**Grant Agreement Number: 257528**

**KHRESMOI**

**www.khresmoi.eu**

# Report on and prototype of the translation support

| | |
|---|---|
| **Deliverable number** | *D3.1* |
| **Dissemination level** | *Public* |
| **Delivery date** | *10 May 2012* |
| **Status** | *Final* |
| **Author(s)** | *Lorraine Goeuriot*<br>*Gareth Jones*<br>*Liadh Kelly*<br>*Sascha Kriewel*<br>*Pavel Pecina* |


*This project is supported by the European Commission under the Information and Communication Technologies (ICT) Theme of the 7th Framework Programme for Research and Technological Development.*

## Abstract

We describe in this deliverable the introduction of translation support within the Khresmoi system. This support was added using ezDL technologies developed by the University of Duisburg Essen, which provides a library manager with a graphical interface, and a translation service provided by Charles University in Prague, which translates from Czech, French, German into English and vice-versa. The goal of this translation support is to allow users to extend their search to the English library without the need for English knowledge. It is important for most of the users and more specifically patients to get medical information in their mother tongue or in a language they can easily understand. But relevant information may not be provided in their language (especially non-English). Thus our system will let users query English libraries without using English: the system will translate their query to English, search for documents in the English library and send them back translated into the user's language.


---

## Table of Contents

# List of Abbreviations

| | |
|---|---|
| API | Application Programming Interface |
| CUNI | Charles University in Prague |
| CZ | Czech language |
| EU | European Union |
| FR | French language |
| GE | German language |
| JSON | JavaScript Object Notation |
| MT | Machine Translation |
| REST | Representational state transfer |
| UDE | University of Duisburg-Essen |

# 1 Introduction

Obtaining relevant and valuable medical information is one of the main concerns of Khresmoi potential users (see [1], [2] and [3]). Moreover, it is very important for users, and especially patients, to get information in their own language. English is the lingua franca of scientific domains and still a dominant language on the web. This constitutes a major problem for non-English speaking users in obtaining information they are looking for.

The Khresmoi system aims at providing EU citizens (general public, medical practitioners and radiologists) valuable medical information. To cope with the unbalanced aspect of available multilingual medical data, translation support is included in the Khresmoi search system to provide access to the medical information available only in English. This will provide support to users with different levels of English, ranging from no English knowledge upwards.

When the user types a query in a language other than English, the system will offer to translate the query into English. If the user selects the query translation option, the English translation of the query will be searched in the English database. Retrieved results will then be translated into the user's original language prior to being returned to the user.

The Khresmoi search system is based on ezDL, developed by UDE, a digital libraries manager and search system, with an adaptable graphical interface. The translation support was integrated into ezDL, using a translation service API, provided by CUNI. This translation service API handles 4 languages and provides the following translations: {CZ, FR, GE} → EN and EN → {CZ, FR, GE}.

The ezDL interface is divided into customisable views, such as query, results, query history, personal library, etc. The goal of the task described in this deliverable was to extend some of these views and link them to the translation service. The views which were extended in order to provide users translation support are: the query view, where the user types the query; the result view, where the list of results is displayed; and the detail view, providing summaries of the document selected in the result list. More specifically, once a query is typed, it is sent to the translation API and possible translations are sent back to the interface. If one of these is selected, the translation support is launched. The translated query is then searched in the English library, and the results are translated back by the API into the user's language.

Section 2 describes the current Khresmoi system, and the main components we are focusing on here, ezDL and the translation service. In Section 3 we give an overview of the translation support system and detail each of its main components: query translation support, and results translation support.

# 2 Current System Description

The Khresmoi project aims to develop a multilingual and multi-modal search and access system for biomedical information and documents. Khresmoi targets three kinds of users. Two of these are groups with general medical interest: general public and general medical practitioners. The other group is an example of clinician with a specific expertise, radiologists. The Khresmoi system can be seen as a set of services allowing any of these users to access medical information. The architecture of the system is composed of three layers (see Figure 1): the application layer, the interface between users and the services; the core service layer, responsible for the services orchestration; and the persistence layer, composed of services and repositories.

The translation support involves 3 elements of this global architecture:

- the translation service,
- the data repositories
- and the end user interface.

Figure 2 shows the sequence diagram of textual search in the system, from the user to Mimir[1], the search and indexing tool. The translation of the query is performed prior to sending the query to the backend and the results are translated after they are sent back to ezDL. Within the query processing flow, translation is performed after spell checking and before disambiguation of the query.

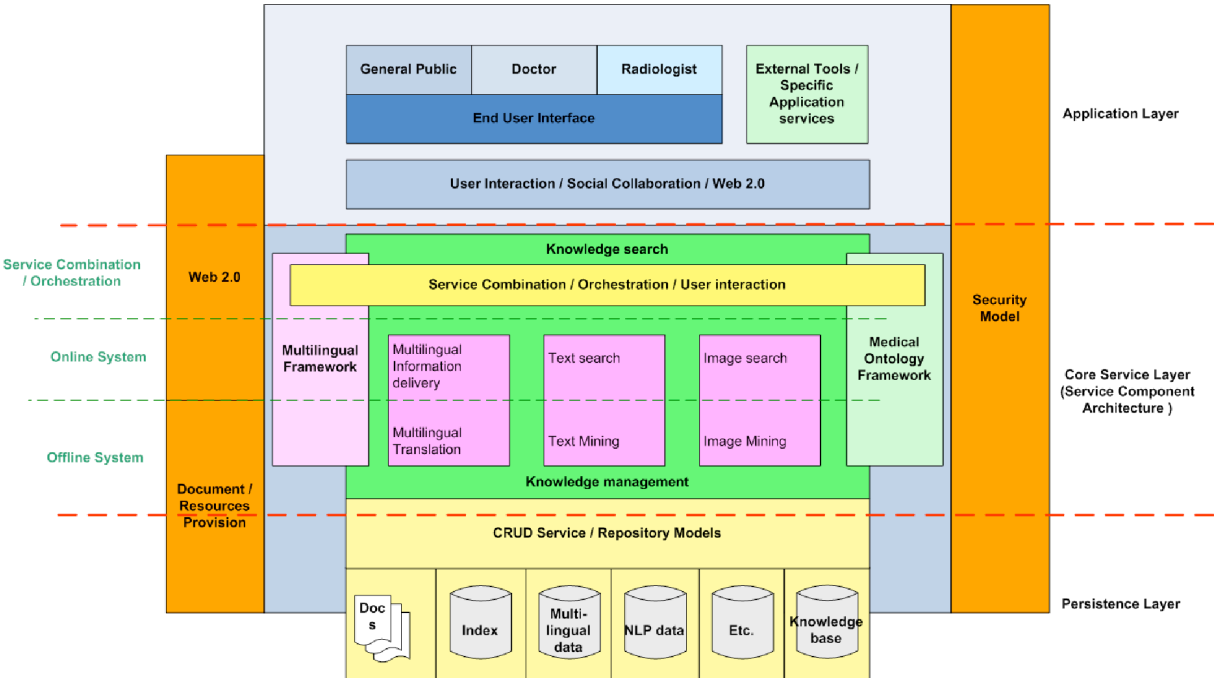


**Figure 1: Khresmoi system architecture**

The remainder of this section describes the two components which were involved in the development of the translation support: the ezDL search application and the translation service API.

## 2.1 EzDL

The user interface of the Khresmoi system is based on ezDL[2], the successor of the Daffodil software [4] developed at the University of Duisburg-Essen. EzDL is a multi-agent search system for heterogeneous data sources and a tool-set for building search user interfaces to support complex tasks. It allows for simultaneous searches in multiple digital libraries through a unified interface and query syntax, and presents a merged and enriched view of the results. The tools provided by ezDL allow users to work with the results and can be arranged in customisable perspectives.

EzDL is composed of a server part consisting of a directory and a large number of agents, and clients that contain a selection of loosely-coupled tools which serve as a user interface to the system (see Figure 3).

[1] https://gate.ac.uk/mimir/

[2] http://www.ezdl.de

The server-side agents connect to the search and query support services provided within Khresmoi, handle user authorisation, user profile management, logging, storage of user data and queries, and the caching of documents. Two basic clients are available within Khresmoi: a search desktop written in Java (see Figure 4), as well as a browser application that uses Java Server Faces. Users can either search as guests or obtain a personal account. A personal account allows for a persistent search history spanning multiple search sessions and offers access to a document depository called `personal library', where a user can store found and uploaded documents, as well as favourite queries and authors, and categorise them with personal tags [5].

Khresmoi is mainly dealing with three libraries: one containing text, one containing 2D images and one containing 3D images. The text library is composed of documents in Czech, English, French, German and Spanish.

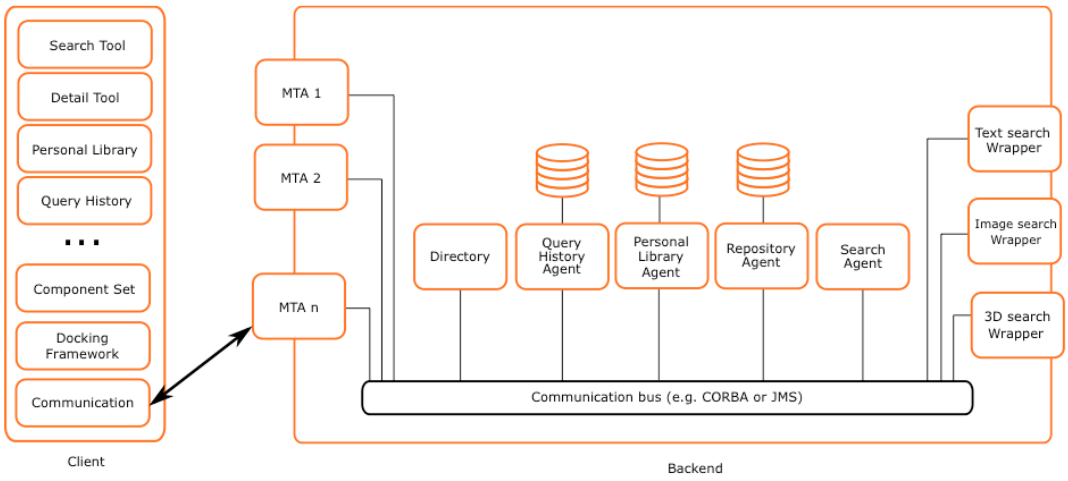


**Figure 2: ezDL architecture**

The search desktop interface used in Khresmoi is composed of customisable views. The views for which translation support is provided are (see Figure 6):

(1) Query: view of the query (text field) and the query functionalities (translation, spell checking, suggestions, definitions),

(2) Results: condensed view of the result list, displaying meta information such as the title, the source, the author(s), snippets, etc.

(3) Details: Detailed view of a selected result. All the meta information is displayed (title, year, author, keywords, publisher, etc.) as well as snippets. Task 4.4 aims at creating summarised views of the document to replace the snippets here.

The process is the following: The user types a query in (1), then clicks the “search” button; the list of results appears in (2), where the user can select relevant ones; a summarized view of the user selected result is displayed in (3).

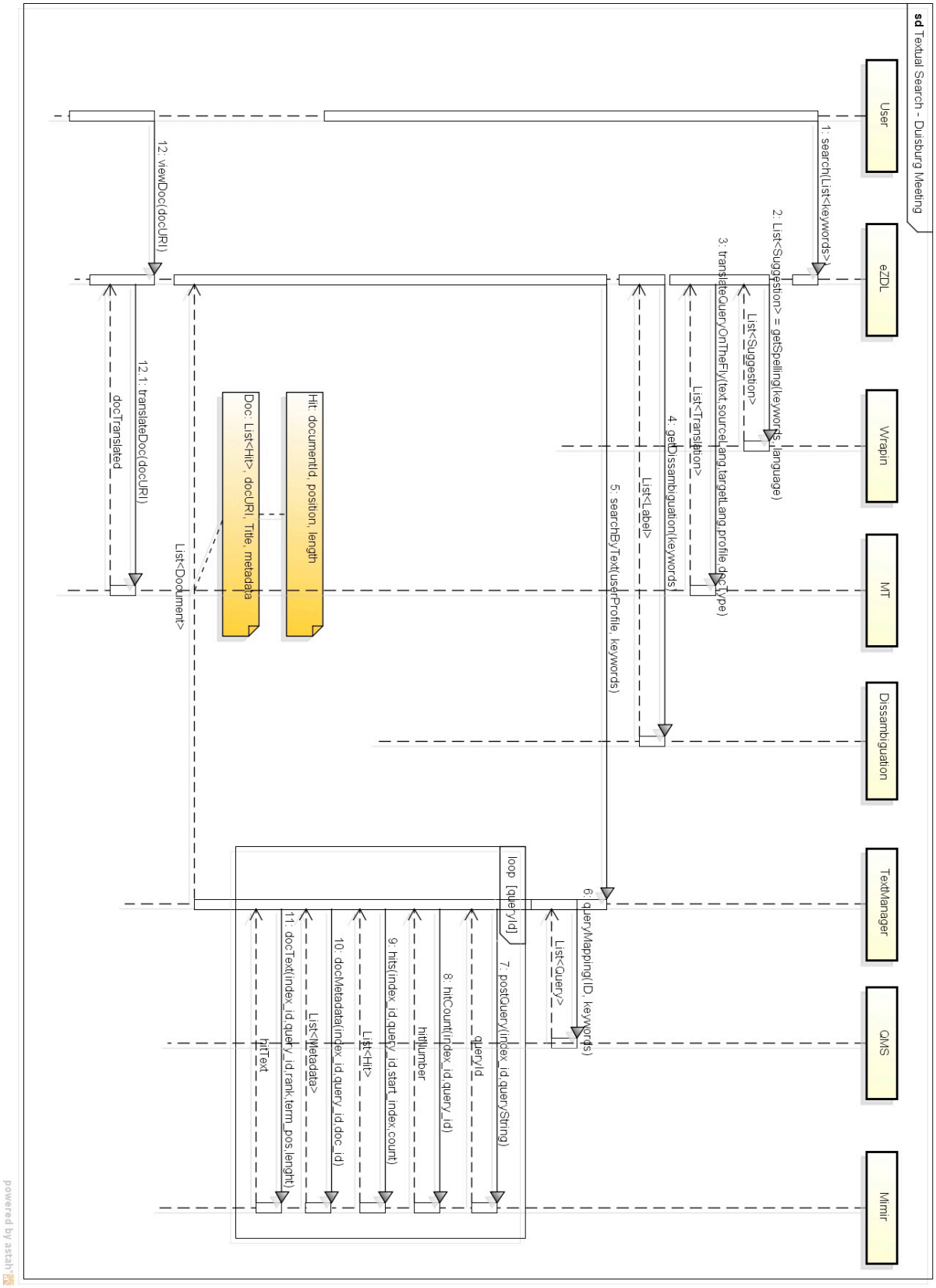


**Figure 3: Textual workflow sequence diagram**

## 2.2 Translation Service

The translation service API is developed by Charles University in Prague. It involves 4 languages: Czech, English, French and German. The service is based on Moses[3]; a statistical machine translation system. Three translators have been trained, for the language pairs CZ-EN, FR-EN and GE-EN. Each of these translators are connected to Khresmoi cloud through a REST API and communicate using JSON requests (see Figure 5).

These translators are phrase based and can handle multiple kind of entries: queries, summaries and whole documents, that are all considered as text.

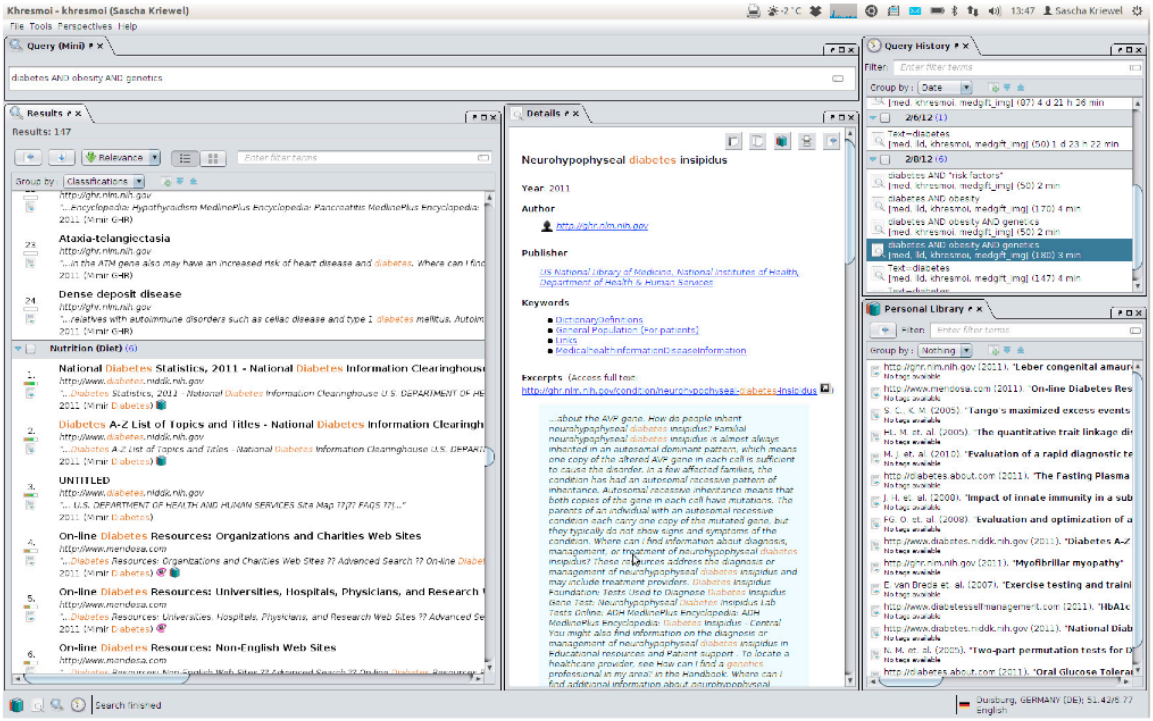


**Figure 4: ezDL snapshot**

A text pre-processing step is required by Moses system and provided by the translation system:

```
Input/output: plain text UTF-8 (with simple tags?)
1.Tokenisation
Identification of tokens and not-to-be translated text
2.Lowercasing
Mapping to lower case to reduce data sparseness
3.Translation
4.Recasing
Reconstruction of letter cases
```

3 http://www.statmt.org/moses

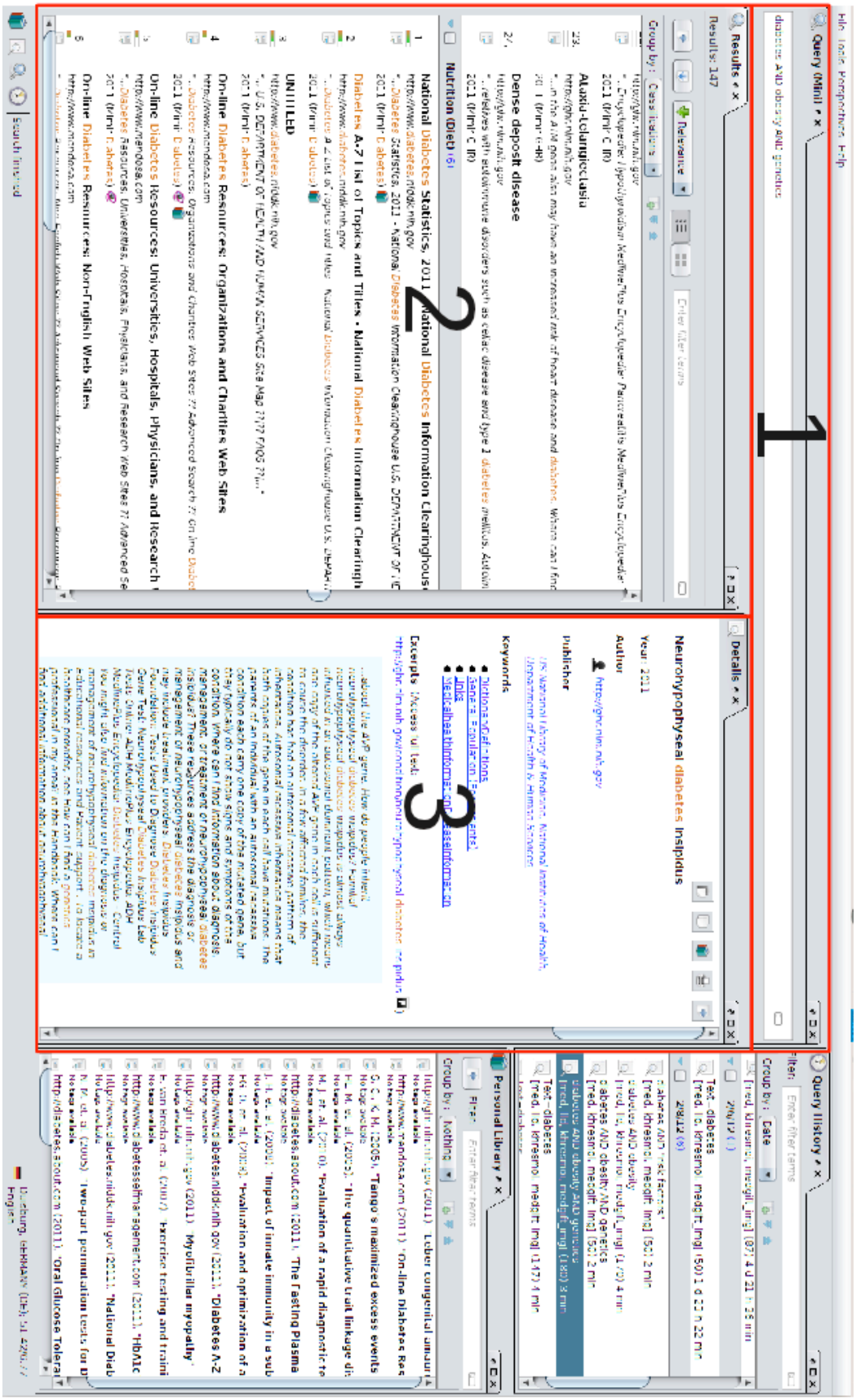

**Figure 5: ezDL views**

5.Detokenisation

Removing spaces before punctuation etc.

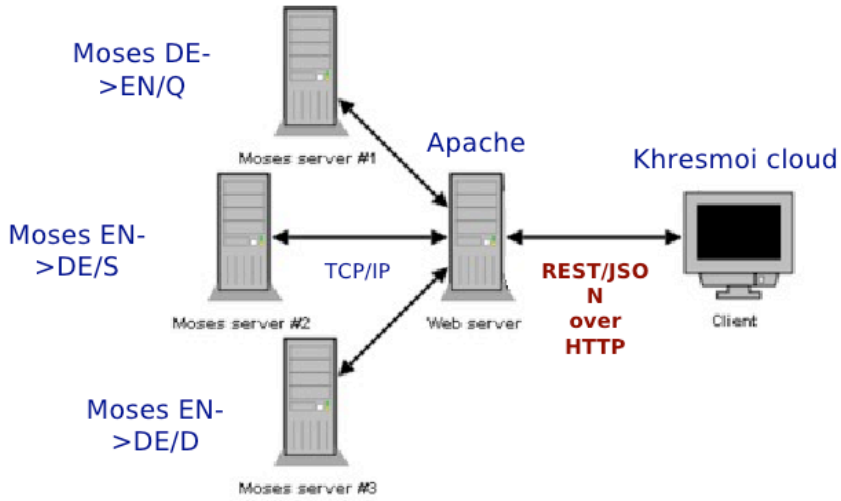


**Figure 6: Translation service architecture**

System inputs and outputs are JSON queries, for example:

```
Input: {
"sourceLang": "en",
"targetLang": "de",
"docType": "query",
"profile": "general public",
"text": "Treat Flu",
"nbest": 1,
"align-info": false,
…
}
```

```
Output: {
"translation": [
{
"translated": "Gönnen Grippe",
"score": 12345,
"translationId":
"adeb8b91-c27f-4e95-a36c-
62a7060e123b",
}]
}
```

# 3 Translation Support System Overview

The purpose of Task 3.3 is to provide translation support by linking ezDL with the translation service API. Because medical data and more generally scientific data is more numerous in English than in other languages, the goal of our support system is to provide users a way to query English databases even if they do not understand English. The system provides all users the option to translate their non-English query once it has been typed. If this option is chosen by the user, it will be sent to the English database and results will be translated into the user's language before they are sent back to the user (As quality of machine translation can be far from human translation, the user will be able to access the original English text). Users are however free not to use this translation option, or query themselves in English.

The process involved in a search using the Kreshmoi system with the integrated translation support is the following (see Figure 6):

(1) In the query view: when the user types a query, a list of possible translations appears. If the user selects one, it replaces the current query.

(2) In the result list view: the results of submitted query are presented, translated into the user's language (of the original query).

(3) In the detail view: when a result is selected from view (2), a more detailed summarized view of it is displayed. By default, the translated version is displayed but the user can also view the original English version.

Figure 8 provides a high-level architectural overview of the translation support operation. The frontend components consist of the three previously mentioned views. The backend components manage messages from the frontend and act as an intermediary between the user (the frontend) and the data or services. When a query is typed, the frontend sends it to the query pre-processing agent that, among other processes, translate it by calling the translation service and return the n-best translations to the frontend. If a translation is selected by the user, it is then sent again to the backend to the search agent, which obtains matching documents from the repositories. The result processing agent then receives the results, translates them (calling the translation service) and sends them back to the frontend.

The remainder of this section describes each of the view functionalities (query, result list and detailed view) in detail.

## 3.1 Query Translation Support

The query view is the start of every search on the system. It is composed of a text field, and a search button. When the user starts typing a query, a drop-down menu shows options such as: spelling correction, and query suggestion. The query suggestion uses medical terms from the knowledge base. Their definition, if available in the knowledge base, is displayed in a pop-up window when the mouse flies over the entry. A list of English translations were added to the drop-down menu, with a message indicating that the English translation may give better results. In case the user has any knowledge in English and to potentially overcome errors from machine translation, the n-best translations are displayed (ranked according to the MT scores). If the user selects an English translation, it replaces the typed query.

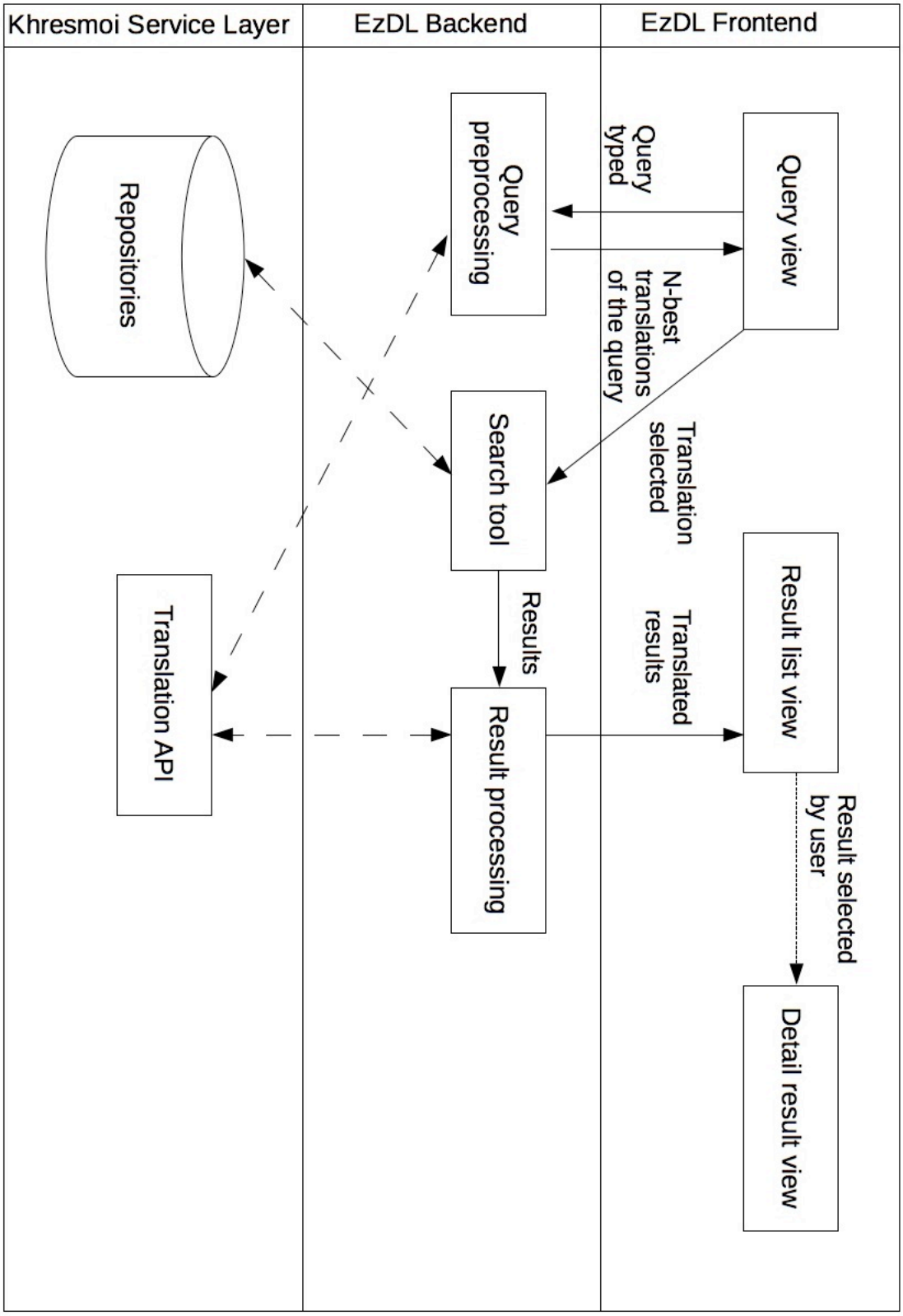


**Figure 7: Translation support architecture**

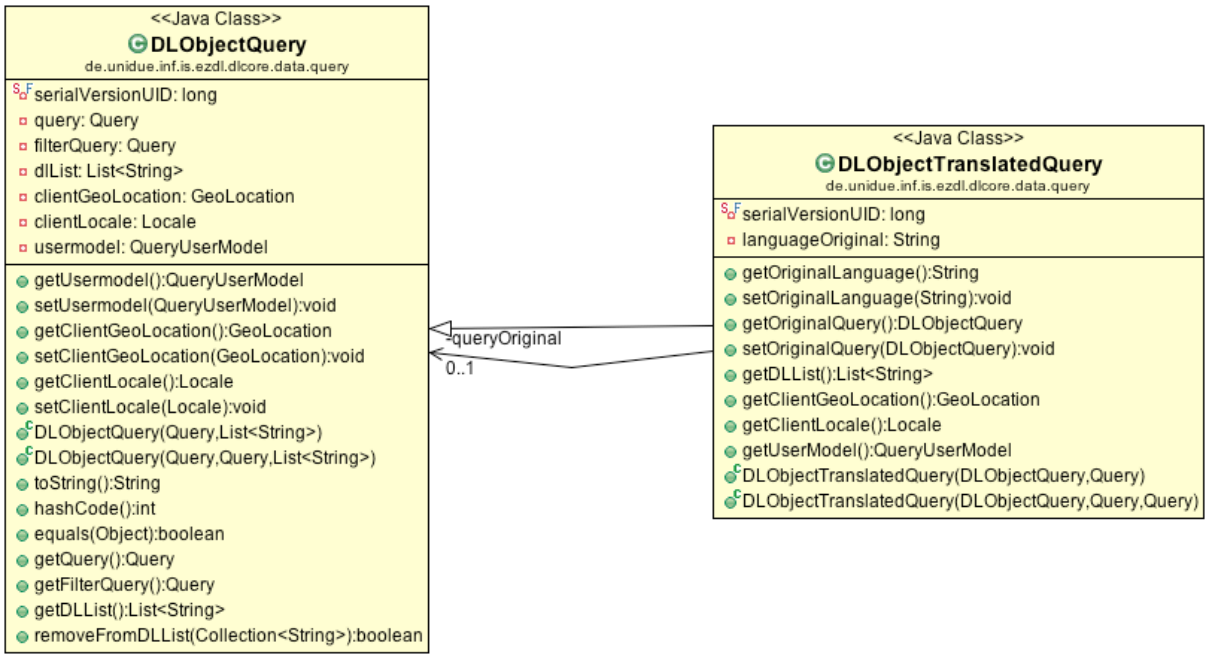


**Figure 8: Query translation class diagram**

To manage translation within ezDL, the existing class `DLObjectQuery` was extended to store both the original text (typed by the user) and the English translated version (see Figure 9).

## 3.2 Result Translation Support

When the query has been sent by the user to the system, it is handled by ezDL backend which sends it to the wrappers, linked to the libraries (through indexes). One wrapper is related to text search, which is linked to the libraries (the data) via indexes (WP1 T1.4). This query handling is performed by the class `DLObjectQueryHandler` in ezDL, which receives the query from the backend, sends it to the wrapper(s), collects the results and sends them back to the frontend. A translated query is handled in the same way as any other query (only the language changes). When the query is a translated one, the only additional operation that has to be performed is the translation of the results before they are sent back to the frontend:

1. Receipt of the query from the frontend
2. Query sent to the wrapper(s)
3. Collection of the results
4. Translation of the results
5. Sending of the results to the frontend

Results are stored in a list containing generic objects that can store different types of documents. Here we are only working with textual documents, which are stored using the class `MimirDocument`. As was done for queries, the document class is extended to handle the translation (see Figure 10).

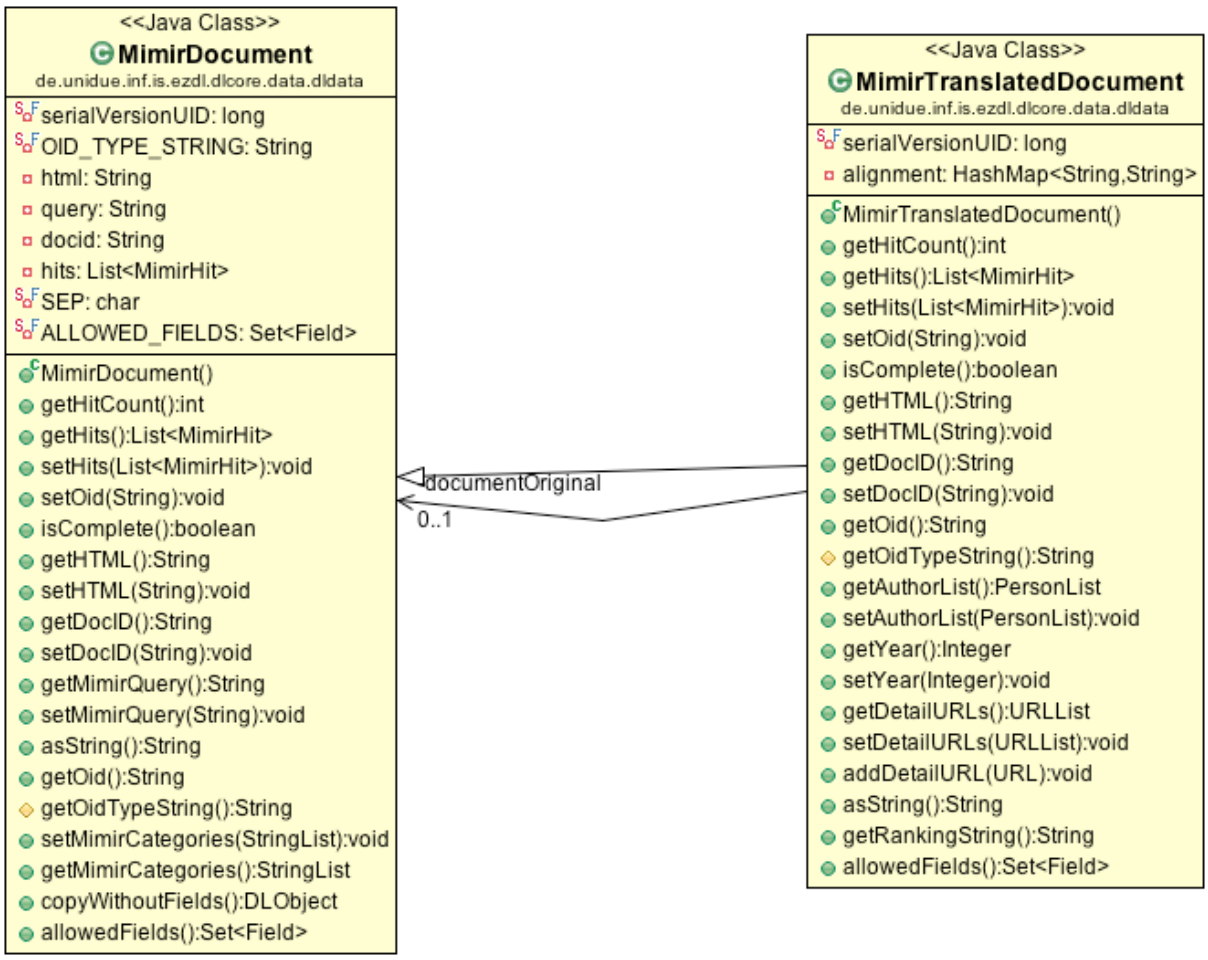


**Figure 9: Document translation class diagram**

After the results are collected from the wrapper(s), the results are translated, using the translation service: varied fields have to be translated in the document, they are stored as new values in the object, while original values are stored in the field `DocumentOriginal` of the object. As explained in Section 2.2, a text and the original and target languages are the only elements required by the translation service. Thus we use a method calling this service for each part of the documents that need to be translated. These translated results are sent back to the frontend as a list `ResultDLObjectList`, containing either `MimirDocument` or `MimirTranslatedDocument` (as an extension any type of document).

The frontend manages the new `MimirTranslatedDocument` as if they were any kind of document in both the result list view and detailed view.

# 4 Conclusion

This deliverable presented the translation support included in the Khresmoi system. The need for this support is based on two facts: firstly, there are more scientific documents written in English than in any other language; secondly, users are usually more comfortable viewing information in their own language. Thus, our system allows users to extend their search to the English libraries without the need for English knowledge. To provide translation support on the interface, we extended ezDL technologies, developed by UDE, with the translation services provided by CUNI. The system currently handles Czech, English, French and German, and will soon include Spanish. The system

works as follows: when a non-English query has been typed, a list of possible English translations are suggested; if the user selects one of the translations, the English libraries are queried and the results are translated back to the starting language before being presented to the user. The translation component will be evaluated in year 3 and also during the forthcoming user-centered prototype evaluations.

# 5 References

[1] Natalia Pletneva and Alejandro Vargas, D8.1.1. Requirements for the general public health search. Khresmoi Project public deliverable, May 2011.

[2] Manfred Gschwandtner, Marlene Kritz and Celia Boyer, D8.1.2: Requirements of the health professional search. Khresmoi Project public deliverable, August 2011.

[3] Henning Müller, D9.1: Report on image use behaviour and requirements. Khresmoi Project public deliverable, May 2011.

[4] Norbert Fuhr, Claus-Peter Klas, André Schaefer, and Peter Mutschke. 2002. Daffodil: An integrated desktop for supporting high-level search activities in federated digital libraries. In Research and Advanced Technology or Digital Libraries. 6th European Conference, ECDL 2002, pages 597–612, Heidelberg. Springer.

[5] Lorraine Goeuriot, Allan Hanbury, Gareth J. F. Jones, Liadh Kelly, Sascha Kriewel, Ivan Martinez Rodriguez, Henning Müller, Miguel A. Tinte (2012) - Supporting Collaborative Improvement of Resources in the Khresmoi Health Information System. In Proceedings of Collaborative Resource Development and Delivery (accepted).