# Leveraging User Interaction and Social Tagging for Improving Cross-lingual Information Access in Digital Libraries

Juliane Stiller

Berlin School of Library and Information Science, Humboldt-Universität zu Berlin,
Dorotheenstr. 26, 10117 Berlin, Germany
juliane.stiller@ibi.hu-berlin.de
http://www.ibi.hu-berlin.de/

**Abstract.** Evaluation of interactive cross-lingual information retrieval systems has been the focus of recent research. The goal is to support the users in formulating effective queries and selecting the documents which satisfy their information needs regardless of the language of the documents. This dissertation aims at harnessing the user-system interaction, extracting the added value and integrating it back into the system to improve the cross-lingual information retrieval system for successive users. To achieve this, user input at different interaction points will be evaluated. This will among others include interaction during user-assisted query translations, implicit and explicit relevance feedback and social tags. To leverage this input, explorative studies need to be conducted to establish user input which might be beneficial and the methods to extract it. The dissertation wants to extend the scope of interactive cross-lingual information retrieval by harnessing user input as a mean for improving cross-lingual information retrieval tools.

**Key words:** Digital Libraries, Interactive cross-language information retrieval, Social tags

## 1 Introduction

Digital libraries which aggregate and provide access to collections in different languages are confronted with challenges of multilingual information access (MLIA). Especially in Europe with its linguistic diversity, research in information retrieval is directed towards finding solutions to overcome language boundaries.

Multilingual access to digital libraries has different facets. A first step towards a multilingual digital library is to provide the interface (static & dynamic pages) in different languages. This can be implemented by either presenting the interface in a default language, which the user can change manually via flags or drop-down menus or by automatically detecting the user's location by IP-address (assuming the user speaks the language of the country he is situated in) or browser language and switch the interface language accordingly.

The most challenging aspect of multilingual information access is the search capability of the system. This is referred to as cross-lingual information retrieval (CLIR). According to Peters and Sheridan [1], CLIR is one component of the broad defined term MLIA and can also be referred to other concepts of information access such as result representation, result filtering and browsing. Throughout this paper the term CLIR will be used to address all aspects of multilingual search.

CLIR is characterized by differences in query and document language [2]. To overcome this language barrier, one could either translate the documents, translate the query or both. A common approach is translating the query using natural language processing tools such as dictionaries. Due to the shortness of queries, the identification of the query language and the disambiguation of the query term in the source and target language are the main challenges in CLIR.

Tools for CLIR such as dictionaries are not universally available in every language needed or in every domain covered in digital libraries. This dissertation aims at harnessing users' input for improving and extending existing CLIR processing tools and consequently CLIR systems in their entirety.

Leveraging the power of users has become a central research issue in many fields (witness the occurrence of the term crowdsourcing in different literatures). In CLIR research, so far little attention has been paid to understanding users as a pool of unlimited language skills which can help to improve machine translations.

The aim of this research project is therefore to harness users' input in a CLIR system by determining interaction points between users and the system which provide valuable opportunities for leveraging the users' input to improve multilingual access to content in digital libraries.

In the next 2 chapters the research field and its related topics will be presented. Particularly studies touching on interactive CLIR and user collaboration such as social tagging and their results will be interpreted to define areas of interest.

The fourth section will outline the proposed approach including studies and experiments. The last section will propose research issues for discussion at the consortium.

## 2 Context of Work

Interest in multilingual access to content dates back decades and was intensified with the emergence of the World Wide Web. The linguistic diversity in Europe increased the demand for accessing content in languages which are not understood by all users. Especially digital library projects like Europeana[1] - aiming at making Europe's cultural heritage digitally available - depend on solutions for CLIR.

This dissertation project is situated between multilingual information retrieval and user interaction research. The goal is to bridge a gap between re-

[1] http://www.europeana.eu/

search in interactive cross-lingual information retrieval and studies focusing on establishing sustainable ways to leverage user input.

Interactive CLIR is the process of finding relevant documents in languages different from the query language satisfying the users' information need by coupling system capabilities with users' input. Oard et al. define several scenarios for user-system interaction within a search process [3]:

- query formulation,
- query translation,
- document selection and
- document examination.

Harnessing user interaction and collaboration for sustainable improvements in CLIR should be a part of the interactive information system research. In the context of this dissertation project, user interaction will be extended to incorporate user collaboration such as social bookmarking and the use of tags, but also other system interaction components like query reformulations and relevance feedback. All these aspects can help to improve cross-language information retrieval on the basis of users' input.

For CLIR, social tags are especially interesting as they represent the users' experience with an object. Linguistically divers tags can broaden access to multilingual resources either through browsing functionalities or through search.

## 3 Related Work

### 3.1 Interactive CLIR

The challenges of CLIR have received considerable attention in the field of information retrieval. Starting with TREC specifically with CLEF (Cross Language Evaluation Forum) & NTCIR, evaluation of CLIR systems components attracted attention. Since the first CLEF campaign in 2000, there has been an increasing interest in user-centered approaches in CLIR which led to the interactive tracks of CLEF. Interactive CLEF (iCLEF) advocated for system evaluation which pays attention to the user and their information-seeking behavior [4]. Of interest is here whether the system manages to support the user in formulating effective queries and in understanding the retrieved documents [2]. It became more and more obvious that interactive components are essential in a CLIR system to satisfy users' information needs appropriately [5].

A number of researchers have evaluated user interaction during the process of term translation which is a unique process for CLIR [3]. This so called user-assisted query translation features different user interaction opportunities:

1. Users determine language of their query,
2. Users select or deselect query translations from alternatives offered by the system (this can happen before the retrieval process starts or after the ranked retrieval offers a list of results),
3. Users add or edit term translations.

These features implemented in different CLIR prototypes were evaluated in interactive studies. One of the conclusions was that users learned from their interactions and adapted their strategies to the machine's performance [6]. In general, users actively manage the possibilities introduced by a system supporting user-assisted query translation [3]. For interactive CLIR, it is also essential that users understand the system's approach, so they can intervene when the system is not performing as expected.

In addition, user-centered studies were conducted within the Clarity Project, establishing a CLIR system for languages with few electronic resources [7]. Here, the key conclusion was that within interactive information retrieval each subtask should be evaluated separately to test each components' capability [8].

The first serious attempt to harness user input for enriching existing dictionaries was made by the CACAO project[2]. The project established a first prototype, which allows users to add their own translations and to select or deselect alternative translations during the query formulation process. Translations added by users will be included in a dictionary after qualitative analysis of an administrator. So far, no reports on the experiences with this approach are given. Not only uses CACAO translations typed into the system by a user for dictionary enrichment, they also analyzed The European Library[3] log files to harvest new translations assuming that users of a multilingual digital library will repeat queries in different languages [9]. This explorative study is one of the first approaches to extract added value from user-system interaction and integrate it back into the system.

Ingeneral, the research in interactive CLIR to date has tended to focus on user-centered evaluation of interactive features for CLIR systems rather than evaluating the added value and benefits user interaction can have on the system. The existing accounts have drawn attention to the interaction design for systems and the evaluation of the results of an interactive process for individual users. Now, research should expand on leveraging interactive features as a source for improving CLIR tools by integrating users' input.

The scope of interactive CLIR research broadens when user input as a source for system improvements is taken into consideration. This dissertation wants to extend the evaluation of interactive CLIR features and develop a holistic approach by harnessing the users' input to enhance retrieval effectiveness and user experience for future users.

### 3.2 User Collaboration - Social Tagging

Another part of user-system interaction which was not researched in the context of CLIR yet is the collaborative content users add to resources for personal information management or the purpose of sharing it within a certain community.

[2] cross- language access to catalogues and on-line libraries, http://www.cacaoproject.eu/

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

A lot of recent research in this area focused on using collaborative input such as social tags as document enrichment and substitution for controlled vocabularies [10]. The applied tags represent the users' interpretation of the documents content [10]. User tags form folksonomies which are an alternative to controlled vocabulary applied by information professionals. This enrichment of documents by folksonomies is accepted to produce added value for users.

Studies of collaborative information systems like del.icio.us[4] as a bookmarking service and Flickr[5] as a media sharing service analyzed tag categories and established models for tag category systems [11][12]. Others aimed at investigating the users' tagging behavior [13].

Most studies in the field examined advantages and possible tradeoffs of folksonomies. Especially the quality and usefulness of tags were investigated [14]. The multilinguality of tags played a minor role in research of folksonomies so far. A few studies were conducted to analyze users' motivation to tag certain objects multilingually [15]. These studies focused on the users and their capability to apply multilingual tags as means for indexing and retrieval. Results emphasize that users tag multilingually if an incentive is given to do so like making their resource universally accessible. In addition, they refer to users' search behavior for objects which were tagged multilingually or users' evaluation of descriptiveness and helpfulness of multilingual tags [16]. One result is that tags need to be understood to have a positive impact on the user experience, even if that means to hide the irrelevant tags [16]. Multilingual tags as means of document representation which enables broader access across languages is not part of the scientific discussion yet.

In the research literature, the aspect of multilingual tags expressing the content of resources in different languages was neglected. How many of the tags are translations of each other? Can they be used to enrich or support document translations? Most of the studies referred to above investigated the implications of multilingual tags on users' behavior if tags were not in the mother tongue of the users.

A very interesting approach was taken by Noh et al. [17] who suggest a translation of tags by comparing the similarities of tag networks. They do not draw the conclusion to use the method to map tags in different languages to enhance existing dictionaries.

Social tagging and folksonomies were also evaluated in the context of CLEF. The aspect focused on here was the use of the multilingual search interface in a photo-sharing environment such as Flickr [18].

## 4 Proposed Research

### 4.1 Setting a Theoretical Framework

In CLIR systems, interactive components are crucial to accomplish search tasks [5]. Studies were targeted on evaluating the capabilities of a system to support

[4] http://delicious.com/
[5] http://www.flickr.com/

users in formulating effective queries. But can these interactive components also support successive users by offering better translations through user input? So far, evaluation of interactive CLIR features focused on satisfying the individual user needs, but feedback from users given at different interaction points during the search process can be beneficial for all future users.

CLIR faces two striking challenges compared to monolingual retrieval. The language of the query term needs to be identified and needs to be disambiguated while translated into the target language(s). The current state-of-the-art in machine translation and available language resources does not offer satisfactory and appropriate query and document translations to support CLIR effectively. Good dictionaries across specific domains and in certain language pairs are not universally available. The users' input as a source for translation and document enrichment through tagging and interaction should be considered here. Harnessing this input to improve the systems capability to search cross-lingually and to meet the challenges associated with it can be economically cheaper than equipping a system with commercial CLIR tools. One scenario could be user-assisted query translations, where the user interacts with the system and adds, enriches, replaces or selects translations for a query. Another possibility is leveraging user-generated content such as multilingual tags and use it in the query translation process or as multilingual document enrichment. There are two main scenarios where the users' input could be incorporated into the system to enhance multilingual information retrieval:

1. Harnessing multilingual tags for enriching metadata and disambiguating query terms and
2. Improving existing dictionaries by adding translations entered into the system by the user, e.g. directly form the user or via log files analysis of different user interaction during the search process.

The first step in this dissertation project is to bridge the gap between interactive CLIR and harnessing the user input for enhancing multilingual access to content. This means establishing a CLIR system with interaction points highlighted by Oard [3] and enriching it by interactive features which can be leveraged for enhancing CLIR tools. First research was conducted in this area but a theoretical framework was not put in place which might unite all these approaches.

### 4.2 Scenarios for User Interaction and Collaboration

Interactive CLIR should not only consider individual users and their search processes but should focus on learning from previous user-system interactions and allowing these experiences to be incorporated into the system. This requires the identification of different interaction points, where users are encouraged to leave their feedback in the form of alternative translations, relevance feedback or tags which can be harnessed and integrated back into the CLIR system.

Table 1 lists the points of interaction in CLIR and assigns tasks of user input which can be harnessed and used for improving language resources. For example,

**Table 1.** Interaction points and according user input

| IR component | Interaction points | User collaboration tasks |
|---|---|---|
| Query | Query formulation | Reformulation |
| | Translation | (De)selecting translations |
| | | Editing/adding translations |
| Document representation | Document selection | (De)selecting |
| | | Relevance feedback |
| | Document examination | Tagging |

the expression of an information need in form of a query leads to query formulation and query translation. User input at these two stages can be: reformulation, selecting or deselecting alternative translations, editing and adding translations. A log file analysis gives information about query reformulations within a session. A language change within this session could be a translation of the same query which should be saved and added to a dictionary or offered as alternative translation next time the user searches for the same query [9].

User interaction should focus on facilitating the users' retrieval process. As mentioned before, many studies are in place, which suggest processes for user interaction and evaluate their results. But none of the studies focused on utilizing users' input for improving the overall information retrieval process. For each interaction point and step within a cross-lingual information system, data analysis should validate the profitableness of users entries and interaction for future retrieval within the same system.

### 4.3 Gathering Data

Preliminary studies aiming at improving translations by analyzing user input [9] achieved promising results, which justify to pursuit research in this area. Users will interact with a CLIR system to satisfy their information needs anyway, so they produce data which can be leveraged to improve the system. Commercial players like Google are improving their functionalities mainly by closely observing the users' behavior.

To gather data, the user input at different points of user interaction will be analyzed and a method to harness this input and to integrate it back into the system will be determined accordingly. One promising approach already mentioned are the query logs [9] but also methods of explicit or implicit relevance feedback have to be considered. A key component is to identify the added value and integrate it back into the system. Europeana - as specifically outlined multilingual digital library - will mainly serve as the research object but also other systems, which offer the desired functionalities or where the desired functionalities can be easily implemented. These studies will be mainly of exploratory nature but they will identify an approach as valuable or not.

Research on multilingual tags will be conducted studying data from already existing collaborative information systems. As a big user base is needed, existing systems can offer more appropriate data than prototypes where the tag functionalities are still underdeveloped. Analysis will focus on the influence of different attributes on the occurrence of multilingual tags in explorative data evaluation. Multilingual tags can occur in three different variants:

1. Tags translated by users,
2. Monolingual documents which are enriched by tags assigned by users in different languages.
3. Multilingual documents which are enriched by tags assigned by users in different languages.

A collaborative information system is a technical system characterized by its main attributes: its users, its tags and its resources; they form a pair wise connection between them [13]. The interdependencies between these parts of an information system will be investigated. How does the interaction of these three major parts influence the occurrence of multilingual social tags? Can information systems be adapted to foster the adoption of multilingual tags? To analyze tagging data, a taxonomy of tagging systems will be established focusing on characteristics which influence the multilinguality of tags. Several taxonomies already exist and can be enriched by components influencing multilinguality. The descriptiveness of tags and the intent they were used for [11] are inherent in most taxonomy categories. On the basis of the established taxonomy, data from different tagging systems will be evaluated. The essential question here is the dependency of resources, tags and users and the influence of this holistic system on the multilingual occurrence of tags. First questions to answer are:

- What is the ratio of different language tags in different systems? Which components influence this ratio?
- Can new translation be inferred if tags in different languages occur on the same object?
- What is the ratio of concepts expressed in different languages?

Unifying the different results from these explorative studies in multilingual tags and other forms of interactive user input will serve as basis for a catalogue of features for interaction design in digital libraries. It should answer the question whether user interaction suggest beneficial results for CLIR. These results will serve as use cases which can easily be prototyped. The project will finish with an analysis of the effectiveness of the features within CLIR system.

## 5 Discussion

This dissertation work aims at leveraging the power of users to improve CLIR systems and its components. Analyzing user data produced while interacting with the system is as valuable as considering the proactive user input such as

social tags of information objects. Both tasks can be understood as user-system interaction which will be used to enrich existing elements of the system.

Especially interesting here is the question of how the user input and interaction with the system can be leveraged to identify potential added value and to integrate this back into the system. Looking at the different steps within the dissertation project, the following issues were identified where feedback and suggestions are appreciated:

**Interactive CLIR**
The user collaboration tasks which should be harnessed to improve CLIR are derived from interaction points listed by Oard [3]. For each interaction point, tasks for user collaboration and interaction were listed (table 1). Are there more tasks of interaction which can be harnessed? Looking at user-assisted query translation, the user collaboration tasks depend on the design of the system. Some systems might allow the editing of suggested query translations, others might only permit the user to select or deselect suggested translations. These conditions will influence the results of the analysis conducted in this research.

In addition, some interaction points were studied thoroughly but others such as document selection and document examination were not analyzed within a multilingual environment. This means that there are not many results to build the research on and establishing a base line might go beyond the scope of this project.

**Gathering data**
The main part of the dissertation will be to gather data from different systems. Working within the project of EuropeanaConnect enables access to data such as log files from Europeana but the design of the system might not involve the features needed for the analysis. Therefore different system should be taken into account, primarily digital libraries in a multilingual environment with cultural heritage objects. How can the data gathering in different systems look like? Log files are the common approach but are there different methods possible?

**Cultural dimension**
For multilingual tags it is also of interest to explore the influence of communities on the use of multilingual tags. Do online communities use the same language? And what about the concepts expressed in tags, do they differ culturally? How can multilingual tags be matched to each other? Can they serve as tool to disambiguate query terms or are they best included as document enrichment?

## References

1. Peters, C., Sheridan, P.: Multilingual information access. In Agosti, M., Crestani, F., Pasi, G., eds.: Lectures on Information Retrieval. Springer (2001) 51–80

2. Oard, D.: Multilingual information access. In Bates, M.J., Maack, M.N., eds.: Encyclopedia of Library and Information Science. 3rd edn. Taylor & Francis (2009)
3. Oard, D.W., He, D.Q., Wang, J.Q.: User-assisted query translation for interactive cross-language information retrieval. Information Processing & Management **44** (2008) 181–211
4. Oard, D.W., Gonzalo, J.: The clef 2001 interactive track. In Peters, C., Braschler, M., Gonzalo, J., Kluck, M., eds.: CLEF '01: Revised Papers from the Second Workshop of the Cross-Language Evaluation Forum on Evaluation of Cross-Language Information Retrieval Systems, London, UK, Springer (2002) 308–319
5. Gonzalo, J.: Scenarios for interactive cross-language retrieval systems. In: Proceedings of the Workshop of Cross Language Information Retrieval: A Research Roadmap Workshop held at the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (2002)
6. He, D., Oard, D.W., Plettenberg, L.: Studying the use of interactive multilingual information retrieval. In: ACM SIGIR Workshop on New Directions in Multilingual Information Access. (2006)
7. Petrelli, D., Beaulieu, M., Sanderson, M., Demetriou, G., Herring, P., Hansen, P.: Observing users, designing clarity: a case study on the user-centered design of a cross-language information retrieval system. J. Am. Soc. Inf. Sci. Technol. **55** (2004) 923–934
8. Petrelli, D.: On the role of user-centred evaluation in the advancement of interactive information retrieval. Information Processing & Management **44** (2008) 22–38
9. A. Bosca, L. Dini: Cacao project at the logclef track. In: Working Notes of the Cross Language Evaluation Forum (CLEF), Corfu, Greece (2009)
10. Peters, I.: Folksonomies: Indexing and retrieval in Web 2.0. Knowledge and Information Studies in Information Science. de Gruyter, Berlin (2009)
11. Golder, S.A., Huberman, B.A.: Usage patterns of collaborative tagging systems. J. Inf. Sci. **32** (2006) 198–208
12. Heckner, M., Neubauer, T., Wolff, C.: Tree, funny, to_read, google: what are tags supposed to achieve? a comparative analysis of user keywords for different digital resource types. In: SSM '08: Proceeding of the 2008 ACM Workshop on Search in Social Media, New York, NY, USA, ACM (2008) 3–10
13. Marlow, C., Naaman, M., Boyd, D., Davis, M.: Ht06, tagging paper, taxonomy, flickr, academic article, to read. In: Hypertext '06: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia, New York, NY, USA, ACM (2006) 31–40
14. Golder, S.A., Huberman, B.A.: The structure of collaborative tagging systems. Technical report, HP Labs (2005)
15. Vuorikari, R.: Can social information retrieval enhance the discovery and reuse of digital educational content? In: RecSys '07: Proceedings of the 2007 ACM Conference on Recommender Systems. (2007) 207–210
16. Vuorikari, R., Ochoa, X., Duval, E.: Analysis of user behavior on multilingual tagging of learning objects. In: Proceedings of the First Workshop on Social Information Retrival in Technology Enhanced Learning (SIRTEL 2007). (2007)
17. Noh, T.G., Park, S.B., Yoon, H.G., Lee, S.J., Park, S.Y.: An automatic translation of tags for multimedia contents using folksonomy networks. In: SIGIR '09: Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, ACM (2009) 492–499
18. Gonzalo, J., Peinado, V., Clough, P., Karlgren, J.: Overview of iclef 2009: Exploring search behaviour in a multilingual folksonomy environment. In: Working Notes of the Cross Language Evaluation Forum (CLEF). (2009)