the "isolated query translation" approach for the TEL system;

- Study of how the "isolated query translation" approach can be integrated into the user interface of the TEL system;

- Description of an approach for multilingual access to bibliographical records using pseudo translation and expansion techniques;

- Demonstration of the pseudo-translation approach on a test collection of over 150,000 records, including expansion;

- Careful analysis of retrieval performance on the test collection and indication of the feasibility of scaling the chosen approach to a functioning system;

A detailed discussion on how the "isolated query translation" approach can be integrated into the current architecture of TEL system and on how this would impact the actual functioning of TEL system is given and the viable alternatives for effectively implementing it are proposed. In addition, different solutions for extending the present user interface of the TEL system in order to offer both simple and advanced MLIA functionalities are illustrated.

[1] http://www.theeuropeanlibrary.org/

The results of the “pseudo-translation of expanded records” approach are given for an in-depth analysis on a sample of 151,700 records using 100 statements of information needs (queries). We present both overall results as well as some query-by-query and single record analysis.

This extended analysis indicates promising cross-language retrieval performance, although the expansion has not had the impact that was desired. It is expected that this could be remedied through using a so-called pilot collection for expansion.

There is evidence that the effectiveness of cross-language retrieval on the sample records used is at around 80% of a monolingual baseline. For a real-world application, this compares very well with the results obtained in CLEF: the Cross-Language Evaluation Forum. State-of-the-art for the best cross-language systems and for languages on which much research has been done is approximately 85% of monolingual retrieval in a more favourable laboratory context [13].

1 Introduction

The growing interest for *Multilingual Information Access* (MLIA) is witnessed by international activities that promote the access, use, and search of digital content available in multiple languages and in a distributed setting, i.e. digital content held in different places by different organisations. In particular, the i2010 Digital Library Initiative[3], a flagship project of the European Commission, clearly states that the improvement of multilingual and multicultural information access and search is one of the key objectives in the drive to provide access to quality digital content for all [15]. Similarly, one of the priorities stated in the programme for the Seventh Framework Programme (FP7) is the development of : "ICT-based systems to support accessibility and use over time of digital *cultural* resources and assets, in a multilingual environment".

In this context, *The European Library* (TEL) has been identified as an existing project which can act as an embryo for a European Digital Library, under the i2010 programme and within FP7, by building upon the TEL-infrastructure to provide a highly visible, multilingual access point to the digital resources of Europe's cultural institutions. A collaboration between the DELOS Network of Excellence on Digital Libraries and TEL was thus initiated in order to investigate how to integrate MLIA functionalities into TEL and to provide guidelines and support for the implementation of MLIA in TEL [16].

This present feasibility study has thus been conducted with the aim of providing a solid basis for the development of MLIA into TEL and of producing:

- Guidelines as to how the TEL infrastructure should be adapted to be ready for the requirements of multilingual information access;
- Guidelines for the adoption of multilingual resources;
- Definition of strategies that should be adopted by TEL in order to enable TEL users to search in their preferred language and retrieve documents in other languages.

These initial goals have been achieved (and in fact expanded on) via the following results:

- Investigation of the "isolated query translation" approach for the TEL system;
- Study of how the "isolated query translation" approach can be integrated into the user interface of the TEL system;
- Description of an approach for multilingual access to bibliographical records using pseudo translation and expansion techniques;
- Creation of a test collection of more than 150,000 pseudo translated records, including expansion;

[3] http://europa.eu.int/information_society/activities/digital_libraries/index_en.htm

- Retrieval experiments on the test collection using 100 statements of information needs (queries) collected from logfiles provided by TEL;

- Indication of the feasibility of scaling the chosen approach to an operational system;

- Directions to support the choice of multilingual and translation resources for both the proposed approaches.

Part of the work described in this document has been also reported in [1], [9], and [25].

The document is organized as follows: Section 2 provides background information about multilingual information access and cross-language information retrieval; Section 3 discusses the present architecture and functioning of the TEL system and the problems it raises with respect to MLIA, and proposes two approaches to address these problems; Section 4 provides an in-depth discussion of the first approach, known as "isolated query translation"; Section 5 explains in detail the second approach, called "pseudo-translation of expanded records"; finally, Section 6 draws some conclusions and makes suggestions for future work.

2 Multilingual Information Access and Cross-Language Information Retrieval

By multilingual information access we usually denote procedures for search on collections of information items (in the context of this report, bibliographic records) that are potentially stored in multiple languages. Usually the term is used for situations in which the user is allowed to query the collection across languages, i.e. retrieving information items in a language that is different from the language used by the user to formulate his/her information need. In this narrower sense, the term "cross-language information retrieval" is often used. The report concentrates on this scenario, as multilingual information access in the wider sense, access to information in multiple languages, is already implemented in TEL. The report uses the terms *Multilingual Information Access* (MLIA) and *Cross-Language Information Retrieval* (CLIR) interchangeably in the following; indicating in both cases the narrower definition of querying across languages.

Approaches for CLIR can be classified according to different schemes. Oard [22] proposes a taxonomy for CLIR approaches in terms of what he calls types (free-text vs. controlled vocabulary) and aspects (knowledge-based vs. content-based). We follow the definition in Braschler et al. [8] which uses a first-level classification according to how the query and information items (documents in the cited paper) are matched across languages – be it by translating the query, the information item, or both. These three basic options can in some situations be extended to include a fourth, which does not use translation on either query or information items, but instead uses matching at sub-word level (see e.g. [20]). Since this option typically relies on lengthy textual representations of queries and information items, it seems to be less suitable for the present problem of matching short bibliographical records, and is not pursued further in the following.

Today, the mainstream research on cross-language information retrieval in Europe is carried out in the confines of the *Cross-Language Evaluation Forum* (CLEF) campaign in Europe [26]. The campaign gives researchers the possibility to compare different approaches to CLIR in a common setting and provides tools for both in-depth analysis and curation of the

experimental result [2], [3]. Most of the experiments in CLEF concentrate on retrieval on lengthy, unstructured full-text documents using a general vocabulary. In such a setting, evaluations have shown that query translation is a good compromise between effectiveness in terms of retrieval quality and efficiency, and query translation is therefore the prevailing method used by participants in the CLEF campaign. An overview of the recent achievements in CLIR can be found in [7], [13], and [14]. Generally, there is a growing sense among the academic community that the CLIR problem as outlined above (lengthy, unstructured full-text documents from a general domain) is fairly well understood from an academic standpoint [10], [6].

3 TEL Architecture and Functioning

Figure 1 shows the architecture of the TEL system. As discussed in [27], 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: which provides users with the TEL portal;
- a central index: which harvests catalogue records from national libraries which support the *Open Archives Initiative Protocol for Metadata Harvesting* (OAI-PMH) [21] and provides integrated access to them via SRU;
- a gateway between SRU and Z39.50[5]: which allows national libraries that support only Z39.50 to be accessible via SRU.

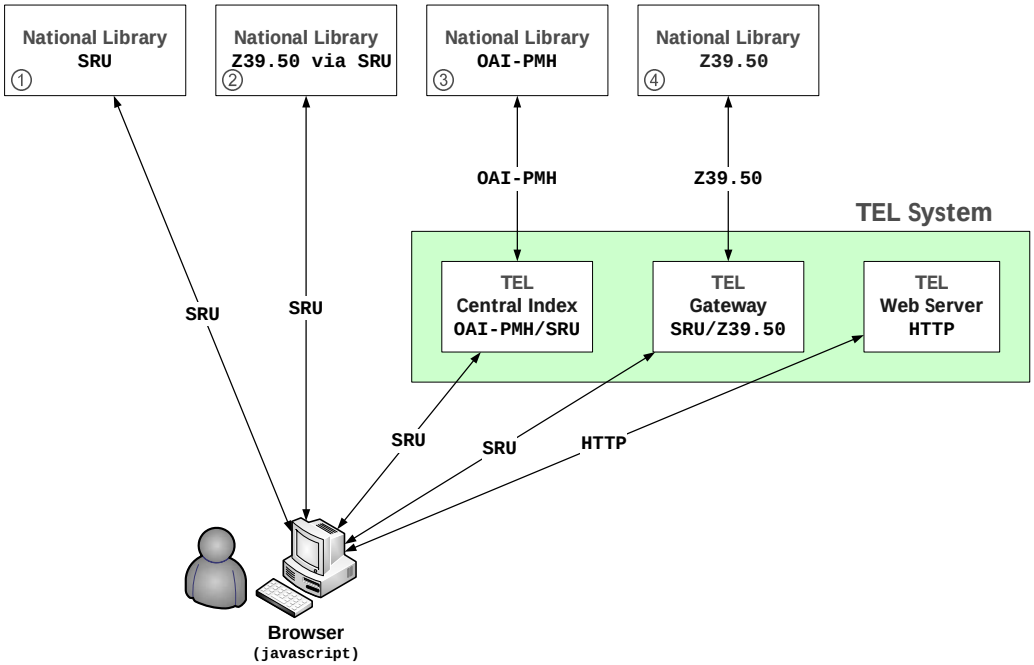

Figure 1: Present architecture of the TEL system.

[4] http://www.loc.gov/standards/sru/
[5] http://www.loc.gov/z3950/agency/

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

- a national library which natively uses SRU can be directly searched by the client;
- 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;
- 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;
- 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 client and the different systems involved interact, because all these factors influence how MLIA/CLIR 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) [24]. 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.

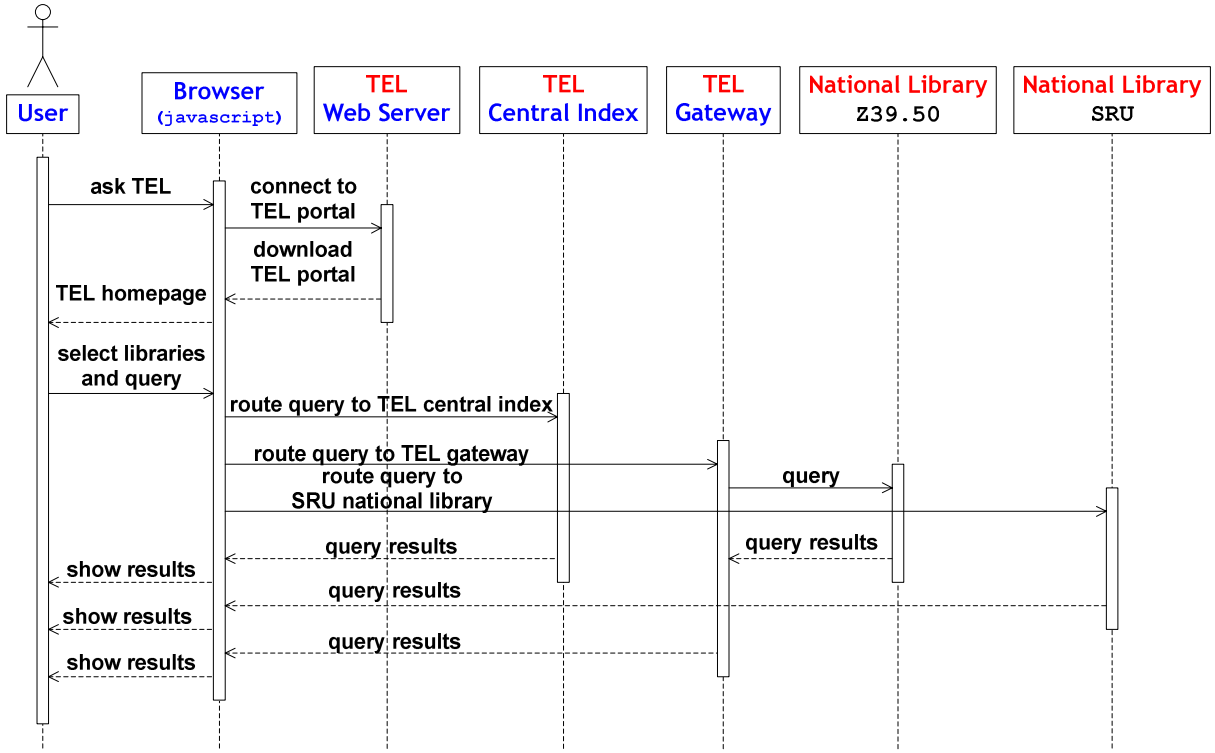

Figure 2: Sequence diagram of the functioning TEL system.

According to Figure 2:

- 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 by using Javascript;

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 the native SRU national library. And waits for the results to be returned;
- the TEL gateway is in charge of actually querying the Z39.50 national library and gathering its results;
- as soon as the queried systems respond, the browser receives the query results from each system and displays them to the user.

3.1 Problems and Approaches

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 would have to be modified. However, this is an unviable option as it would require considerable effort and disregards the “low barrier of entry” criteria adopted when designing the TEL system.
In order to avoid the problem discussed above, while still offering some MLIA functionalities, we plan to introduce an “isolated query translation” step in the query processing, as discussed in Section 4.
On the other hand, the TEL central index harvests catalogue records from national libraries, which in addition to catalogue 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. Unfortunately, the situation in the TEL system is substantially different from the ideal “mainstream setting” for CLIR. The large majority of information items are very short. Similarly, the expressions of information needs by the users, i.e. the queries, tend to be very short as well (see below for numbers). These considerations are addressed by the “pseudo-translation on expanded records” approach, described in Section 5.

4 Isolated Query Translation

In order to avoid the problems discussed above, we suggest the introduction of an isolated query translation step in the query processing, in order to translate a user query from a source language to a target language. Figure 3 shows the modifications necessary to the current TEL architecture to support this feature.

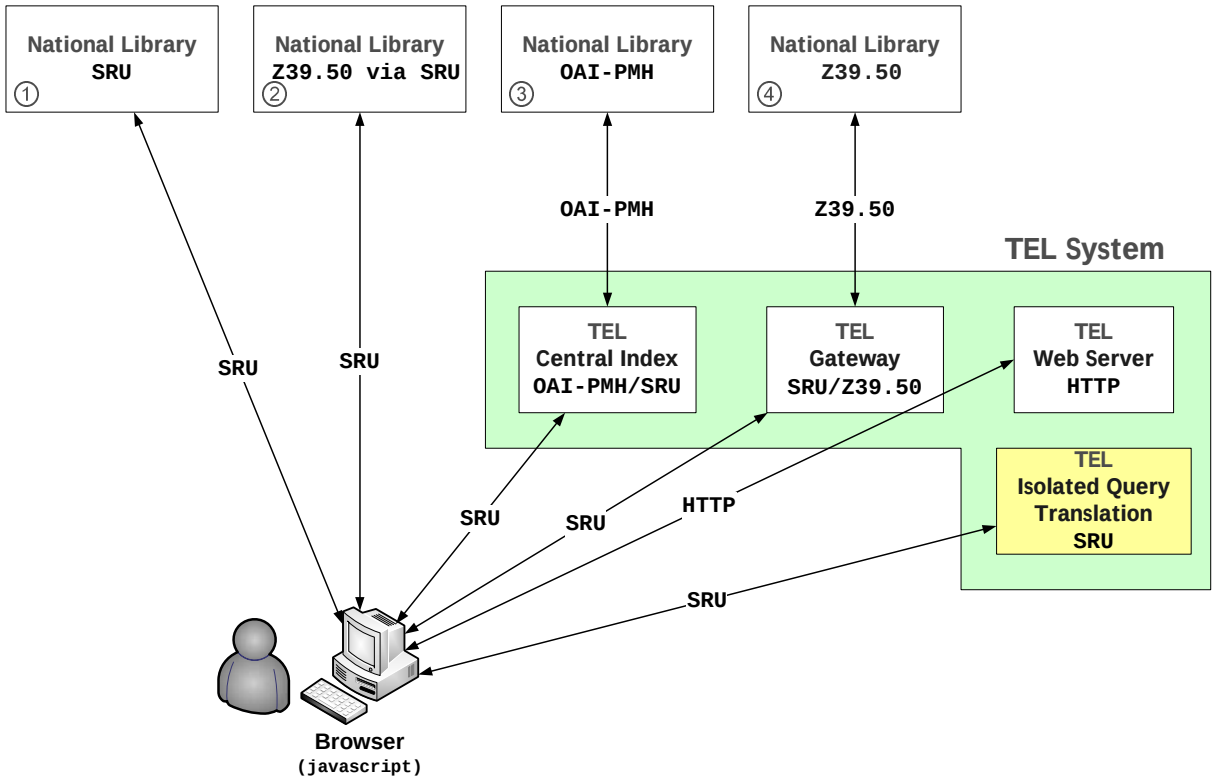

Figure 3: Architecture of the TEL system with the new "Isolated Query Translation" functionality.

The new "isolated query translation" component can be directly accessed by the client browser using the SRU protocol. "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 one.

Figure 4 presents a modified version of the sequence diagram of Figure 3 and shows how the functioning of the TEL system is affected by the "isolated query translation" feature:

- before actually submitting the query, as shown in Figure 2, the user asks the browser to translate it;
- the browser sends the query via SRU to the "isolated query translation" component, which takes care of translating it and, if necessary, applies 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 is initiated 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. The national library systems do not require any modification and this new functionality can be transparently applied when querying them, even though it is actually performed within the TEL central system.

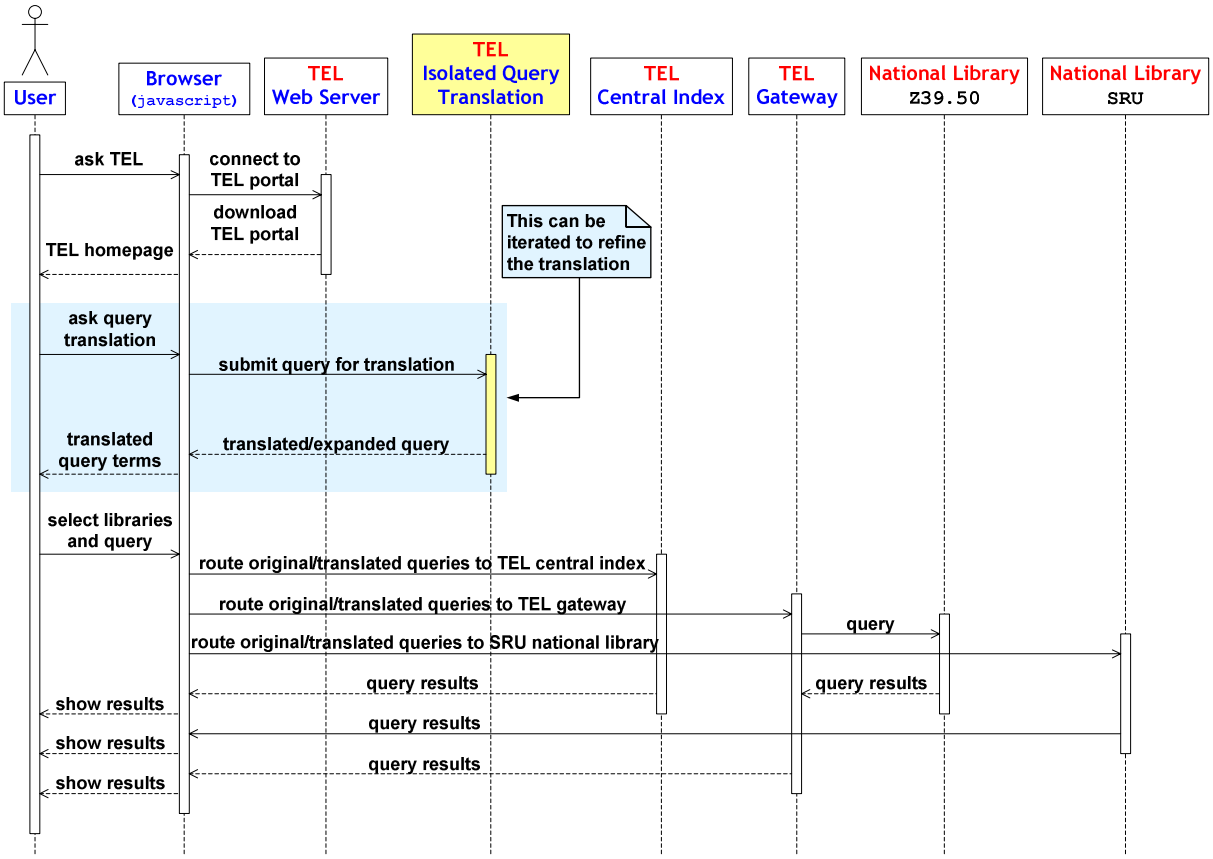

Figure 4: Sequence diagram of the functioning of the TEL system when using the "Isolated Query Translation" feature.

“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, 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; this has to be repeated for each pair of source/target language the system is going to support. Finally, the query expansion mechanism has to be generic and cannot be tailored on the collections queried, since the “isolated query translation” component does not interact with the national library systems.

4.1 Modifications to the TEL System User Interface for the “Isolated Query Translation” Feature

In this section, we discuss how the current user interface of the TEL system could be modified to introduce the “isolated query translation” feature.
We distinguish between two cases: the first concerns simple search functionality, while the second concerns advanced search functionality.

4.1.1 Simple Search

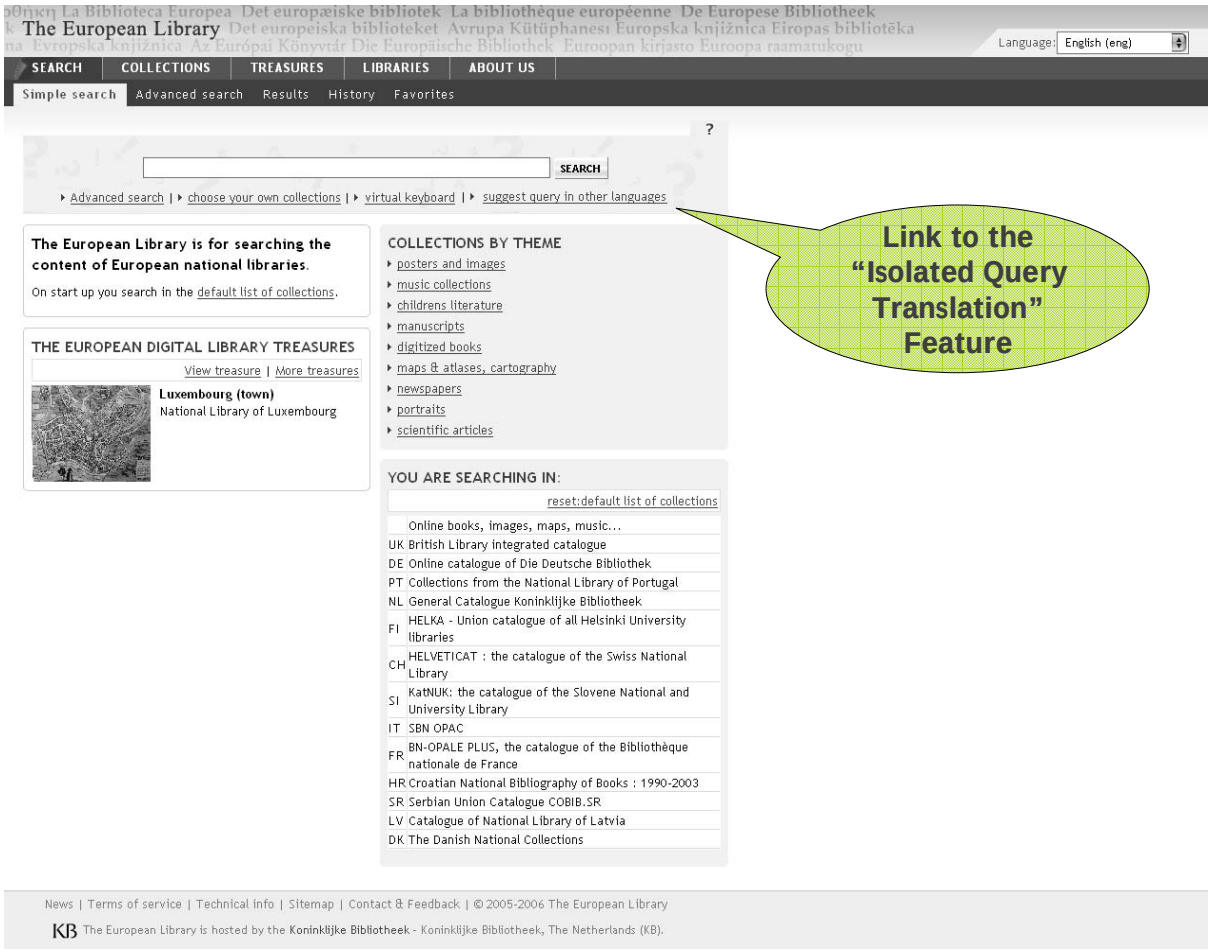

Figure 5: Link to the "Isolated Query Translation" feature in the simple search.

Figure 5 shows the TEL user interface for the simple search with an additional link to the "Isolated Query Translation" feature.

When the user clicks on the "suggest query in other languages" link, as shown in Figure 6, a box with the supported target languages for the translation appears below the search input box. The user can now check the languages for which he wants a translation of the query. Note that the behaviour of this box is similar to that of the "virtual keyboard" link, already implemented in the TEL system.

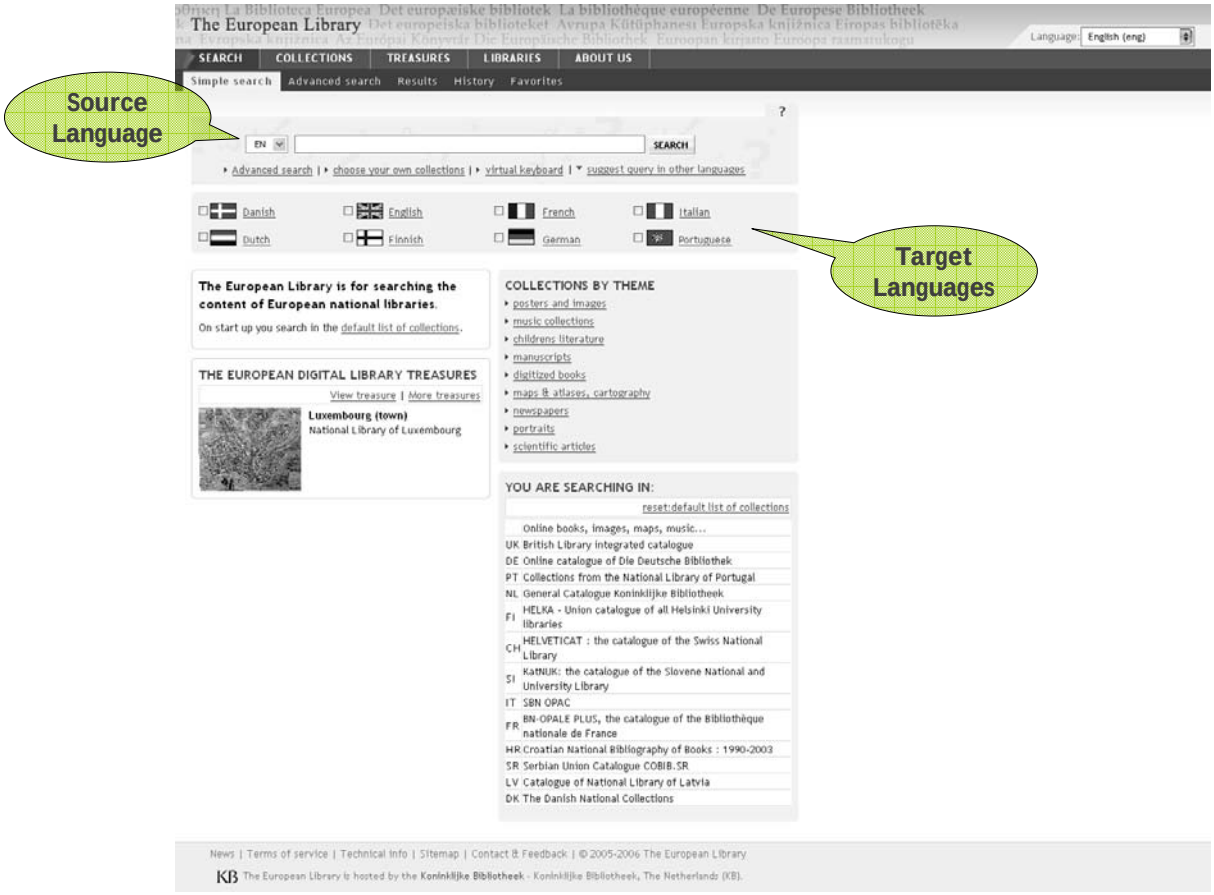

Figure 6: Selection of the source and target languages in the simple search.

Moreover, on the left of the search input box, a list with the possible source languages of the query is now shown, so that the user can specify the language of his original query. Note that this list differs from the list of the languages for the user interface in the upper right corner of the user interface mainly for two reasons. Firstly, users could not select their own language because, for example, they are able to understand English and they want to avoid the "extra-clicks" needed to change the language of the user interface. As a consequence, the actual selection in the list of languages for the user interface could be an erroneous indicator of the source language of the query. Secondly, the "isolated query translation" component could support the translation functionalities for a partially different set or a subset of languages with respect to the list of the languages available for the user interface. Thus, for both these reasons we need two separate language lists: one for the user interface and the other for the source language of a query.

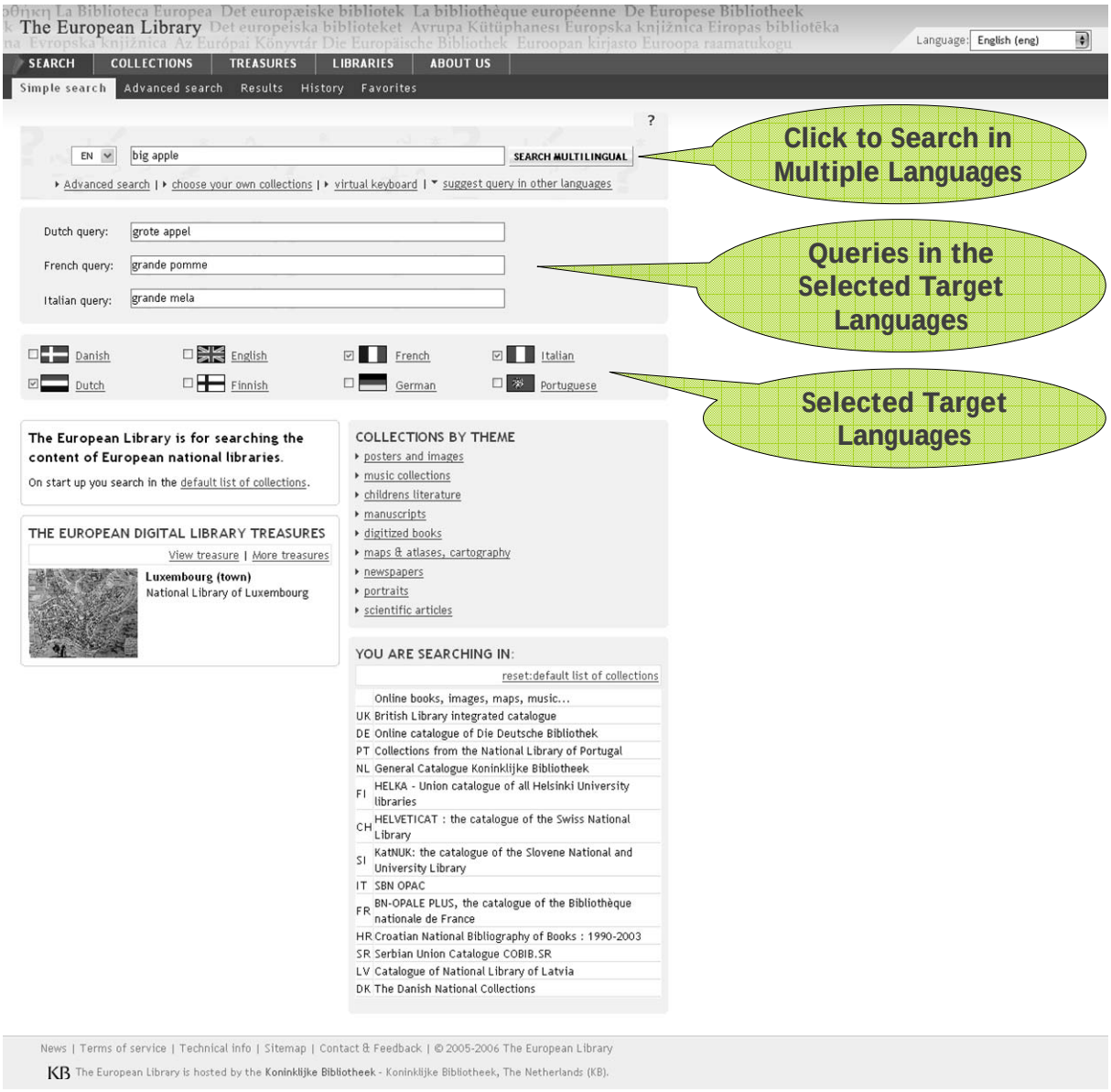

Figure 7: Query suggestions in other languages in the simple search.

As shown in Figure 7, for each target language selected by the user, a new text input box appears below the search input box containing the translation of the query in that language.
There are different possibilities for managing the user interaction when the translation of the query is shown. A first possibility would be to add a button “Suggest” so that the user presses it and the input boxes with the translation of the query appear. In this way, the user needs to explicitly request the translation. Another possibility would be a more *Asynchronous JavaScript Technology and XML* (AJAX) like style of interaction where the input box with the translation appears as soon as the user selects the language in the list of the target languages. In this way, the user does not need to explicitly request the translation and the system would be more proactive. In any case, both ways of interaction comply with the current approach of the TEL system in developing the user interface, which already exploits AJAX.
Once the translations have been obtained, there is the problem of managing the user interaction if the user needs to modify or refine the translations. Since the users of the simple search probably prefer an easy and intuitive way of interacting with the system, the translation refinement step should also be as simple as possible, even if some precision or some expressive power is lost. For this purpose, two alternatives can be offered to the user:

1. the user could directly edit each text input box in order to modify the translated query until it meets his needs. This means that the user can delete or add words to the translation.
   Note that if the user has no knowledge of a target language, he will not be able to refine the translated query and he will have to use the query suggested by the system without modifications.
   In addition, some attention has to be paid when multiple translations are possible because they have all to be listed in the input box and thus some visual clue should be provided to help the user in distinguishing between multiple alternatives.
2. if the translation greatly differs from user's expectations, there is the possibility of modifying the source query by adding or deleting words to it thus obtaining a new translation in the target languages, which can be modified as described above.

Once the various translations of the query are correct, the user can click the "Search Multilingual" button to perform a search in both the original and the selected target languages.

4.1.2 Advanced Search

The advanced search acts in a similar way to the simple search with respect to the "isolated query translation" feature, except that now there is the possibility of asking a suggestion for each field of the advanced search. Thus, the considerations made in the previous section hold also in this case, even if they may need to be adapted a little bit to the particular characteristics of the advanced search functionality.
Figure 8 shows the links to the "isolated query translation" feature for each field of the advanced search. Note that translating the query does not make sense for all the possible fields, neither is it needed. For example, for fields whose values are taken from a controlled vocabulary, there is no need to perform an on-the-fly translation; translations can be prepared in advance and simply shown in a list.
Figure 9 shows the selection of the source and target languages in the case that the "suggest query in other languages" link of the "Title" field is clicked.
Finally, Figure 10 shows the translated query for the "Title" field; if the user is satisfied that the translation is correct, they can perform a multilingual search by clicking on the "Search" button, otherwise they can refine the query directly in the text input boxes.
Since the advanced search offers search capabilities across multiple fields, suggestions for translations of the query in other languages can be performed for the different fields of the advanced search. For example, Figure 11 shows the case of two different suggestions for the "Title" and "Subject" fields, respectively: the query for the "Title" field is translated into English, while the query for the "Subject" field is translated into Portuguese.

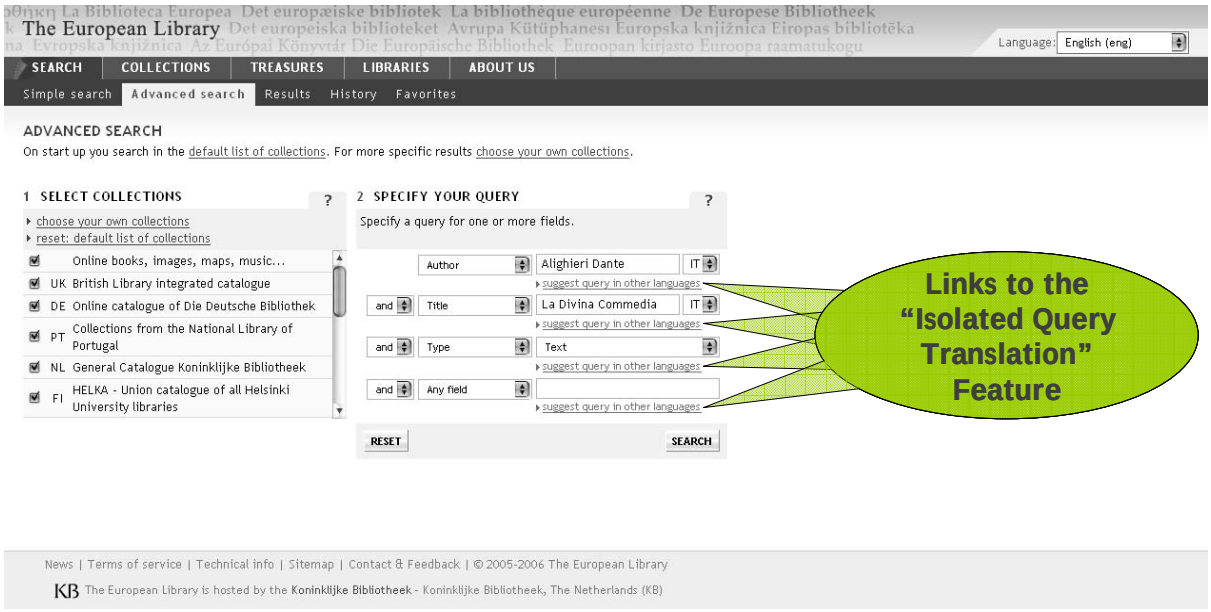

Figure 8: Links to the "Isolated Query Translation" feature in the advanced search.

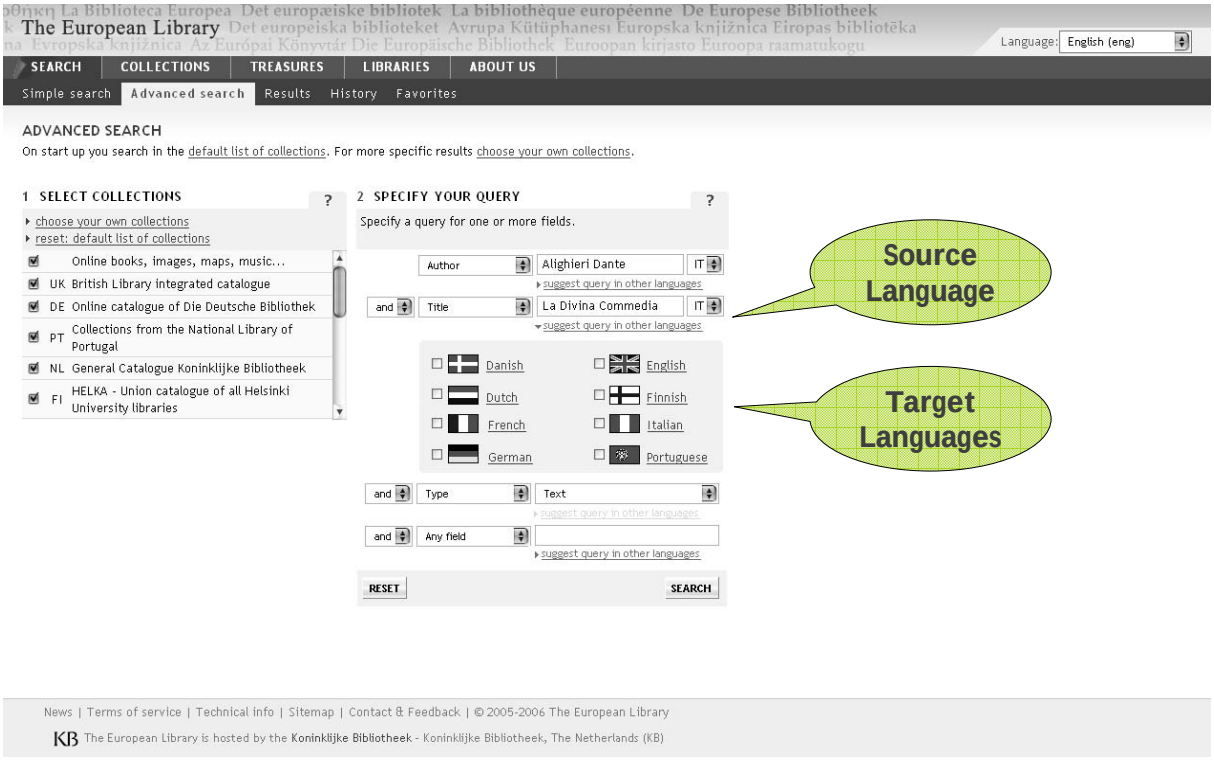

Figure 9: Selection of the source and target languages for the "Title" field in the advanced search.

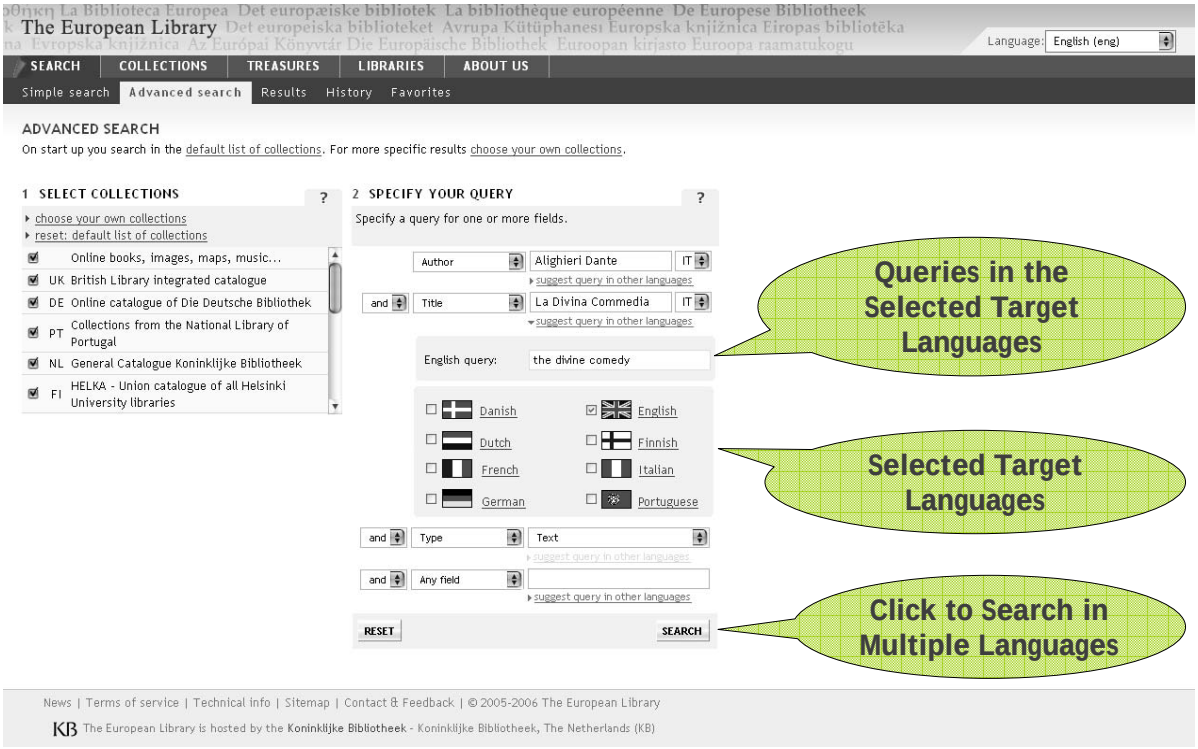

Figure 10: Query suggestions in other languages for the "Title" field in the advanced search.

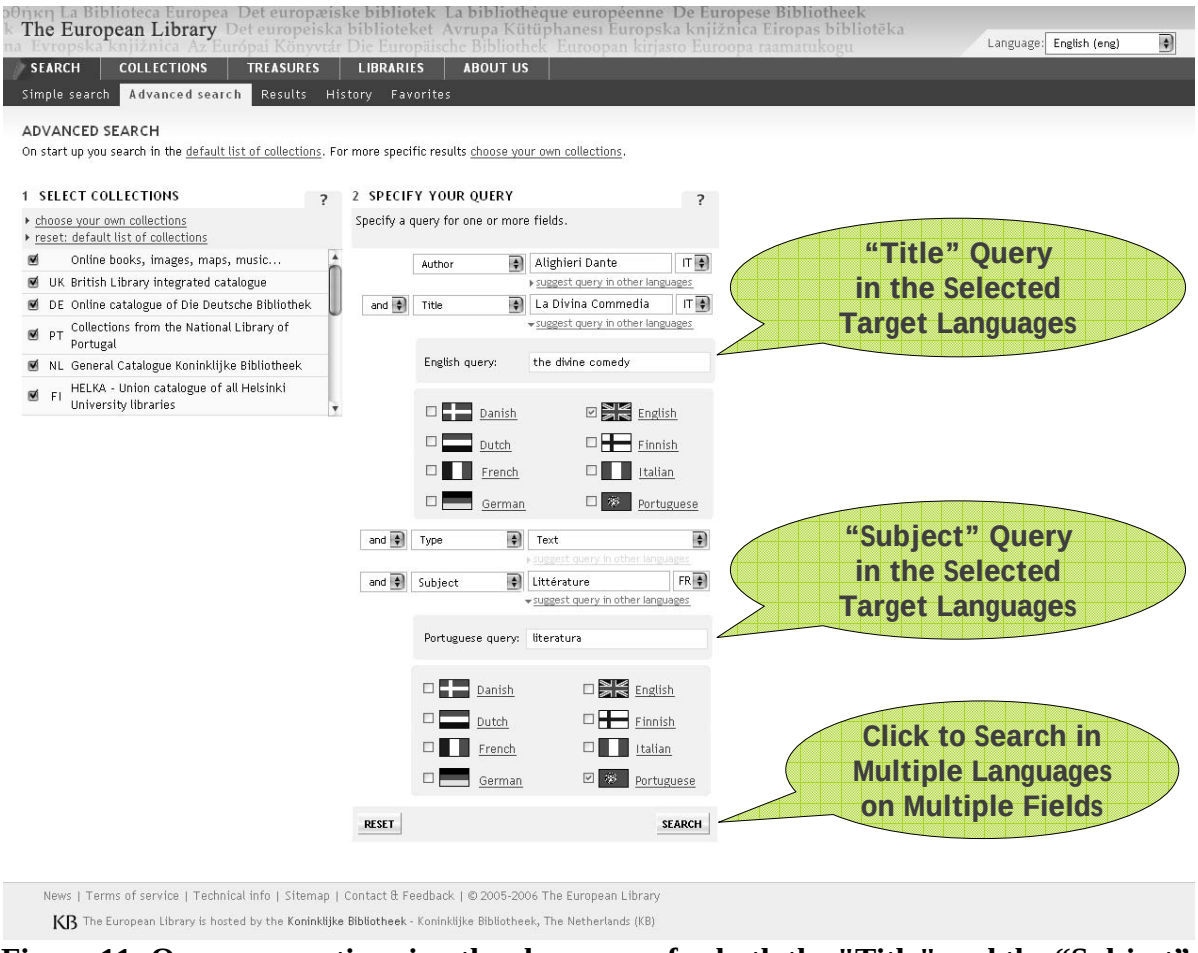

Figure 11: Query suggestions in other languages for both the "Title" and the “Subject” fields in the advanced search.

4.2 Alternatives for Implementing the Isolated Query Translation Component

With respect to the implementation of the "isolated query translation" component, shown in the architecture of Figure 3, some possible alternatives can be envisioned. The following sections describe these alternatives in more detail together with their pros and cons. Note that these sections do not aim at being a fully exhaustive survey of all the possible products available on the market but, on the contrary, they only aim at highlighting some possible solutions. Each solution, if chosen, would then require a follow-up deep investigation.

4.2.1 Commercial Software

A commercial software package like Systran[6] can be purchased and used as a basis for the "isolated query translation" component.
This solution would require almost no competence in MLIA and a limited effort for wrapping the purchased software in order to make it accessible through the SRU protocol. Moreover, the "isolated query translation" component can be kept updated by buying subsequent releases of the software. Finally, this solution would be similar to the solution adopted for the TEL central index, which is implemented using the Verity[7] software.
The main drawback is that commercial software usually acts as a "black box", which offer little or no control on the MLIA process. Moreover, data structures and information resources are usually in a proprietary format and not easily accessible. As a consequence, there possibilities of modifying or interacting with the software are limited and it could be difficult to re-use the data structures and information resources for other purposes in the future. Finally, the costs for licenses have to be taken into account.

4.2.2 Commercial Online Services

There are commercial online services that offer translation functionalities, such as Wordlingo ServiceAPI Development Tool[8]. In this case, a proprietary Web Application Program Interface (API) can be used to access the translation functionalities.
Considerations similar to the case of the commercial software also hold in this case: almost no MLIA competence is needed but there is no control on the MLIA process nor possibility of interacting with the data structures and information resources. Again the licensing costs have to be considered.
With respect to the actual implementation there are two possibilities, discussed in the following.
The first possibility is to implement a gateway, which receives the translation request from the client via SRU and properly forwards it to the translation service using the Web API. This solution would make the "isolated query translation" component very similar to the SRU/Z39.50 gateway already implemented in TEL.
The other possibility is that clients directly access the online translation service using the Web API. This solution would require no "isolated query translation" component at all in the TEL system and, as a consequence, the infrastructural/hardware cost would be lowered. In addition, this solution would be very similar to the way in which SRU national libraries are queried today, expect that the proprietary Web API would have to be used instead of SRU. Note that

[6] http://www.systransoft.com/index.html
[7] http://www.autonomy.com/content/home/index.en.html
[8] http://www.worldlingo.com/en/products/worldlingo_api.html

this solution might pose some problems with respect to the licensing policies of the online service, which would be accessed not by TEL but directly by the clients.

4.2.3 Free Online Services

This proposed solution uses a free online translation service, such as Babelfish by Altavista[9] or Google Language Tools[10]. Considerations similar to the case of the commercial online service also hold in this case, with the exception of the costs, since the service would be free.
Unlike the previous case, a Web API might not exist and the necessary parameters would have to be sent using a *HyperText Transfer Protocol* (HTTP) request, as with *HyperText Markup Language* (HTML) forms. In this case, the source code of the Web user interface of the translation service has to be inspected to identify the correct parameters to be sent. Similarly, the translation could be received as a HTML page which needs to be parsed to extract the actual translation. This makes the interaction with the translation service dependent on the layout of the HTML pages used for the user interface.
Moreover, the provider of the translation service should be informed that the service is going to be accessed in this way, because the number of requests could be higher than usual and TEL would be using a third-party service to improve its functionalities.
However, this solution could be an interesting option for a proof of concept implementation of the "isolated query translation" feature in order to evaluate its functioning and its effectiveness. Once the "isolated query translation" feature has proved to be useful, one of the other options described in this section should be used for the final deployment to end-users.

4.2.4 Ad-hoc Implementation

In this case, the "isolated query translation" component can be implemented by developing an ad-hoc MLIA engine which makes use of proper MLIA algorithms and techniques, such as those described in [19] and [23].
This solution guarantees the full control on the MLIA process which can be fine-tuned to the needs of the TEL community. Moreover, all the data structures and the information resources used by the MLIA engine would be available for future purposes. The same MLIA engine could also be used for implementing the "pseudo-translation" of the records contained in the TEL central index, as described in [8].
This solution would require purchasing the necessary linguistic resources, such as *Machine Readable Dictionaries* (MRD), and a certain effort would be needed to implement the MLIA engine. Furthermore, the effor needed to continuously update the lexical resources should not be forgotten. On the other hand, it would offer the possibility of building MLIA competence and background within the TEL team, which would be a solid basis also for future development of the system. Moreover, the effort required to create the necessary MLIA competence and to develop the MLIA engine could be reduced by addressing them in the context of a European research project, as suggested by the i2010 Digital Library Initiative [16], where other partners with stronger MLIA competence can cooperate with TEL people and the resources built can be reusable in other European funded contexts.

[9] http://babelfish.altavista.com/
[10] http://www.google.com/language_tools?hl=en

5 Pseudo-Translation of Expanded Records

5.1 Situation

Today, the large majority of all records available through the search facility in TEL contain bibliographical metadata only (no abstracts or full text). Only short segments of text are thus available for free-text search by the user.

While potentially a problem in monolingual search as well, the brevity of the available text exacerbates the problems usually encountered in multilingual information access situations.

5.2 Expansion Techniques

The solution that was chosen for overcoming the lack of textual content in the information items is expansion of the content fields. The approach used is derived from techniques used in classical information retrieval for query expansion, such as relevance feedback [15] and local context analysis [28]. In their classical form, such techniques rely on a user identifying a number of information items that are relevant to his/her information need. These additional items are used as a source to extract terms that are thought to statistically discriminate well between them and other items in the collection. Such terms are usually close in meaning to the original search terms used in the query, they are often "quasi-synonyms" or related terms. Use of the techniques in a system requires the user to go through an extra interactive step: initially, he/she starts a search by formulating an original query, and then he/she identifies some of the information items that are returned as relevant, in order to allow the query refinement by the system. Such techniques have been repeatedly shown to be beneficial in boosting retrieval effectiveness of monolingual search systems.

These techniques can also be applied without any additional involvement of the user. In such cases, the system automatically assumes that the top items returned in response to user requests are relevant, and uses these items for the term extraction, in place of items manually selected by the user. Such a technique is called "blind feedback", and has been proven to be beneficial in many CLIR settings [1], [5].

It is possible to use the same techniques independently of specific information needs, by expanding the information item itself instead of the query. While usually not applicable to retrieval on lengthy documents, such an approach was identified as promising at the Paris and The Hague 2006 DELOS/TEL meetings, and thus stands at the center of our work.

Two of the main problems involved in CLIR are lack of vocabulary coverage (i.e. terms can not be translated due to insufficiently large translation resources) and word sense ambiguity (i.e. terms have potentially different meanings, and it is unclear how to determine the correct one, leading to wrong translations).

By using expansion techniques, we intended to address both problems. Additional terms added during expansion tend to be from a more general vocabulary, as term frequency influences term selection. The new, longer representation of the record also makes it less likely that none of the terms can be translated. Since the record is used in its entirety to drive expansion, the presumption is that the added terms correctly represent the overall concept of the record, thus helping to disambiguate polysemous terms.

5.3 Pseudo-Translation

In this proposed solution, we cross the language boundary by translating the documents to match the query. While the less frequent choice compared to query translation in academic literature, document translation has been found to be very competitive [5] in some general settings. The main reason for the scarse adoption of the document translation techniques can be attributed to problems of scalability. It is currently very hard to translate large sets of lengthy documents with appropriate efficiency. However, this problem is much less pronounced in the case of current case of TEL; the brevity of the records should make document translation applicable even to large numbers of records, e.g. in the order of multiple millions of records. Furthermore, the gain in efficiency during querying of the system is considerable and likely of high importance for the TEL system. To translate the records, and make the translations searchable, access to the actual records and the creation of a new search index is necessary. This makes the approach directly suitable for integration with the TEL central index; the same approach could however also be deployed in other search systems of the national libraries that are accessed via TEL.

Document translation has been chosen in order to make best use of the data available for matching records and queries; since the records, while short, are slightly longer than typical queries to the TEL system, it was believed that they would less suffer from vocabulary coverage problems. Additionally, document translation can be performed offline, thereby largely independent of the retrieval system later used for search; decoupling the tools used for the feasibility study from the system used by TEL.

Using the translated records for matching with queries only, and not for presentation means that we can use “pseudo-translations”, i.e. to potentially leave terms untranslated or translate them into multiple different terms in the target language. There is no need to produce a syntactically correct or semantically unambiguous document as the translation will remain hidden to the end user. This approach of using “rough” translations for retrieval is both cheaper to implement and often more effective for retrieval, as multiple translation alternatives can be retained during the process. A machine-readable dictionary is suitable for pseudo-translation; but we chose a machine translation system for licensing reasons. This choice of translation resource will be further commented on later in the report.

5.4 Outline of Approach

The outline of this approach, called in the following “pseudo-translation of expanded records”, or “pseudo-translation” for short, is thus:

- the unstructured content fields of the record (usually only the title, as well as some keyword fields that can be treated as unstructured content) are expanded by additional terms that are statistically similar to the original terms in those fields
- these additional terms are derived by searching for records that are similar in content to the record that is to be expanded, and then extracting from these additional records the best terms according to well-known blind feedback techniques
- the expanded terms are then used during the retrieval phase
- the expanded records are translated in their entirety using a simple translation resource
- retrieval takes place on the expanded, translated records

Figure 12 shows a modified architecture of the TEL system that includes both the “isolated query translation” approach and the “pseudo-translation” approach discussed in the following.

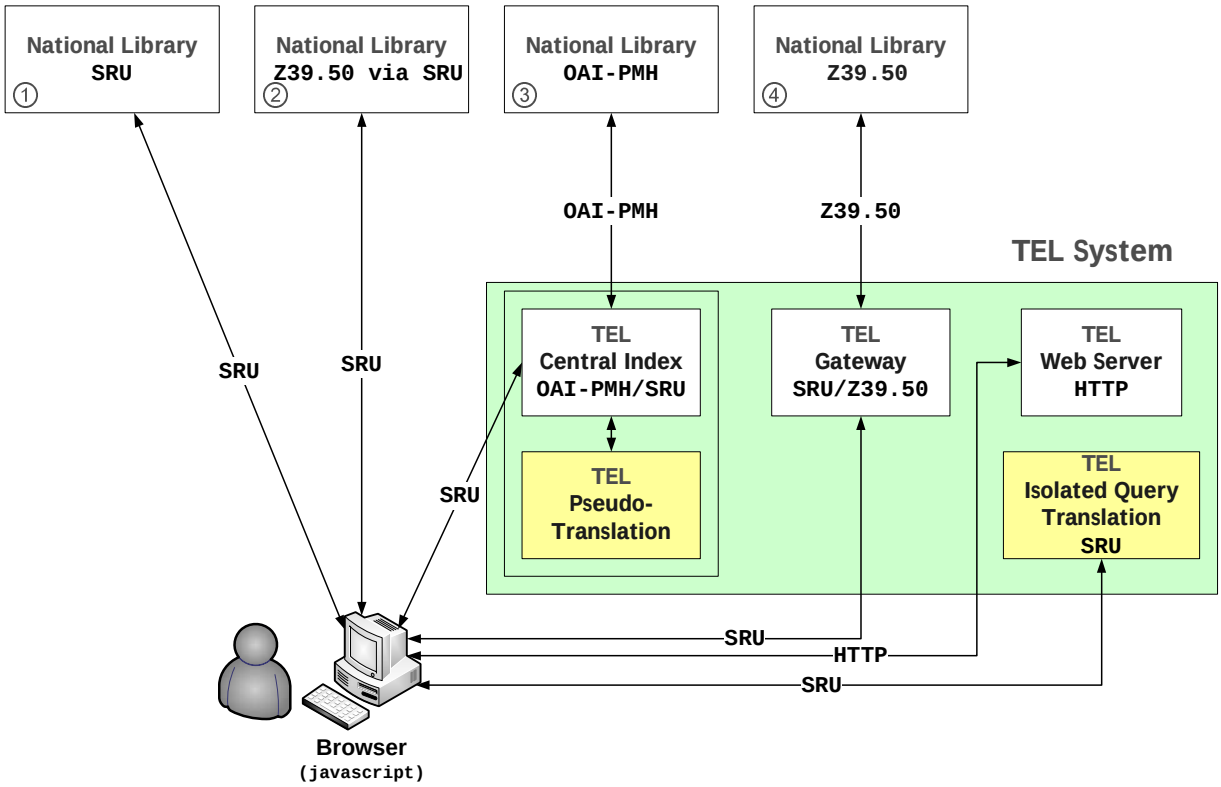

Figure 12: Architecture of TEL system extended by "pseudo-translation" approach (integrated with the central index) and "isolated query translation".

5.5 Experiments

The intention of the feasibility study was to demonstrate how expanded records could be represented, how they would look in their pseudo-translated state and to analyze whether they could be expected to be usable for implementing CLIR in the TEL system.
Fast progress and the availability of key software tools allowed us to redefine these objectives to a more ambitious and powerful evaluation of pseudo-translation and expansion techniques.
Thus, a full evaluation on a sample of 151,700 bibliographical records from TEL using nearly 100 queries derived from logfiles to represent typical information needs was carried out. The retrieval system used was Terrier, an open-source information retrieval system developed by the University of Glasgow. Terrier is available under the Mozillla Public License. Note that the choice of system for the experiments is secondary, as much of the procedures described would be implemented off-line in an operational system. Terrier was chosen as it has modern feedback techniques implemented. Any other probabilistic information retrieval system offering similar feedback algorithms could be used in its place. Translation of the records was done using the PROMT[11] off-the-shelf machine translation system. Again, as described later, a variety of different translation resources could be used in support for the chosen CLIR approach.

5.5.1 System

The Terrier[12] system is freely available under the Mozilla Public License. We used version 1.0.2 for our experiments. Terrier was chosen as a prototype system supporting both state-of-the-art probabilistic retrieval and query expansion techniques. Terrier could be replaced in an

[11] http://www.e-promt.com/
[12] http://ir.dcs.gla.ac.uk/terrier/

operational version of the chosen "pseudo-translation" approach by any other system offering similar retrieval features. We found Terrier to be generally suited to the task, with two major exceptions:

1. no appropriate support for non-US-ASCII characters. This had to be circumvented by pre-coding the records to only use 7bit characters
2. a bug in handling query expansion in certain situations that only arises when using large numbers of very short documents (as is unfortunately the case for the experiments in this report). This bug was fixed by changing the Terrier source code. The fix was reported back to the developers through their support forum

5.5.2 Test Collection

The experiments of the study use a "test collection". In information retrieval, effectiveness of retrieval systems is most commonly measured by employing the so-called Cranfield paradigm [12]. Tests are conducted using a fixed collection of documents (bibliographical records in this case), a number of information needs represented by queries, and relevance judgments linking the information needs to the relevant items in the document collection.

The building of such document collections is costly, because in theory one relevance judgment for every document-query pair is need. For 1,000,000 documents and 100 queries this means 100,000,000 judgments, all to be made by expert human relevance assessors. This is clearly unpractical, and there are different alternative procedures that are widely accepted by the research community. We have concentrated on building the test collection in terms of documents and queries, and have used various techniques to keep the relevance judgment work to a minimum (see below).

5.5.3 "Document" Base

For our experiments, 151,700 records from the British Library that were provided to us by Bill Oldroyd served as "documents". These records are included in the TEL central index as well, and so serve as a sample of the English language records present in the TEL system. The experiments had to be restricted to English for technical reasons, it was not possible for TEL to provide a sufficiently large number of records in other languages as a database dump to work with, and the respective other national library contacts were not able or willing to help with the study.

The records were delivered in two batches, in MARCXML[13] format. This was transformed using *XSL Transformations* (XSLT) and some custom-built scripts to an encoding compatible to the Search/Retrieve Web Service (SRW)[14], and enriched with unique document identifiers. The resulting *eXtensible Markup Language* (XML) file was stripped of fields that had no relevance to the further analysis. Only the fields for title and subject keywords were retained. For samples of the resulting XML records see below in the analysis of results.

Only one of the batches delivered by the British Library had language identifiers. For the other batch, a simple heuristic using a list of common words of the English language (so-called stop words) was used to filter out non-English records. The resulting extract seems to be sufficiently clean. After all conversion and filtering steps, a collection of 151,700 records resulted for further analysis.

[13] `http://www.loc.gov/standards/marcxml/`
[14] This is a companion protocol to the SRU protocol used in TEL. See http://www.loc.gov/standards/sru/srw/

5.5.4 "Document" Characteristics

Number of documents: 151,700
Number of documents with title field: 151,700
Number of documents with subject keywords: 69,924
Average number of tokens per document: 10.9
(note that these are term occurrences, not unique terms. Stopwords are excluded)
Average number of unique terms per document: 7.6

5.5.5 Queries

The queries were selected to represent information needs by users of the TEL system. We worked with query logfiles provided by Eric van der Meulen from TEL. The logfiles covered roughly 3 months (March-May 2006) with 637,118 query instances. Of these, 285,407 query instances were of the same query that is known to be used as test query by the TEL administrators. These instances were excluded from further consideration, leaving 351,711 query instances.

Using these logfiles, 100 English-language queries were selected for the experiments. We aimed to select around half the queries as one-word statements and another half as multi-word statements based on frequency in the logfiles. The distinction was made as frequent queries tend to be very short (typically one-word), and selection based only on frequency would not have yielded many multi-word queries. The average query length in TEL is 2.2 words (2.5 words if only unique queries are considered). Unfortunately, a query was duplicated, leaving effectively only 99 unique queries for further testing (see Table 1).

Table 1: 100 queries extracted from the TEL logfiles and used for subsequent experiments. Given are the English query, the German hand-made translation, and the number of occurrences of the query in the logfiles

ID	English	German	Frequency
1	bible	Bibel	1025
2	art	Kunst	788
3	computer	Computer	385
4	music	Musik	274
5	islam	Islam	188
6	marketing	Marketing	179
7	architecture	Architektur	148
8	religion	Religion	111
9	history	Geschichte	92
10	psychology	Psychologie	91
11	sport	Sport	87
12	medicine	Medizin	86
13	economics	Wirtschaft	82
14	love	Liebe	76
15	network tourism	Netzwerk Tourismus	69
16	beer	Bier	65
17	modern language notes	Anmerkungen zu moderner Sprache	56
18	education	Bildung	53
19	business	Geschäft	52
20	computer science	Informatik	50
21	biology	Biologie	47
22	database systems	Datenbanksysteme	43
23	coins	Münzen	37
24	public health	Volksgesundheit	36
25	internet	Internet	36
26	geography	Geographie	35
27	nature conservancy council	Naturschutz Kollegium	35
28	simulations, evaluations models	Simulationen, Evaluationsmodelle	34
29	chess	Schach	33
30	police	Polizei	33
31	laser devices applications	Anwendungen von Lasergeräten	33

ID	English	German	Frequency
32	sociology	Soziologie	32
33	food	Nahrung	31
34	globalization	Globalisierung	31
35	dog	Hund	30
36	agriculture	Landwirtschaft	29
37	newspapers	Zeitungen	29
38	insurance	Versicherung	29
39	homoeopathy	Homöopathie	24
40	liquid crystals	Flüssigkristalle	23
41	artificial intelligence	künstliche Intelligenz	22
42	development law international finance	Entwicklung Gesetz internationales Finanzwesen	21
43	marketing research	Marktforschung	21
44	European parliament	Europäisches Parlament	21
45	photography	Fotografie	21
46	development law international finance	Entwicklung Gesetz internationales Finanzwesen	21
47	oceans	Ozeane	20
48	anatomy	Anatomie	20
49	tactics skills	taktische Fertigkeiten	20
50	world higher education database	Weltweite Datenbank höherer Bildung	19
51	angels	Engel	18
52	music in the renaissance	Musik in der Renaissance	18
53	programming	Programmierung	17
54	issues in holocaust education	Themen Holocaust Unterricht	17
55	medicine health through time	Medizin Gesundheit über die Jahre	17
56	energy	Energie	17
57	world war	Weltkrieg	17
58	the social psychology of groups	Sozialpsychologie von Gruppen	14
59	computer science technology	Informatik Technologie	13
60	how languages are learned	Wie Sprachen gelernt werden	12
61	guide to job interview answers	Handbuch zu Antworten für Bewerbungsgespräche	12
62	world war debt settlements	Weltkrieg Schuldenausgleich	11
63	short history of the German people	Kurze Geschichte des deutschen Volkes	11
64	emerging market capital flows	Wachstumsmärkte Kapitalfluss	10
65	environmental engineering international	Umweltingenieurwissenschaften international	9
66	survival development of humanity	Überleben Entwicklung der Menschheit	9
67	behaviour management in schools	Verhaltensmanagement in Schulen	9
68	Scandinavian economic history review	Überblick der skandinavischen Wirtschaftsgeschichte	9
69	building design and quality of life	Gestaltung von Gebäuden und Lebensqualität	9
70	current legal problems	Aktuelle juristische Probleme	8
71	novelty detection research idea	Neuheiten Erkennung Forschungsidee	8
72	air quality air condition	Luftqualität Luftzustand	8
73	electronic navigation systems	Elektronische Navigationssysteme	6
74	application artificial intelligence finance	Anwendung künstliche Intelligenz Finanz	6
75	importance of scientific research	Bedeutung wissenschaftlicher Forschung	6
76	human rights law journal	Menschenrechte Gesetz Journal	6
77	risk management it projects	Risikomanagement IT Projekte	6
78	modern methods in mathematics	Moderne Verfahren in der Mathematik	6
79	young parents education Scotland	Junge Eltern Bildung Schottland	6
80	what kind of eagle is the national bird of the usa	Welche Adlerart ist der Nationalvogel der USA?	6
81	the future of city tourism in Europe	Die Zukunft des Städtetourismus in Europa	6
82	noise as a public health problem	Lärm als ein Volksgesundheitsproblem	6
83	using tax incentives to conserve enhance biodiversity in Europe	Verwendung von Steueranreizung zum Erhalt der Biodiversität in Europa	6
84	ecotourism in the national parks of Latin America	Ökotourismus in den Nationalparks Lateinamerikas	6
85	china's place in global geopolitics	Chinas Platz in der globalen Geopolitik	6
86	how to become a clinical research associate	Wie man medizinischer wissenschaftlicher Mitarbeiter wird	6
87	state of the art of surgery ...	Aktueller Forschungsstand in Chirugie	6
88	theories of mental development the problem	Theorien über geistige Entwicklung das Problem	5
89	war society in late medieval Britain	Kriegsgesellschaft im spätmittelalterlichen Grossbritannien	5
90	HIV aids global governance	HIV AIDS Globale Lenkungsformen	4
91	what is worth teaching	Was sich zu lehren lohnt	3
92	why we have law	Wieso gibt es Gesetze?	3
93	European royal family benefits to maintaining royal family	Europäische Königsfamilien Vorteile des Erhalts königlicher Familien	3
94	how to do things with words	Wie man Dinge mit Worten erreicht	3

ID	English	German	Frequency
95	banking their activity regarding the loans	Bankwesen ihre Aktivitäten betreffs Krediten	3
96	German commanders of world war ii	Deutsche Kommandanten des Zweiten Weltkriegs	3
97	saving our students saving our schools	Die Studenten retten die Schulen retten	3
98	approach to English literature for students abroad	Zugang zu englischer Literatur für Studenten im Ausland	3
99	hierarchy of norms in community legal order	die Hierarchie der Regeln in der Rechtsordnung von Gemeinden	3
100	human body what the human body is made up of	Menschlicher Körper Woraus besteht der menschliche Körper	3

The queries were manually translated into German for later cross-language retrieval experiments.

5.5.6 Expansion

The Terrier system was used to expand the 151,700 records by using each record in turn as a query and running it against the whole collection to determine the set of most similar records. Based on this set of records, the statistically most discriminating terms were determined.
This process was run overnight. It is an off-line process, which can be redone periodically when the collection has significantly altered (in the case of a system like TEL, redoing the process every few months should be sufficient).
Different parameters for expansion were tried:

a) 10 best-ranked documents, maximum of 5 expansion terms, leading to 1.8 terms added on average
b) 10 best-ranked documents, maximum of 10 expansion terms, leading to 2.1 terms added on average
c) lower threshold for terms inclusion, leading to 3.7 terms added on average

These numbers constitute a roughly 25-45% expansion over the length of the original records in terms of unique, content-bearing terms. Closer inspection showed that the best additional terms clearly came from parameter setting a), with b) also contributing interesting terms. Setting c) was too “aggressive” and not pursued further.

Not all records benefit from this expansion process. For some records, no new statistically associated terms can be found, and the records remain unexpanded. In all, 44,647 out of 151,700 records (~29%) were not expanded. This number would likely drop significantly as the available data collection grows, i.e. as more records are added to the index that is used for expansion.

5.5.7 Translation of Records

On the basis of the document collection and queries described, the system was set up for cross-language retrieval. We expanded the 151,700 records as outlined above, and pseudo-translated them from English to German using the PROMT machine translation system. While as a machine translation system PROMT tries to produce grammatical output in the target language, we have to argue that the translation was still not of superior quality compared to using dictionary lookup, as titles seldom lend themselves to good translation (containing lists of keywords, additional information on publishers and authors, etc.).
In fact, it was discovered that PROMT tried much too aggressively to take a German linguistic phenomenon called “compounding” into account. In German, it is possible to form compound nouns by stringing together a number of simple nouns. Such compound nouns express complex concepts and usually are represented as phrasal expressions in English. PROMT clearly uses an algorithmic procedure to produce compound nouns in its translated output,

producing many wrong (and sometimes hilarious) results. Generally speaking, the resulting compounds are too long and unwieldy. Unfortunately, such compound words are detrimental for retrieval, as matches between search terms and records would have to cover the whole compound, unless there is a procedure built into the retrieval system that provides for partial matches (e.g. by splitting the compounds: on the use of "decompounding" for splitting such words, see e.g. [11]).

Some examples for bad compound formation caused by PROMT are given in Table 2 and include:

Table 2: Examples for bad compound formation during translation in PROMT. Note that the automatic translation also suffers from the fact that the original expressions are not necessarily grammatical sentences, but can be lists of key words.

Record Position	English	German tanslation by PROMT	potentially correct hand-made translation
25	Non-domestic cleaning substances & machines	Nichtinnenreinigungssubstanz-Maschinen	industrielle Reinigungsmittel und –ma-schinen
283	Diabetes technology & therapeutics	Zuckerkrankheitstechnologietherapeutik	Technologien und Therapien zu Diabetes
1089	market reinsurance news forum	Marktrückversicherungsnachrichtenforum	Rückversicherung von Märkten Nachrichtenforum
1165	FOOD INDUSTRY KEY NOTE	NAHRUNGSMITTELINDUSTRIE-SCHLÜSSELZEICHEN	Nahrungsmittelindustrie (grundlegende Gedanken)
7855	Hydro-thermal unit commitment	Hydrothermaleinheitsengagement	hydrothermische Einheit Haftung
10140	Gravitational scatter theory	Gravitationsstreuungstheorie	Theorie der Streuung der Gravitation
16140	multi-bodies compensation systems	Mehrkörperentschädigungssysteme	Kompensationssysteme für Mehrkörperschaften
17875	renal transplant population	Nierenverpflanzungsbevölkerung	Nierentransplantation Bevölkerung
18496	Motion sickness incidence on sea-going passenger vessels	Reisekrankheitsvorkommen auf Hochseepersonenbehältern	Vorkommen von Reisekrankheit auf Hochsee-Passagierschiffen
18884	Stress corrosion cracking	Betonungskorrosionsknacken	Belastung Korrosion Rissbildung
20807	Growth hormone pharmacology	Wachstumshormonarzneimittellehre	Pharmakologie Wachstumshormone
22010	business information databases	Geschäftsinformationsdatenbanken	Datenbanken für Geschäftsinformation
22215	Trade facilitation information	Handelserleichterungsinformation	Handelserleichterungen Information
91598	Schools Students Behaviour problems	Schulstudentenverhaltensproblem	Schulen Verhaltensprobleme von Studenten

The problem was countered by forcing PROMT to proceed word by word, thus preventing compound formation. This was clearly beneficial for retrieval.

Some examples of pseudo-translated records. Shown is also the difference when compound formation in PROMT is suppressed (see <title2> field vs. <title> field):

Orginal Record:

```
<srw_dc:dc>
<recordPosition>1612</recordPosition>
<title>Pioneer landowners in the Lane Cove district 1794-1796
/</title>
<subject>Landowners New South Wales Lane Cove History 18th
century.</subject>
</srw_dc:dc>
```

Translated Record:

```
<srw_dc:dc>
<recordPosition>1612</recordPosition>
<title>Pioniergrundbesitzer im Gasse-Bezirk der Kleinen Bucht 1794-
1796/</title>
```

```
<title2>Bezirk, 1796, 1794, Pionier, Gasse, Grundbesitzer, kleine
Bucht, </title2>
<subject>Grundbesitzer das Neue Südliche 18. Gasse-Geschichts-
Jahrhundert der Kleinen Bucht von Wales.</subject>
<extendedTerms></extendedTerms>
</srw_dc:dc>
```

Orginal Record:

```
<srw_dc:dc>
<recordPosition>1785</recordPosition>
<title>Human rights and women of North East India /</title>
<subject>Women's rights India, Northeastern Congresses.</subject>
<subject>Human rights India, Northeastern Congresses.</subject>
<subject>Women India, Northeastern Social conditions
Congresses.</subject>
</srw_dc:dc>
```

Translated Record:

```
<srw_dc:dc>
<recordPosition>1785</recordPosition>
<title>Menschenrechte und Frauen des Nordöstlichen Indiens/</title>
<title2>Rechte, Indien, Mensch, Norden, Osten, Frauen, </title2>
<subject>Frauenrechte Indien, Nordöstliche Kongresse.</subject>
<subject>Menschenrechte Indien, Nordöstliche Kongresse.</subject>
<subject>Frauen Indien, Nordöstliche Soziale
Bedingungskongresse.</subject>
<extendedTerms></extendedTerms>
</srw_dc:dc>
```

The resulting 151,700 pseudo-translated records were loaded into the Terrier system for retrieval.

5.5.8 Cross-Language Retrieval

The 100 queries were hand-translated into German and used to retrieve the top 10 records for each query from the pseudo-translated German records. This constitutes a cross-language retrieval experiment, as each pseudo-translated record can clearly be matched with the original English version it represents in the search index. As a baseline for comparison, we ran the same 100 queries in their original English version against the original English records.

5.6 Analysis of results

Clearly, it was not feasible to do extensive relevance assessments for all 100 queries (resulting in 151,700 * 100 assessments, with a typical time estimate of roughly 1 minute per judgment – an effort of 28.9 person years consisting of 24 hour days!). We used an alternative method called “overlap analysis”. The monolingual baseline of using English queries on English records represents the same information needs as the cross-language experiment of using the German queries on German pseudo-translated records. It is therefore possible to use the monolingual English baseline as a “gold standard”, assuming that the results from that retrieval experiment have sufficient quality.

Using this “gold standard”, any retrieval result for a query from the cross-language experiment that is sufficiently similar to the monolingual result is assumed to be acceptable. This assumption is further strengthened by the fact that the two runs essentially form two different sources for evidence that the same records should be highly ranked. We have considered the top 10 ranked records to determine the similarity between the monolingual and the cross-language experiment. We have used this technique to exclude such queries from further

analysis, therefore driving down overall workload. In all, 30 of the 100 queries had sufficiently similar results, and were thus excluded.

The remaining 70 queries have results that significantly differ from the monolingual baseline. This, however, does not necessarily indicate that these queries have poor performance. For further analysis, four cases need to be distinguished:

Case 1: good monolingual result; good, but different, cross-language result
Case 2: good monolingual result; bad cross-language result
Case 3: bad monolingual result; good cross-language result
Case 4: bad monolingual result; bad, but different, cross-language result.

The four cases can be supplemented with the previously described case:

Case 0: monolingual and cross-language result similar; assumed to be good

We have attempted to classify the remaining 70 queries to Cases 1-4 based on relevance assessments of the top 10 records for both the monolingual and cross-language case. This would have in total meant 70 * 20 = 1400 judgments. In combination with the actual analysis of the results, it was not possible to process all 70 remaining queries. A total of 18 queries had to be excluded from further processing due to lack of resources.

Overall, the queries were classified as indicated in Table 3.

Table 3: Classification of queries

Similar retrieval result, assumed to be good	Case 0: queries 1,2,4,5,6,7,8,10,11,14,16,20,21,25,29,32,34,37,39,41,43,45,46,48,52,59,74,83,90,96 (30)	
	Good cross-language retrieval	**Bad cross-language retrieval**
Good monolingual retrieval	Case 1: queries 3, 9, 12, 33, 35, 36, 38, 47, 51, 53, 55, 56, 57, 67 (13)	Case 2: queries 13, 15, 18, 22, 28, 84, 31, 42, 44, 58, 65, 69, 75, 88 (14)
Bad monolingual retrieval	Case 3: queries 54, 66 (2)	Case 4: queries 17, 24, 26, 40, 49, 50, 60, 61, 62, 63, 64, 68, 70, 71, 82, 86, 87, 89, 91, 92, 93, 94, 97 (23)

We thus analyzed a total of 82 queries.

We argue that Case 0, 1, and 3 provide evidence for good retrieval results, whereas Case 4 at least indicates that the cross-language result is not necessarily worse than the monolingual result (for distribution of the cases, see also Figure 12).

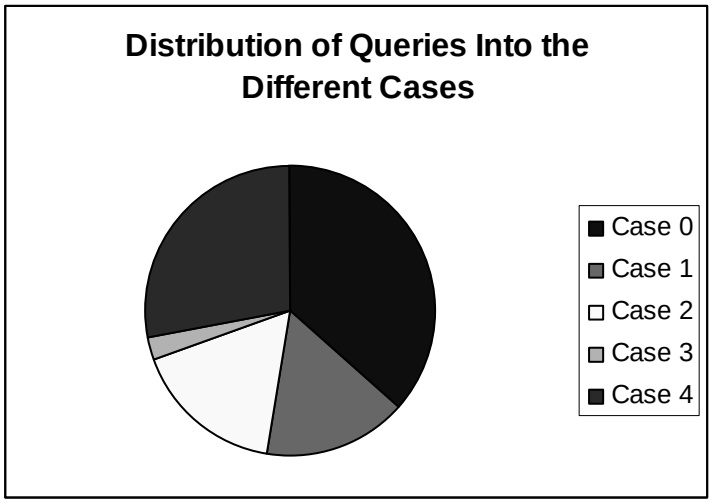

Figure 12: Graphical distribution of the queries into the different cases for analysis. Cases 0, 1 and 3 are assumed to be good (blue coloring), while Case 4 is assumed to be not substantially worse than monolingual (purple coloring)

In all, this means that 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, it is in line with what has been reported as state-of-the-art for CLIR in the CLEF campaign on lengthy documents [6]. Of course, the number has to be treated with care, as it is based on various assumptions (see above) and as the number of relevance judgments it is based on is severely limited (see also above). There is also reason to believe that the approach will actually benefit in terms of effectiveness when scaling up to larger collections (see below in the Conclusions).

5.6.1 Query-by-Query-Analysis

In the following section we delve deeper into some results by an analysis of single queries. Query-by-query analysis makes it possible to get a more detailed look at the different issues that arise during retrieval. The classical approach of looking at overall performance can obscure the behavior of individual queries due to calculation of averages. We believe that robustness of performance across as many queries as possible is an important characteristic for operational CLIR/MLIA systems (see also [6]).

5.6.1.1 Compound problems

As mentioned earlier, PROMT forms compound words in German very aggressively. These words do not match search terms that cover only parts of the compounds. We give some examples of queries that degraded in effectiveness due to translation problems with compounds.

Query English: laser device application
Query German: Anwendugen von Lasergeräten

English record, relevant to query:

```
<srw_dc:dc>
<recordPosition>83773</recordPosition>
```

```
<title>Laser applications.</title>
<subject>Lasers.</subject>
<subject>Lasers</subject>
</srw_dc:dc>
```

German pseudo-translation. The compound is wrong and does not match. Partial matches on the constituent word “Laser” would be possible by employing compound splitting.

```
<srw_dc:dc>
<recordPosition>83773</recordPosition>
<title>Laseranwendungen.</title>
<subject>Laser.</subject>
<subject>Laser</subject>
<extendedTerms>Halbleiter, </extendedTerms>
</srw_dc:dc>
```

Query English: behaviour management in schools
Query German: Verhaltensmanagement in Schulen

English record, relevant to query:

```
<srw_dc:dc>
<recordPosition>91598</recordPosition>
<title>Strategies for managing behavior problems in the classroom
/</title>
<subject>Classroom management.</subject>
<subject>Schools Students Behaviour problems Treatment</subject>
</srw_dc:dc>
```

Bad translation to German (bad formation of compounds). Again, best would be to avoid this formation of compounds. Otherwise, partial matches would have to be forced by using a compound splitting component (e.g. on Schule, Problem, etc.)

```
<srw_dc:dc>
<recordPosition>91598</recordPosition>
<title>Strategien für Betriebsverhaltensprobleme im
Klassenzimmer/</title>
<subject>Klassenzimmer-Management.</subject>
<subject>Schulstudentenverhaltensproblem-Behandlung</subject>
<extendedTerms>Disziplin, positiv, Schul-, heute, </extendedTerms>
</srw_dc:dc>
```

5.6.1.2 Queries that benefit from translation

The following query benefits from translation, as the translation process solves a synoym problem (agriculture/farming) as a “by-product”

Query English: agriculture
Query German: Landwirtschaft

English record, NOT found in monolingual setting:

```
<srw_dc:dc>
<recordPosition>67103</recordPosition>
<title>Profitable organic farming /</title>
<subject>Organic farming.</subject>
</srw_dc:dc>
```

German pseudo-translation. Terms match.

```
<srw_dc:dc>
<recordPosition>67103</recordPosition>
<title>Gewinnbringende organische Landwirtschaft/</title>
<subject>Organische Landwirtschaft.</subject>
<extendedTerms></extendedTerms>
</srw_dc:dc>
```

Another example (humanity/mankind):

Query English: survival development of humanity
Query German: Überleben Entwicklung der Menschheit

English record, not retrieved in monolingual setting, relevant:

```
<srw_dc:dc>
<recordPosition>102799</recordPosition>
<title>Progress and survival : An essay on the future of
mankind.</title>
</srw_dc:dc>
```

Pseduo-Translation to German: matches (synonymy mankind/humanity resolved).

```
<srw_dc:dc>
<recordPosition>102799</recordPosition>
<title>Fortschritt und Überleben : Ein Aufsatz auf der Zukunft der
Menschheit.</title>
<extendedTerms>Christentum, </extendedTerms>
</srw_dc:dc>
```

5.6.1.3 Stemming

A number of problems were caused because we did not use any stemming in our experiments. Stemming is used to reduce terms to their base forms, allowing matches during retrieval even if different word forms are used in query and record.

Query English: angels (Mehrzahl)
Query German: Engel

English record, not found in monolingual setting, due to missing stemming:

```
<srw_dc:dc>
<recordPosition>157364</recordPosition>
<title>The Angel Steps in.</title>
</srw_dc:dc>
```

German pseudo-translated record. The plural of „Engel“ is „Engel“, thus there is a match in German.

```
<srw_dc:dc>
<recordPosition>157364</recordPosition>
<title>Der Engel tritt Ein.</title>
<extendedTerms></extendedTerms>
</srw_dc:dc>
```

5.6.1.4 Queries with no relevant records

Some queries have no relevant records associated. The retrieval results between monolingual and cross-language will differ, because it is essentially determined at random which records will match in absence of any good “hits”. None of the results, monolingual or cross-lingual, is “better”, they are essentially both not useful.

Examples with no relevant records:
Query English: how languages are learned
Query German: Wie Sprachen gelernt werden

Query English: european royal family benefits to maintaining royal family
Query Deutsch: Europäische Königsfamilien Vorteile des Erhalts königlicher Familien

5.6.1.5 Large number of relevant records

Some queries have many relevant records associated with them. Again, it is not surprising that monolingual and cross-language systems do not necessarily provide similar results: both systems retrieve a different subset of relevant records at high ranking positions. As the analysis is by necessity limited to highly-ranked documents, it is impossible to tell which system performs better (please note that for many users that are "precision-oriented", i.e. that search for a limited number of highly relevant items, the systems are in fact equally useful)..

Examples with many relevant records
Query English: medicine
Query German: Medizin

But note: the monolingual system in such scenarios has potentially returned relevant documents that are not found by the multilingual system (e.g. due to translation errors)

5.6.1.6 Problems with weighting during retrieval

Example
Query English: theories of mental development the problem
Query German: Theorien über geistige Entwicklung das Problem

Weighting problems are the reason for poor retrieval if result lists are dominated by individual terms from the query, whereas the original information need is only expressed adequately by the query as a whole (but there is reason to believe that good records do exist; the other system may have provided proof of this). More weighting problems were also caused due to a technical glitch in our setup, which prevented elimination of certain non content-bearing words from retrieval.

5.6.1.7 British English vs. American English

Translation can smooth out issues of British vs. American English, as exemplified by the following query (behaviour/behavior → Verhaltens-)

Query English: behaviour management in schools
Query German: Verhaltensmanagement in Schulen

English record, not found, but relevant:

```
<srw_dc:dc>
<recordPosition>49919</recordPosition>
<title>Behavior management in the schools : a primer for parents
/</title>
<subject>School discipline.</subject>
<subject>Parent-teacher relationships.</subject>
<subject>Child rearing.</subject>
<subject>Behavior modification.</subject>
<subject>Behavior Therapy in infancy & childhood.</subject>
<subject>Child Behavior Disorders therapy.</subject>
<subject>Parent-Child Relations.</subject>
<subject>Schools.</subject>
</srw_dc:dc>
```

German pseudo-translated record.

```
<srw_dc:dc>
<recordPosition>49919</recordPosition>
<title>Verhaltensmanagement in den Schulen : eine Zündvorrichtung fu4r
Eltern/</title>
<subject>Schuldisziplin.</subject>
```

```
<subject>Elternteillehrer-Beziehungen.</subject>
<subject>Kindererziehung.</subject>
<subject>Verhaltensmodifizierung.</subject>
<subject>Verhaltenstherapie in der Säuglingsalter-Kindheit.</subject>
<subject>Kinderverhaltensunordnungstherapie.</subject>
<subject>Elternteilkinderbeziehungen.</subject>
<subject>Schulen.</subject>
<extendedTerms>Kinder, Handbücher, affective, emotional, das
Verstehen, </extendedTerms>
</srw_dc:dc>
```

5.6.1.8 Problems of Synonymy

We found some queries where the translation can be mapped to a number of synonyms. Below we show a query referring to “job interviews”. Both “Vorstellungsgespräch” and “Bewerbungsgespräch” are perfectly reasonable translations. Unfortunately, PROMT decides on one of these alternatives (in this case, on the one that is different from the translated query) and does not allow the inclusion of both translations. The use of a machine-readable dictionary instead of machine translation could remedy such situations, as would using a synonym list during indexing phase.

Example:
Query English: guide to job interview answers
Query German: Handbuch zu Antworten für Bewerbungsgespräche

English record, relevant:

```
<srw_dc:dc>
<recordPosition>73502</recordPosition>
<title>The 250 job interview questions you'll be most likely be asked
: and the answers that will get you hired!.</title>
<subject>Employment interviewing.</subject>
<subject>Applications for positions.</subject>
</srw_dc:dc>
```

German pseudo-translation. Not found due to use of alternative synonym (Vorstellunsgespräch vs. Bewerbungsgespräch).

```
<srw_dc:dc>
<recordPosition>73502</recordPosition>
<title>Die 250 Vorstellungsgespräch-Fragen werden Sie am
wahrscheinlichsten sein gefragt zu werden: Und die Antworten, die Sie
anstellen lassen werden!.</title>
<subject>Das Arbeitsinterviewen.</subject>
<subject>Anwendungen für Positionen.</subject>
<extendedTerms>Radio, </extendedTerms>
</srw_dc:dc>
```

5.6.1.9 Quality of expanded terms

We give below examples of queries that benefit from inclusion of new terms during our expansion step.

Query English: issues in holocaust education
Query German: Themen Holocaust Unterricht

English record, relevant, not found:

```
<srw_dc:dc>
<recordPosition>146644</recordPosition>
<title>A Theology of Auschwitz.</title>
</srw_dc:dc>
```

Pseudo-translated, expanded German record. The inclusion of “holocaust” enables the retrieval of the record, which cannot be directly found in the unexpanded English representation.

```
<srw_dc:dc>
<recordPosition>146644</recordPosition>
<title>Eine Theologie von Auschwitz.</title>
<extendedTerms>doktrinell, blackwell, Holocaust, </extendedTerms>
</srw_dc:dc>
```

Query English: survival development of humanity
Query German: Überleben Entwicklung der Menschheit

English record, relevant, not found:

```
<srw_dc:dc>
<recordPosition>103899</recordPosition>
<title>Private power : Multinational corporations for the survival our
planet.</title>
</srw_dc:dc>
```

German record, pseudo-translated, expanded. The inclusion of “Entwicklung” (development) as an expanded term boosts the score of the record, ensuring a high ranking.

```
<srw_dc:dc>
<recordPosition>103899</recordPosition>
<title>Private Macht : Multinationale Vereinigungen für das Überleben
unser Planet.</title>
<extendedTerms>Entwicklung, Welt, </extendedTerms>
</srw_dc:dc>
```

5.6.1.10 Problems with expanded terms

In contrast to the good examples shown in the last section, there are examples of bad expansion.

Query English: energy
Query German: Energie

English record, not relevant:

```
<srw_dc:dc>
<recordPosition>61636</recordPosition>
<title>Menu-converter.</title>
</srw_dc:dc>
```

Pseduo-translated expanded record, irrelavant. Probably “converter” was associated with “energy”.

```
<srw_dc:dc>
<recordPosition>61636</recordPosition>
<title>Menükonverter.</title>
<extendedTerms>Energie, </extendedTerms>
</srw_dc:dc>
```

Query English: economics
Query German: Wirtschaft

English record, not relevant:

```
<srw_dc:dc>
<recordPosition>154973</recordPosition>
<title>The New Social Order in China.</title>
</srw_dc:dc>
```

German pseudo-translated, expanded record. The connection between “social order” and “economics” is probably too weak to justify retrieving this record in response to the query. However, the expanded terms leads to exactly such retrieval.

```
<srw_dc:dc>
<recordPosition>154973</recordPosition>
<title>Die Neue Gesellschaftsordnung in China.</title>
<extendedTerms>Wirtschaft, </extendedTerms>
</srw_dc:dc>
```

5.6.1.11 Translation Problems

Example of translation problems

There are few outright translation errors, and they tend not to influence retrieval too much. However, there are numerous issues with translation ambiguity and the lack of stemming.

Query English: world war
Query German: Weltkrieg

English record, relevant:

```
<srw_dc:dc>
<recordPosition>52108</recordPosition>
<title>The causes of World War Three /</title>
</srw_dc:dc>
```

Bad German pseudo-translation (wrong word form, plural): Could be solved by using both stemming and decompounding.

```
<srw_dc:dc>
<recordPosition>52108</recordPosition>
<title>Die Ursachen des Weltkrieg***s*** Drei/</title>
<title2>drei, Welt, Krieg, Ursachen, </title2>
<extendedTerms>diplomatisch, 1939, 1945, München, </extendedTerms>
</srw_dc:dc>
```

Query English: behaviour management in schools
Query German: Verhaltensmanagement in Schulen

English record, relevant:

```
<srw_dc:dc>
<recordPosition>91598</recordPosition>
<title>Strategies for managing behavior problems in the classroom
/</title>
<subject>Classroom management.</subject>
<subject>Schools Students Behaviour problems Treatment</subject>
</srw_dc:dc>
```

Bad German pseudo-translation. The query term “Schulen” (schools) does not match unless dempounding is introduced. The extended term “Schul-“ has potential to alleviate the problem, but only if stemming is added to the system.

```
<srw_dc:dc>
<recordPosition>91598</recordPosition>
<title>Strategien für Betriebsverhaltensprobleme im
Klassenzimmer/</title>
<title2>Klassenzimmer, Probleme, Strategien, das Handhaben, Verhalten,
</title2>
<subject>Klassenzimmer-Management.</subject>
<subject>Schulstudentenverhaltensproblem-Behandlung</subject>
<extendedTerms>Disziplin, positiv, Schul-, heute, </extendedTerms>
</srw_dc:dc>
```

Query English: environmental engineering international
Query German: Umweltingenieurwissenschaften international

English record, relevant:

```
<srw_dc:dc>
<recordPosition>36003</recordPosition>
<title>Environmental engineering.</title>
<subject>Environmental engineering.</subject>
</srw_dc:dc>
```

Pseudo-translated German record, inappropriate terminology. Again, the compound prevents any matches based on “Umwelt” (environment).

```
<srw_dc:dc>
<recordPosition>36003</recordPosition>
<title>Umwelttechnik.</title>
<title2>Technik, Umwelt-, </title2>
<subject>Umwelttechnik.</subject>
<extendedTerms>cibse, Gebäude, </extendedTerms>
</srw_dc:dc>
```

5.6.2 Single Record Analysis

Here below we give tables with more examples of some of the problems we have encountered during retrieval.

5.6.2.1 More examples of good expansion of records

The idea of expanding short records to increase the likelihood of good retrieval matches is central to the approach described in this report. Table 4 shows more good examples for expansion of records that we came across in our analysis. The list is obviously far from complete; it is impossible with the resources available for the study to achieve an exhaustive listing.
Please note how the terms inside the “<extendedTerms>” tag, which were added during expansion, complement the original “<title>” keywords.

Table 4. Additional examples for good record expansion. We give the XML representations of the expanded records. Orginal data is inside the “<title>” and “<subject>” tags, whereas expanded terms are enclosed by “<extendedTerms>” tags.

Record Position	English
501	<srw_dc:dc> <recordPosition>501</recordPosition> <title>STATISTICAL DIGEST - CENTRAL BANK OF BELIZE</title> <extendedTerms>banks, banking, quarterly, financial, review, </extendedTerms> </srw_dc:dc>
507	<srw_dc:dc> <recordPosition>507</recordPosition> <title>International encyclopaedia of laws.</title> <subject>Environmental law, International.</subject> <extendedTerms>policy, cooperation, </extendedTerms> </srw_dc:dc>
9978	<srw_dc:dc> <recordPosition>9978</recordPosition> <title>Testing for HIV and AIDS : The next five years.</title> <subject>[AIDS testing market trends]</subject> <extendedTerms>infections, disease, global, aspects, </extendedTerms> </srw_dc:dc>
18151	<srw_dc:dc> <recordPosition>18151</recordPosition>

Record Position	English
	<title>The implementation of a COBOL language enhancement feature.</title> <extendedTerms>programming, computer, program, languages, structured, </extendedTerms> </srw_dc:dc>
20234	<srw_dc:dc> <recordPosition>20234</recordPosition> <title>Business guide to avoiding environmental liability /</title> <extendedTerms>responsibility, products, management, practice, </extendedTerms> </srw_dc:dc>
22329	<srw_dc:dc> <recordPosition>22329</recordPosition> <title>Nano : the emerging science of nanotechnology : remaking the world - molecule by molecule /</title> <extendedTerms>electron, molecular, collisions, molecules, </extendedTerms> </srw_dc:dc>

5.6.2.2 Examples of bad vocabulary coverage

One of the key concerns in proposing a methodology for CLIR/MLIA in TEL was vocabulary coverage during translation. As expected, the vocabulary used especially in the titles of the records is very rich. Inevitably, some of these terms cannot be translated. As can be seen in the expanded, pseudo-translated records, there is additional "material" available for retrieval. As long as the extended terms are good additional descriptors of the record, retrieval of the record may be possible by matching on them instead of the original title terms.

Table 5. Additional examples of failed translation due to lacking vocabulary coverage. Original English records and pseudo-translated German are given. Out-of-vocabulary terms are highlighted.

Record Position	English	German
6291	<srw_dc:dc> <recordPosition>6291</recordPosition> <title>The resistance welding of coated steel.</title> <subject>[Steel weldability]</subject> </srw_dc:dc>	<srw_dc:dc> <recordPosition>6291</recordPosition> <title>Das Widerstand-Schweißen von gekleidetem Stahl.</title> <subject>[Stahl **weldability**]</subject> <extendedTerms>Korrosion, gmaw, mild, das Schmelzen, fleckenlos, </extendedTerms> </srw_dc:dc>
7005	<srw_dc:dc> <recordPosition>7005</recordPosition> <title>Genetics of pigmented secondary metabolites in Streptomyces coelicolor.</title> </srw_dc:dc>	<srw_dc:dc> <recordPosition>7005</recordPosition> <title>Genetik von **pigmented sekundärem metabolites** in **Streptomyces coelicolor**.</title> <extendedTerms>â, molekular, 2, Gen, genetisch, </extendedTerms> </srw_dc:dc>
9406	<srw_dc:dc> <recordPosition>9406</recordPosition> <title>Substitution of polyhalogenoaromatic compounds by sterically hindered nucleophiles.</title>	<srw_dc:dc> <recordPosition>9406</recordPosition> <title>Der Ersatz von Polyhalogenoaromatic-Zusammensetzungen durch **sterically** hinderte

Record Position	English	German
	<subject>[Aromatic compound substitution]</subject> </srw_dc:dc>	nucleophiles.</title> <subject>[Aromatischer zusammengesetzter Ersatz]</subject> <extendedTerms>**organosilicon, nucleophilic**, Reaktionen, mechanistisch, Dose, </extendedTerms> </srw_dc:dc>
16237	<srw_dc:dc> <recordPosition>16237</recordPosition> <title>Caledonian tectonics from stratigraphy and isotope geochemistry of Lower Palaeozoic successions.</title> <subject>[Geology of S. Scotland]</subject> </srw_dc:dc>	<srw_dc:dc> <recordPosition>16237</recordPosition> <title>Kaledonische Tektonik von **stratigraphy** und Isotop-Geochemie von Niedrigeren Paläozoischen Folgen.</title> <subject>[Geologie von S. Schottland]</subject> <extendedTerms>Colorado, stabil, System, Studien, </extendedTerms> </srw_dc:dc>
16477	<srw_dc:dc> <recordPosition>16477</recordPosition> <title>Solid phase compaction of polymeric powders.</title> <subject>[Plastic powder compaction]</subject> </srw_dc:dc>	<srw_dc:dc> <recordPosition>16477</recordPosition> <title>Feste Phase **compaction** polymerer Puder.</title> <subject>[Plastikpuder compaction]</subject> <extendedTerms>Explosivstoff, Boden, Metall, </extendedTerms> </srw_dc:dc>

6 Conclusions and Future Work

We have described the results and the findings of a feasibility study carried out to determine how multilingual information access functionalities could be added to the TEL system, in compliance with the recent research directions recommended by the European Commission in the i2010 Digital Library Initiative.

We studied the present architecture of the TEL system and its functioning and, according to the results of this analysis, 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.

It should be noted that the two proposed approaches are neither separate nor mutually exclusive. On the contrary, they aim at addressing two aspects of the TEL system that can be considered as distinguishing features. On the one side the TEL capability of directly querying national libraries can be enhanced with MLIA functionalities by using the “isolated query translation” approach; on the other side the search functionalities of the TEL central index, which harvests catalogue records from national libraries, can be extended with MLIA capabilities by adopting the “pseudo-translation” approach. As a consequence, the two

proposed approaches complement each other and offer a coherent multilingual extension of the TEL characteristics mentioned above.

6.1 Considerations on the "Isolated Query Translation Approach"

The "isolated query translation" approach allow us to easily extend the current TEL infrastructure in a transparent way for the participating national libraries, still offering both simple and advanced MLIA functionalities to TEL users.

In the following sub-sections, we make some more detailed observations on this approach.

6.1.1 Architectural Aspects

This "isolated query translation" can be implemented complying with the "low barrier" design approach characteristic of TEL, since the national digital libraries participating in TEL do not need to be aware of this solution in order to benefit from it. In addition, it has been demonstrated how this solution can be integrated into the TEL architecture in a transparent way. As a consequence, both national libraries, which have already joined TEL, and national libraries, which will join TEL in the future, are not required to perform any modifications to their systems. Therefore, this solution not only has very low impact on the TEL partners, since it neither requires changes by existing partners nor raises the barriers for new partners to join TEL, but also it represents a great benefit for the TEL user community, by offering multilingual query support.

Many different alternatives are available to readily put this approach into practice, ranging from adopting off-the-shelf components and services (both commercial and free) to developing components tailored to the specific needs of TEL. As a consequence, a twofold strategy can be adopted: off-the-shelf components may be used in a first step to have a quick feedback from the user community, while a solution, which is more advanced and customized for the TEL needs, may be developed in a second step. Please note that in this second step a higher interaction with the "pseudo-translation" solution can be exploited, because the system can be designed and developed to share components and resources between the "isolated query translation" and the "pseudo-translation" solutions.

6.1.2 User-Interaction Aspects

Two different alternatives for extending the TEL user interface in order to offer both simple and advanced search functionalities have been studied.
It has been demonstrated how the TEL user interface can be extended in a simple and coherent way to make these new features accessible to end-users and how the style of interaction with the user adopted so far in TEL can be maintained and successfully adopted also in the case of multilingual queries.

On the other hand, particular care should be paid in designing and developing the user interaction, because TEL users can range from people who may not be familiar with foreign languages and so may have troubles in understanding whether a translation is correct, to people with high multilingual capabilities, able to understand the translations of a query and to fine-tune them.

Since both kinds of users have to be supported, the user interface and interaction have to be designed with care, so that basic functionalities are easily accessible for non-expert users while, for expert user, it should be possible to quickly evaluate the translation alternatives and choose among them.

6.1.3 Translation and Retrieval Effectiveness Aspects

In the “isolated query translation” solution the term “isolated” highlights the fact that the translation problem is separated from actual retrieval .

In particular, as explained above, for the translation problem an interactive approach has been adopted, so that the user can iteratively see and refine different translation proposals. Once the user considers the translation correct, the translated query is sent to national library systems for searching. Thus, retrieval effectiveness mainly depends on the performances of the national library systems, since at this point the TEL system is no longer involved.

On the other hand, evaluating the quality of the translation is more a natural language processing problem than an information retrieval one and the results may vary depending on the resources adopted. A study of this kind is out-of-scope for the purposes of the present work. Nevertheless, the considerations and findings about the issues encountered with the PROMT system can provide us with some insights of what may happen with the "isolated query translation". Indeed, using a MT system for translating the user queries will have similar problems to those ones encountered with PROMT, and described in Section 5.6.

It should be noted that we deal with an iterative and interactive translation process. This is in contrast to the “pseudo-translation” approach, where we deal with a kind of “one-shot” translation, since records are translated only once. As a consequence, the translation problems highlighted in the “pseudo-translation” solution should have a lower impact in the case of the “isolated query translation” solution, because the user has the possibility to see and modify the translated query before sending it to the national libraries.

6.2 Considerations on the “Pseudo-Translation Approach”

In the “pseudo-translation” approach, we moved a step further and proposed a solution able to effectively exploit the records held in the TEL central index, which is going to increase in the future, in order to provide more fine-tuned MLIA facilities to TEL users. Our tests indicate a good effectiveness for cross-language retrieval using this method.

In the following sub-sections, we make some more detailed observations on this approach.

6.2.1 Overall effectiveness of CLIR using Pseudo-translation

We were able to significantly expand on the initial goal of providing some sample records for the pseudo-translation approach and have crafted a test collection of 151,700 records and 100 queries in order to evaluate CLIR effectiveness of our pseudo-translation approach. For reasons of practicability, the experiments were evaluated using a monolingual baseline as gold standard and limited to judgment of the top-ranked records.

The necessary limitations in the evaluation mean that it is difficult to make comparisons with the CLIR evaluation campaigns such as the Cross-Language Evaluation Forum (CLEF).

Nevertheless, the evaluation provides solid evidence for assessing retrieval effectiveness, much better than initially planned for the feasibility study.

We have found evidence that at least 55% of queries should be considered of good retrieval quality after cross-language retrieval, and more than 80% can be assumed to be of roughly equal performance to the monolingual "gold standard" (for interpretation of these numbers, please note that there do not necessarily exist relevant records for all the queries). This is encouraging, and compares fairly well to state-of-the-art CLIR on lengthy documents [6]. Pseudo-translation was found to work well.

It should also be noted that these procedures tend to benefit from larger data collections. We found clear evidence of this when setting up our test collection. Initial trials on 10,000 records were not encouraging, and it quickly emerged that the collection would have to be scaled up to get meaningful results. The same approach could thus perform better when applied to the whole TEL databases.

6.2.2 Problems in procedure

A number of problems have been identified in the procedures used, some of which could be fixed fairly easily in subsequent setups:

- the lack of stemming and decompounding (for German) was likely detrimental for overall retrieval quality. A future system should include both features
- the PROMT machine translation system provided good vocabulary coverage, but was not able to give multiple translation alternatives for terms. A machine-readable dictionary, if obtainable with similar coverage, may be better suited
- small technical problems with Terrier. The method described is in no special way bound to the Terrier system, and should integrate well with the present TEL system, provided probabilistic, ranked retrieval can be added to TEL.

6.2.3 Expansion on external pilot collection

Notwithstanding the encouraging overall result, there remain issues with expansion. We are convinced that expansion should be helpful for retrieval quality. Unfortunately, the effect was hard to measure on the query set, and resources prohibited a direct retrieval comparison between expanded and unexpanded records. We believe that expansion is yielding too few extra terms as it is implemented now. There is too little content in the individual records in order to effectively identify similar records. While the number of terms should increase as a result of scaling up to larger sets of records, we propose extending the analysis also using an external "pilot collection" [18]. By using an appropriate set of full-text documents for expansion, short bibliographical records could be heavily expanded (by 100% and more). This is consistent with the plans of TEL to expand their system with more electronic full-text documents in the future. The procedures described in this report should readily adapt to such a setting.

6.3 Final Remarks

It is important to note that both the proposed 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.

A combination of the two approaches would lend itself naturally to a system using an interlingua setup. In such cases, instead of having support and multilingual resources for all the possible pairs of source and target languages, one language is selected as *pivot* and all the translations are made to and from this pivot language. For example, if we need to translate from Portuguese to Bulgarian, instead of performing a direct translation, we may choose English as the interlingua, and perform a translation from Portuguese to English and from English to Bulgarian. This solution would allow us to avoid an exponential growth of the multilingual resources, if many different partners join TEL. In a combination approach, one of the two translation steps would be implemented using "pseudo-translation", while the other translation step is carried out during retrieval using "isolated query translation". This would address issues with translation errors ("noise") that are typically amplified when multiple translation steps are used. The expansion and the user interaction, respectively, can be expected to mitigate such noise problems.

On the whole, we can envision the following evolutionary scenario:

- short-term: the "isolated query translation" solution is a first step for adding MLIA functionalities to TEL and represents a quick way to give TEL users and partners a multilingual experience. In parallel, steps should be undertaken to take more advantage of the growing central index in TEL and prepare for additional CLIR components such as "pseudo-translation"
- mid-term: the implementation and deployment of a "pseudo-translation" solution is a second step which allows to better exploit the information directly managed by the TEL central index;
- long-term: the adoption of an inter-lingua approach allows for scaling up the system, provides the means for better integrating the "isolated query translation" and the "pseudo-translation" solution and facilitates sharing components and resources among them. The implementations can be specifically tailored to address the needs of TEL.

Finally, the suggested evolutionary scenario is in line with the research directives for MLIA outlined in the i2010 Digital Library Initiative, since they aim at the development of multilingual access functionality for the digital resources of Europe's cultural institutions.

7 Acknowledgments

We thank Marco Dussin for his work on the study of the modifications to the TEL system user interface for introducing the "isolated query translation" feature.
Many thanks are alsodue to Bill Oldroyd of the British Library for his assistance in obtaining the set of records used for the experiments on pseudo-translated, expanded records. Similarly, we would like to thank Eric van der Meulen of the TEL office for providing us with the log file data on which we based the set of test queries for these experiments.

8 References

[1] Agosti, M., Braschler, M. & Ferro, N. (2006). A Study on how to Enhance TEL with Multilingual Information Access. In C. Thanos, editor, DELOS Research Activities 2006, pp. 115-116, ISTI-CNR At Gruppo ALI, Pisa, Italy.

[2] Agosti, M., Di Nunzio, G. M., and Ferro, N. (2006). A Data Curation Approach to Support In-depth Evaluation Studies. In Gey, F. C., Kando, N., Peters, C., and Lin, C.-Y., editors, Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006), pages 65–68. http://ucdata.berkeley.edu/sigir2006-mlia.htm [last visited October 2006].

[3] Agosti, M., Di Nunzio, G. M., and Ferro, N. (2006). Scientific Data of an Evaluation Campaign: Do We Properly Deal With Them? In Nardi, A., Peters, C., and Vicedo, J. L., editors, Working Notes for the CLEF 2006 Workshop. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/agostiCLEF2006.pdf [last visited October 2006].

[4] Ballesteros, L. & Croft, W. B. (1997). Phrasal translation and query expansion techniques for cross-language information retrieval. In N. J. Belkin, D. Narasimhalu, P. Willett (Eds.), Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 84-91).

[5] Braschler, M. (2004). Combination Approaches for Multilingual Text Retrieval. In Information Retrieval, Volume 7, Issue 1/2, 183-204, Kluwer Academic Publishers.

[6] Braschler, M. (2004) Robust Multilingual Information Retrieval. Dissertation. Institut Interfacultaire d'Informatique, Université de Neuchâtel.

[7] Braschler, M., Di Nunzio, G. M., Ferro, N., and Peters, C. (2005). CLEF 2004: Ad Hoc Track Overview and Results Analysis. In Peters, C., Clough, P., Gonzalo, J., Jones, G. J. F., Kluck, M., and Magnini, B., editors, Multilingual Information Access for Text, Speech and Images: Fifth Workshop of the Cross–Language Evaluation Forum (CLEF 2004) Revised Selected Papers, pages 10–26. Lecture Notes in Computer Science (LNCS) 3491, Springer, Heidelberg, Germany.

[8] Braschler, M., Krause, J., Peters, C. & Schäuble, P. (1999). Cross-Language Information Retrieval (CLIR) Track Overview. In E. M. Voorhees and D. K. Harman (Eds.), Information Technology: The Seventh Text REtrieval Conference (TREC-7), NIST Special Publication 500-242 (pp. 1-8).

[9] Braschler, M., Ferro, N. & Verleyen, J. (2006). Implementing MLIA in an existing DL system. In F. C. Gey, N. Kando, C. Peters, and C.-Y. Lin, editors, Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006), pp- 73-76, http://ucdata.berkeley.edu/sigir2006-mlia.htm [last visited October 2006].

[10] Braschler, M. & Peters, C. (2004). Cross-Language Evaluation Forum: Objectives, Results, Achievements. In Information Retrieval, Volume 7, Issue 1/2, 7-31, Kluwer Academic Publishers.

[11] Braschler, M. & Ripplinger, B. (2004). How Effective is Stemming and Decompounding for German Text Retrieval?. In Information Retrieval, Volume 7, Issue 3/4, 291-306, Kluwer Academic Publishers.

[12] Cleverdon, C. W. (1967). The Cranfield tests on index language devices. Aslib Proceedings, 19, 173-192. Reprinted in (Sparck Jones and Willett, 1997)

[13] Di Nunzio, G. M., Ferro, N., Jones, G. J. F., and Peters, C. (2006). CLEF 2005: Ad Hoc Track Overview. In Peters, C., Gey, F. C., Gonzalo, J., Jones, G. J. F., Kluck, M., Magnini, B., Müller, H., and de Rijke, M., editors, Accessing Multilingual Information

Repositories: Sixth Workshop of the Cross–Language Evaluation Forum (CLEF 2005). Revised Selected Papers, pages 11–36. Lecture Notes in Computer Science (LNCS) 4022, Springer, Heidelberg, Germany.

[14] Di Nunzio, G. M., Ferro, N., Mandl, T., and Peters, C. (2006). CLEF 2006: Ad Hoc Track Overview. In Nardi, A., Peters, C., and Vicedo, J. L., editors, Working Notes for the CLEF 2006 Workshop. http://www.clef-campaign.org/2006/working_notes/workingnotes2006/dinunzioOCLEF2006.pdf [last visited October 206].

[15] European Commission (2006). Commission Recommendation of 24 August 2006 on the digitisation and online accessibility of cultural material and digital preservation. Official Journal of the European Union, OJ L 236, 31.8.2006, 49:28–30, 31 August 2006, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2006:236:0028:0030:EN:PDF [last visited October 2006].

[16] European Commission Information Society and Media (2006). i2010: Digital Libraries. http://europa.eu.int/information_society/activities/digital_libraries/doc/brochures/dl_brochure_2006.pdf [last visited October 2006].

[17] Frakes, W. B. & Baeza-Yates, R. (1992). Information Retrieval. Data Structures & Algorithms. Prentice-Hall.

[18] Lam-Adesina, A. & Jones, G. J. F. (2003). Exeter at CLEF 2002: Experiments with Machine Translation for Monolingual and Bilingual Retrieval. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Advances in Cross-Language Information Retrieval, Third Workshop of the Cross-Language Evaluation Forum, CLEF 2002, Lecture Notes in Computer Science, Vol. 2785 (pp. 127-146), Spinger Verlag.

[19] Levow, G. A., Oard, D. W. & Resnik, P. (2005). Dictionary-based techniques for cross-language information retrieval. Information Processing & Management, 41(1):523–547.

[20] McNamee, P. & Mayfield, J. (2002). JHU/APL Experiments at CLEF: Translation Resources and Score Normalization. In C. Peters, M. Braschler, J. Gonzalo, M. Kluck (Eds.), Evaluation of Cross-Language Information Retrieval Systems, Second Workshop of the Cross-Language Evaluation Forum, CLEF 2001, Lecture Notes in Computer Science, Vol. 2406 (pp. 193-208), Springer.

[21] OAI (2004). The Open Archives Initiative Protocol for Metadata Harvesting – Version 2.0. http://www.openarchives.org/OAI/openarchivesprotocol.html [last visited October 2006].

[22] Oard, D. W. (1997). Alternative approaches for cross-language text retrieval, In AAAI Symposium on Cross-Language Text and Speech Retrieval. American Association for rtificial Intelligence. Electronic working notes at (AAAI, 1997).

[23] Oard, D. W. & Resnik, P. (1999). Support for interactive document selection in cross-language information retrieval. Information Processing & Management, 35(1):363–379.

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

[25] Peters, C. (2006). Multilingual Information Access for Digital Libraries: The Impact of Evaluation on System Development. In C. Thanos, editor, DELOS Research Activities 2006, pp. 105-107, ISTI-CNR At Gruppo ALI, Pisa, Italy.

[26] Peters, C. & Braschler, M. (2001) European Research Letter: Cross-language system evaluation: The CLEF campaigns, Journal of the American Society for Information Science and Technology, Volume 52, Issue 12, 1067-1072, John Wiley & Sons.

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

[28] Xu, J. & Croft, W. B. (1996). Query expansion using local and global document analysis. Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, Switzerland, pages: 4 – 11, 1996