CLIR-Based Collaborative Construction of a Multilingual Terminological Dictionary for Cultural Resources

Mohammad DAOUD, Mohammad.Daoud@imag.fr [1]
Asanobu KITAMOTO, Kitamoto@nii.ac.jp [2]
Christian BOITET, Christian.Boitet@imag.fr
Mathieu MANGEOT, Mathieu.Mangeot@imag.fr

Keywords: multilingual term database, dictionary initialization, community-based translation, cross-lingual information retrieval, automatic terminology translation.

Abstract

We will describe ongoing work in developing a collaborative environment to construct a CLIR-based multilingual *terminological dictionary* dedicated to the Digital Silk Road project and web site launched and managed by NII (National institute of Informatics, Japan-Tokyo). A considerable amount of cultural resources has been digitized, including 95 rare books written in 10 different languages. In order to make them searchable and accessible easily by the visitors of the site, themselves multilingual as well, a cross lingual information retrieval system is being built. As these books are very rich in specialized terms, an important part of that endeavour is to gather these terms in many languages in a *terminologicial dictionary* (a database of terms contianing some information potentially usable to later build a real terminological database). For that purpose, we use a participative approach, where visitors of the online archive are the main source of the terms used in the languages they know, while multilingual online resources are used to initialize the term base through a process that depends on the archived textual data.

---

[1] The first, the third, and the fourth authors work at Grenoble Informatics Laboratory, GETALP, Université Joseph Fourier (Grenoble, France).

[2] The second author works for the National Institute of Informatics (Tokyo, Japan).

1 Introduction

The Digital Silk Road project (ONO, KITAMOTO et al. 2008) is an initiative started by the National Institute of Informatics (Tokyo) in 2002, to archive cultural historical resources along the Silk Road, by digitizing them and making them available and accessible online. One of the most important sub-projects is the Digital Archive of Toyo Bunko Rare Books (NII 2008) where tens of old rare books available at Toyo Bunko library have been digitized using OCR (Optical Character Recognition) technology. The digitized collection contains books from different languages (English, French, Russian…), all of them related to the historical Silk Road, like the 2 volumes of the Ancient Khotan by Marc Aurel Stein, and the "*Mission Scientifique dans la Haute Asie*" by Jules-Léon Dutreuil de Rhins.

In this paper we are presenting our work in developing a collaborative multilingual *terminological dictionary*[3] dedicated to these digitized resources, that will interact with a Cross-Lingual Information Retrieval system (CLIR). This companionship between the dictionary and the CLIR system will achieve two results: (1) trigger users who are browsing and searching the archive to contribute to the dictionary, (2) translate search requests using the dictionary, so that both systems will help each other.

The next section will describe the problems and show some related work, then we will propose our solution in section three. In the fourth section, we will present the design of the system and its components, and in the fifth section we will describe the process of seeding the dictionary, and present the current prototype. A conclusion and some perspectives will follow.

2 Problems and Related Work

Producing a domain-specific multilingual terminological database of high quality is a very difficult and complicated task, and depends heavily on human terminologists (Cabre and Sager 1999).

Traditionally, this task starts by studying the domain, finding its logical elements and sub-domains, then a team of terminologists starts analyzing specialized textual material to find the most relevant and interesting terms, and then teams of terminologists translate each term into the targeted languages, and they would define it, and add the necessary descriptive information, to be standardized and adopted. This approach needs a lot of resources, particularly, human

[3] We use this somewhat unorthodox term to denote a collection of terms for concepts in some knowledge area, possibly containing information such as definitions, contexts, domains & sub-domains, and examples of use. While that information may be useful for helping readers access a specialized document in a foreign language, it is by no means of the kind and quality of what professional terminologists would put in a terminological database such as IATE (http://www.iate.europa.eu) — although they might use it as « raw material » for their work.

resources. Only huge organizations are able to conduct such kind of work. Table 1 provides some examples of online multilingual terminological databases built and made available by large organizations and standardization bodies.

Table 1 : some existing multilingual terminological online databases

Name	**Number of multilingual terms**	**Languages**	**Domain**	**Provider**
IATE (IATE 2008)	8.4 million terms	The 23 EU official languages	General, 155 domains	EU
UNTerm (IATE 2008)	80,000 terms	The six UN official languages	100 subjects related to the UN	UN
FAOTerm (FAO 2008)	58,000 terms	7 languages	FAO related domains and organizational bodies	FAO
Electropedia (IEC 2008)	20,000 terms	9 languages	electrical and electronic terminology in 75 categories	International Electrotechnical Commission (IEC)
The Great Terminological Dictionary (La Grand Dictionnaire Terminologique) (OQLF 2008)	3 million terms	French, English, and Latin	200 categories	Quebec board of the French language (Office québécois de la langue française)

As shown in table 1, the providers have mature resources and experience to build such data bases. In fact, those online systems are continuations to efforts started decades ago, and they have been compiled using material from older and existing databases.

Not only are the conventional approaches in building a multilingual terminological data base very expensive, but it is usually difficult to achieve good coverage (either informational or linguistic), especially in particular on specific domains. Besides, terminologists are more prone than domain experts to introduce inaccuracies. In this situation, a possible and (we think) necessary solution is to depend on volunteers knowing quite well the domain at hand, and let them contribute to the access multilinguization process through a collaborative environment. For example, *ITOLDU* (Bellynck, Boitet et al. 2005) collected 17000 English-French terms in 20 technical domains from 250 French students (learners of Englich). *Yakushite.net* (Murata, Kitamura et al. 2003) is another

example, where users contribute to bilingual dictionaries (organized following a domain hierarchy) that is used to enrich both the online Pensée machine translation system and the human translation aids. Also, Papillon (Sérasset 1994; Boitet, Mangeot et al. 2002; Sérasset 2004) is a Jibiki-based (Mangeot 2006) general purpose collaborative multilingual dictionary.

The problem with building terminological databases collaboratively is that it is difficult to attract domain experts to contribute: in the examples mentioned above, one can not expect massive contribution from normal people who are only visiting the dictionary, and one can even less expect volunteers to replace professional terminologists. A volunteer could translate a term, but s/he may not be able to give full descriptive information about a term (its definition, usage, domain, context…), and, if s/he may, it will certainly not be in the way a professional terminologist would.

Another point to be considered is that such a database should be *seeded*, so that visitors can find initial data to start the contribution. For that, using online resources seems to be a very promising option. Projects such as MultiMatch (Jones, Fantino et al. 2008), and PanImage (Etzioni, Reiter et al. 2007) use Wikitionaries (Wikitionary 2008), Wikipedia (Wikipedia 2008), and other online dictionaries for this problem. A similar approach will be used in our case with the consideration of the DSR's data and needs.

3 Proposed Solution

3.1 Overview

Our proposal is to build an easy to use collaborative environment where normal online archive visitors are oriented to contribute spontaneously by translating related terms in the languages they know, and possibly, what they are translating is the search terms that they use to browse the archived data.

As shown in figure 1, historical physical books have been digitized and indexed into a SOLR-based search engine. And we analyzed the output OCR text to initialize our term database.

We expect users to send monolingual search requests in any language supported by our system to get multilingual answers. Having a term base of multilingual equivalences could achieve this (Chen 2002) (Oard 1999). A bilingual user who could send a bilingual search request could be a valid candidate to contribute, in fact the same bilingual request could be a valid dictionary contribution, and so the multilingual request. We plan that users who use our search engine will use the terminological dictionary to translate their requests and will be able to edit and add new entries to the dictionary spontaneously.

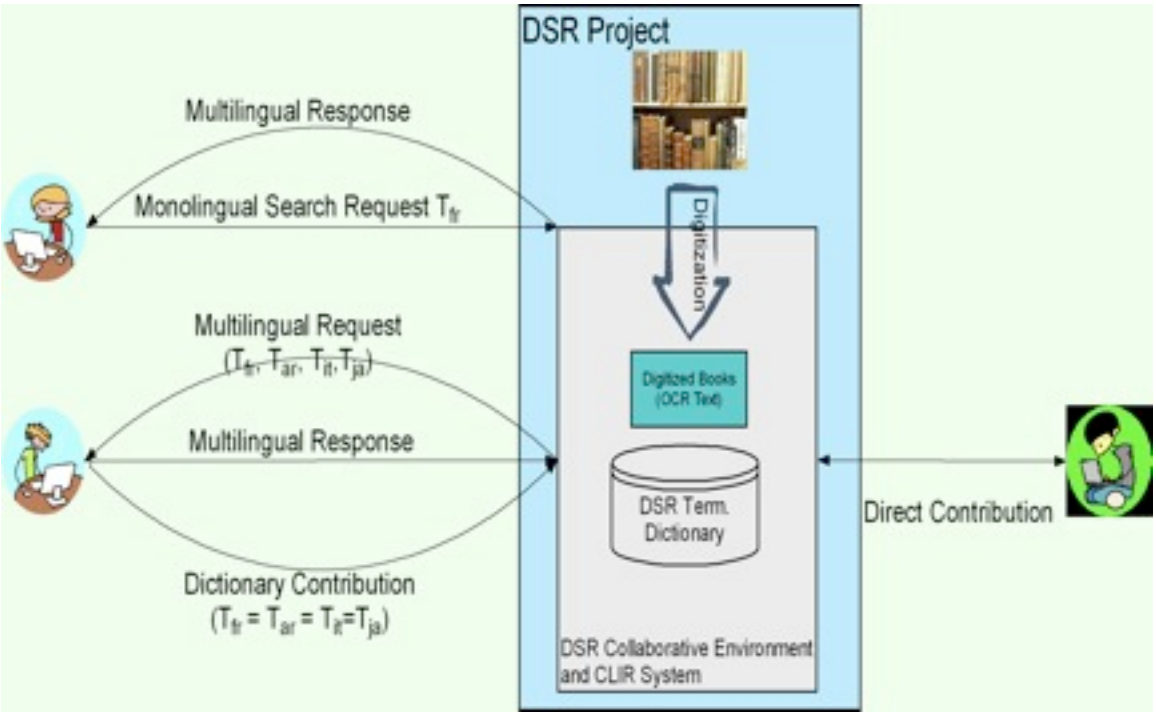

Figure 1 : general view of the proposed solution

Note that the collaborative dictionary could have at the same time direct contributors and visitors.

3.2The Online Collaborative System

The search engine we use for indexing the OCR books is Solr-Apache (Apache 2008), an open source search server based on the Lucene Java search library, configurable to be used for languages other than English. Its availability and advanced features make it a good choice for our experiment. As shown in figure 2, the OCR text will be indexed to an instance of Solr. As the online archive contains scanned images and associated OCR text of each book, users could be more interested in the scanned images, while digitized OCR text can make these images searchable and improve accessibility of the books.

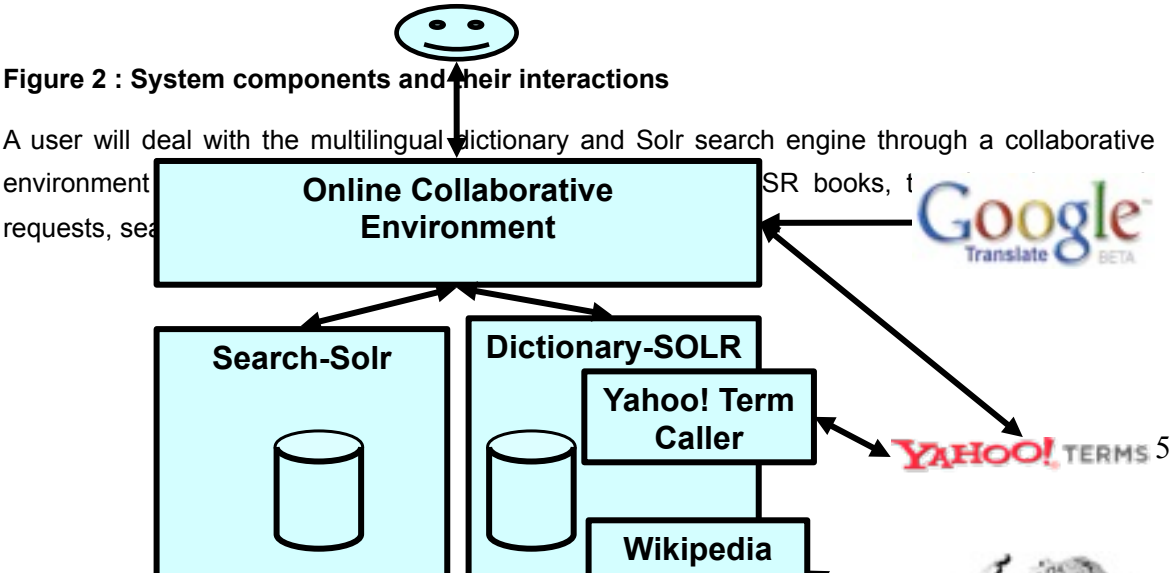

Figure 2 : System components and their interactions

A user will deal with the multilingual dictionary and Solr search engine through a collaborative environment SR books, t requests, sea

Volunteers will be equipped with some online reference data and assistance such as showing the suggested translation by Google Translate (Google 2008) and any initial translation available in the database.

Offline tools are interacting with some online multilingual resources to prepare the initial multilingual database as we will describe later.

3.3Multilingual Dictionary Structure

Figure 3 shows the architectural design of the collaborative multilingual terminological dictionary we are developing.

User layer will interacts with users by a simple set of HTML pages and web forms that will interpret the contribution and search logic developed at the business layer.

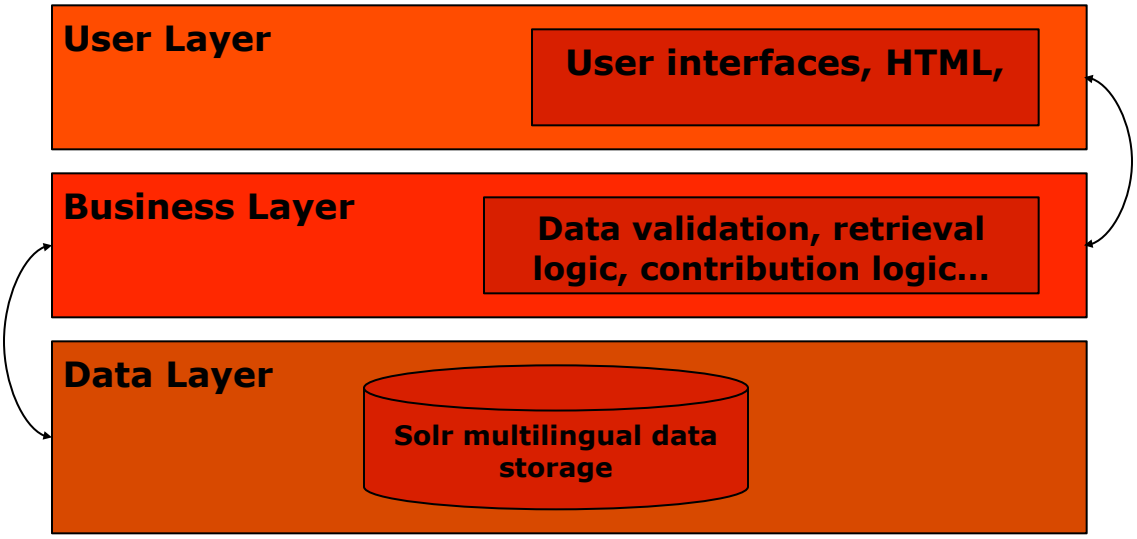

Figure 3 : 3 tiers architecture of the terminological dictionary

At the data layer, multilingual dictionary entries will be indexed into a Solr index, which will give mature dictionary lock-up facilities (provided by a powerful search engine).
Each entry will be indexed as a solr XML document, each field of the document will be configured to have different indexing and querying analyzer based on its language, here is a simple multilingual document.
Note that the structure can be changed dynamically to include any kind of information needed later. Its simplicity will make it easy for users to contribute, their contribution will be automatically transfer into an XML document and indexed into the dictionary, contributors are not required to provide descriptive information, while it is very important for a term base, it is not in the context of a multilingual search engine.

```
<doc>
      <field name="id">27932</field>
      <field name="en">kharosthi script </field>
      <field name="fr">Alphabet kharo□□hî</field>
      <field name="ja">カローシュティー文字</field>
      <field name="ar">الخاروستي النص </field>
      <field name="th">อักษรคโรษฐี</field>
      <field name="de">Kharoshthi-Schrift</field>
      <field name="zh">佉卢文</field>
      <field name="pt">N/A</field>
      <field name="ru">Кхароштки</field>
      <field name="it">Kharoshthi</field>
      <field name="se">Kharosti</field>
      <field name="sa">N/A</field>
      <field name="av">N/A</field>
      <field name="Q">***</field>
      <field name="OCRTextUsage">Ancient Khotan : vol.1 / page 213:
      The conclusion to be drawn from this current use of an Indian
      language is greatly strengthened by the Kharosthi script of the
      records; for we know that within India this script was peculiar
      to that region of which Taxila and the adjoining Gandhara were
      the historical and cultural centres for centuries before and
      after the commencement of our era. </field>
      <field name="subject"> Afghanistan, Pakistan </field>

</doc>
```

Figure 4 : a multilingual entry to be indexed into Solr-dictionary

4 Dictionary Initialization Experiment

In this experiment, we are trying to imitate the typical manual construction of a terminological database. This process usually contains two main time consuming steps: (1) document consultation, during which a terminologist tries to find the important terms available in a relevant set of documents, and (2) terminology translation. This process has been used to construct the initial manually developed database, which contains around 700 terms, available in up to 8 languages. In this experiment we associated these terms with our database by transforming the entries into the appropriate xml format.

Our process will do a similar job automatically, but the results will only have the status of a proposal “raw material”). Figure 5 shows the main tools and steps of our approach.

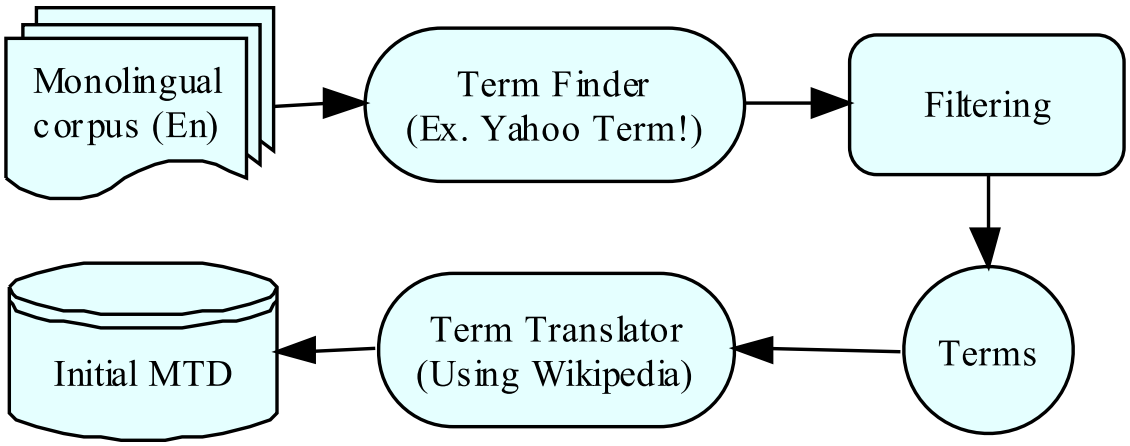

Figure 5 : the process of seeding the database

Each page of an English OCR book is sent to Yahoo! Terms (Yahoo 2008) to find the most important terms, assuming these terms are good candidates to be in our database. After filtering them, a tool will translate them using Wikipedia.

For each term provided by Yahoo! Terms, we form a Wikipedia URL to the term's article at the English Wikipedia, for example, the term "cuneiform script" would give the following URL: http://en.wikipedia.org/wiki/Cuneiform_script

We retrieve that article and analyze it to construct a multilingual entry. As an English article usually contain links to equivalent articles in different languages, we use these links to translate the term, and we use also the categorization information to associate a simple descriptive classification to the entry. From such an article, we find the relevant terms and we translate them again using the same method.

As a first experiment, 80000 English terms have been extracted from the historical books. More than 22000 terms have been multilingualised (they are now in 1 up to 20 languages) automatically using Wikipedia and other cultural glossaries.

5 Prototype

5.1Implementation

The first prototype of the system has been developed using Java Server Pages; the server is running over Apache-Tomcat 5.x. Users will interact with the system using a very simple web interface. When they will search the Digital Silk Road Archive in their own language, their request will be multilingualised and sent to the Solr search engine.

Search results will be shown to the user in an HTML page, from which volunteers will be allowed to change the translation(s) of their request, or add new translations in the dictionary. Here lies an interesting synergy between a search tool and a contribution platform.

Users can already add new terms through a very simple form where they are asked to add the translation of the term in any of the languages they know, and the system provides them with the term translated using Google Translate and other online dictionaries as well as the available translations in the local database.

The dictionary is implemented to serve as an independent entity as well: users can access it directly, search its database and add new entries.

5.2 Screen shots

The first screen shot, figure 6, shows the main user interface. A user can input his search term in one of the fields corresponding to his language, and then he can translate the term into other languages to send a multilingual search request.

In the example of figure 6, a French user wrote “lapis-lazuli” into the French field and clicked on “Translate search term” to have it translated into 11 languages (the current interface shows only 11 languages).

Figure 6 : main web interface, where users can add their search terms, and get or put their translations in the dictionary

The user can click into “Add suggestions” to add or edit translations of the term. As shown in figure 7, the user is provided with the translation from Google Translate and the entries available in the local database. The user can add his own contribution and save it using that simple and easy interface.

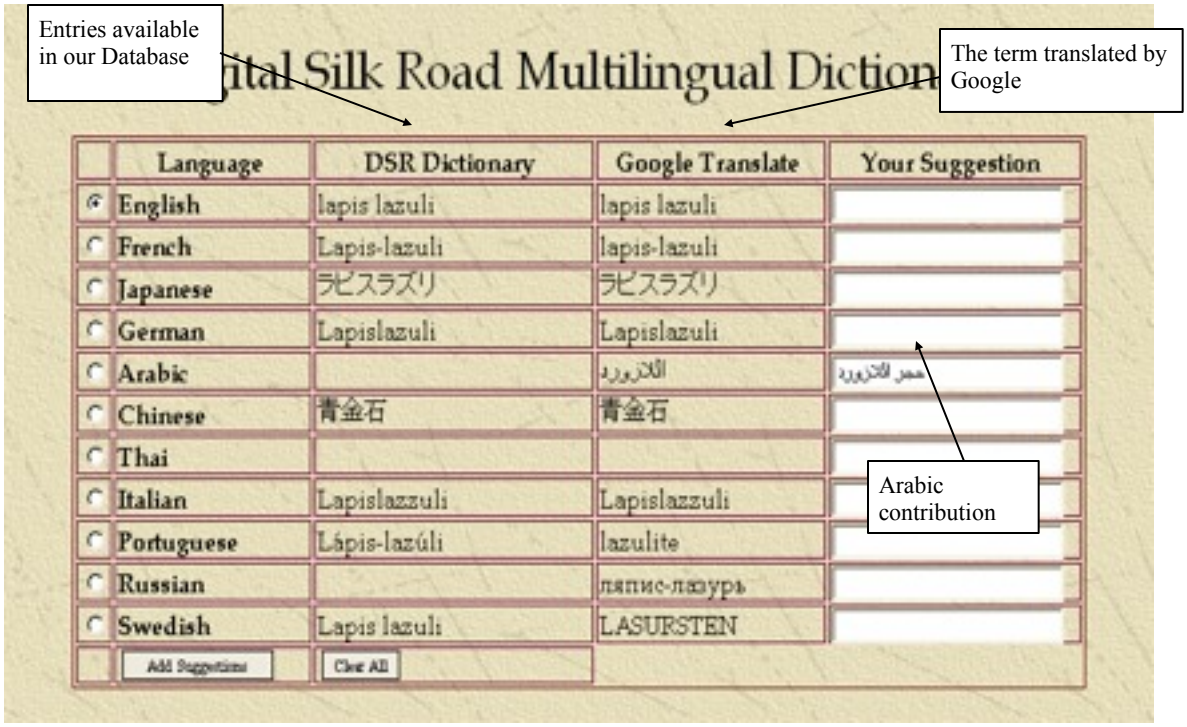

Figure 7 : a contribution interface where the user adds his contribution

Figure 8 shows multilingual search results for the term “lapis lazuli” after it has been translated.

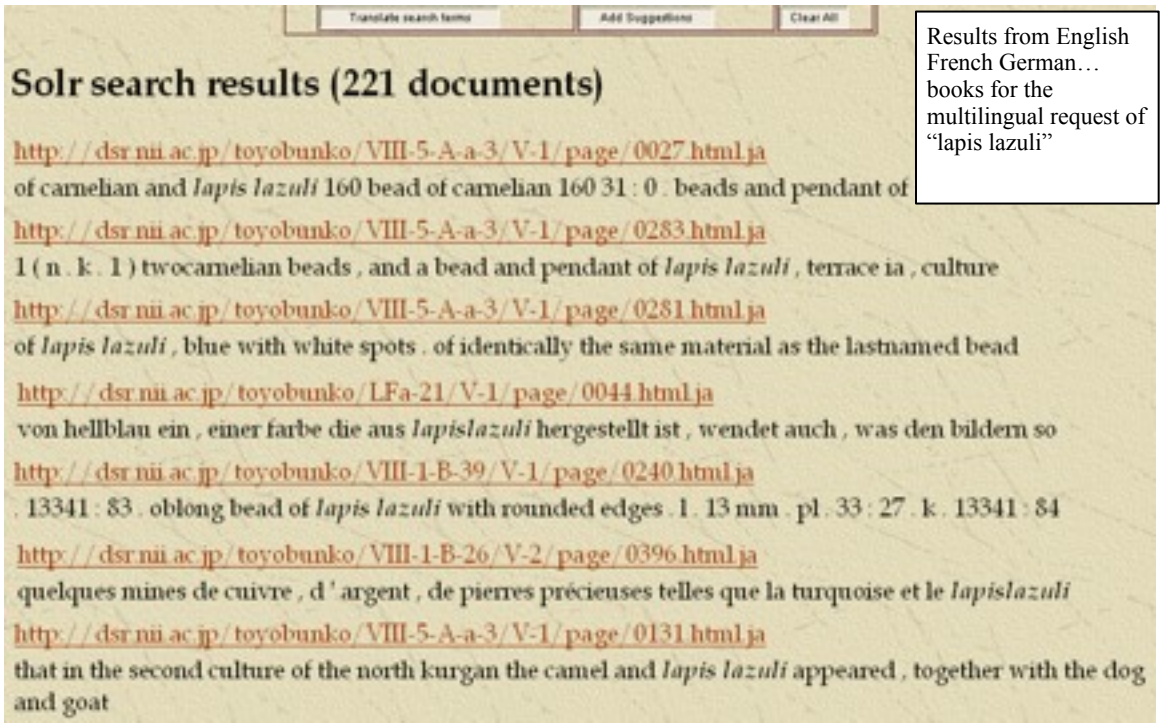

Figure 8: multilingual search results for “lapis lazuli”

The results contain links to English, French, and German books available in our archive that contains the term lapis lazuli, or its equivalences in other languages. At any time, a user can choose to add his new suggestion.

6 Conclusion

In this paper we presented our work in developing a collaborative multilingual term base dedicated to a digital archive, which works as a companion to a dedicated CLIR-based system.

Building terminological multilingual databases is very expensive and complicated task; it needs resources from huge organizations, while smaller communities can not afford to build their own multilingual databases, even collaboratively, as collaborative environments are designed only for sophisticated users (mostly terminologists, which brings the problems of traditional approaches). Ignoring the needs of normal users (likely subject matter experts) will almost surely lead to a failure. Another important problem is how to make the collaborative process attractive enough, especially for people who are interested in the community (website) itself.

The collaborative approach we are proposing and developing will consider all the contributors needs, specially the attractiveness and easiness of the contribution, by attaching the collaborative environment to a multilingual search tool, where users search requests are translated by the local multilingual database, and then they can edit the translation or add it, if it is not available.

To initialize the dictionary, we suggested a process that uses the archived data to extract the important terms and translate them from online accessible resources such as Wikipedia. As an experiment, we extracted 80000 monolingual key terms from the digitized historical books we have, using the Yahoo Term! API. After suitable filtration, we obtained translations for more than 22000 terms (into 1 up to 20 languages).

The initially developed prototype demonstrates the main functionalities of the proposed collaborative environment. Another version is under development currently, to meet all the needs of contributors and DSR's online archive visitors. In the next phase we will try to make the collaborative environment and the construction process generic and dynamic enough so that any community can create an instance of the platform and dedicate it to their needs and data. Similar communities should be helped to communicate and use their resources along with available online multilingual resources. As for the Digital Silk Road database, new languages will be added, and care will be given to languages along the Silk Road, especially to poorly equipped languages (pi-languages) and "frozen" languages like Sanskrit which was included in the first experiment.

7 References

Apache. (2008). "Solr." Retrieved 20-3-2008, from http://lucene.apache.org/solr/.

Bellynck, V., C. Boitet, et al. (2005). ITOLDU, a Web Service to Pool Technical Lexical Terms in a Learning Environment and Contribute to Multilingual Lexical Databases Springer Berlin / Heidelberg.

Boitet, C., M. Mangeot, et al. (2002). The PAPILLON project: cooperatively building a multilingual lexical data-base to derive open source dictionaries \& lexicons. Proceedings of the 2nd workshop on NLP and XML - Volume 17, Association for Computational Linguistics.

Cabre, M. T. and J. C. Sager (1999). Terminology: Theory, methods, and applications J. Benjamins Pub. Co.

Chen, A. (2002). "Cross-Language Retrieval Experiments at CLEF 2002." in CLEF-2002 working notes,.

Etzioni, O., K. Reiter, et al. (2007). Lexical translation with application to image searching on the web. MT Summit XI, Copenhagen, Denmark.

FAO. (2008). "FAO TERMINOLOGY." Retrieved 1/9/2008, 2008, from http://www.fao.org/faoterm.

Google. (2008). "Google Translate." Retrieved 1 June 2008, 2008, from http://translate.google.com.

IATE. (2008). "Inter-Active Terminology for Europe." Retrieved 10/10/2008, 2008, from http://iate.europa.eu.

IEC. (2008). "Electropedia." Retrieved 10/10/2008, 2008, from http://dom2.iec.ch/iev/iev.nsf/welcome?openform.

Jones, G. J. F., F. Fantino, et al. (2008). Domain-Specific Query Translation for Multilingual Information Access Using Machine Translation Augmented With Dictionaries Mined From Wikipedia. Proceedings of the 2nd International Workshop on Cross Lingual Information Access - Addressing the Information Need of Multilingual Societies (CLIA-2008), Hydrabad, India.

Mangeot, M. (2006). Dictionary Building with the Jibiki Platform. Software Demonstration. Proc. of EURALEX 2006, Torino, Italy.

Murata, T., M. Kitamura, et al. (2003). Implementation of collaborative translation environment 'Yakushite Net' MT Summit IX. New Orleans, USA.

NII. (2008). "Digital Archive of Toyo Bunko Rare Books." Retrieved 1 June 2008, 2008, from http://dsr.nii.ac.jp/toyobunko/.

Oard, D. (1999). Global Access to Multilingual Information. Fourth International Workshop on Information Retrieval with Asian Languages. Taipei-Taiwan.

ONO, K., A. KITAMOTO, et al. (2008). Memory of the Silk Road -The Digital Silk Road Project-. Proceedings of the Conference on Virtual Systems and Multimedia (VSMM08), Project Papers, Limassol, Cyprus.

OQLF. (2008). "Le grand dictionnaire terminologique." Retrieved 1/9/2008, 2008, from granddictionnaire.com/.

Sérasset, G. (1994). "Interlingual lexical organisation for multilingual lexical databases in nadia." COLING-94 **volume 1**: pages 278-282.

Sérasset, G. (2004). A Generic Collaborative Platform for Multilingual Lexical Database Development. COLING 2004 Multilingual Linguistic Resources, pages 73-79, Geneva, Switzerland, Aug. 2004. .

Wikipedia. (2008). "Wikipedia." Retrieved 1 June 2008, 2008, from http://www.wikipedia.org/.

Wikitionary. (2008). "Wikitionary." Retrieved 1/9/2008, 2008, from http://en.wikipedia.org/wiki/Wiktionary.

Yahoo. (2008). "Yahoo Terms." Retrieved 20-3-2008, from http://developer.yahoo.com/search/content/V1/termExtraction.html.