User Behavior and Evaluation of Multilingual Information Access in Digital Libraries

Maria Gäde

Berlin School of Library and Information Science,
Dorotheenstr. 26, 10117 Berlin, Germany
maria.gaede@ibi.hu-berlin.de

Abstract. While the importance of multilingual access to information systems is undoubted, few truly operational systems exist and can serve as examples. This dissertation addresses the issue of what the user expectations and the consequences for system development are in a multilingual information environment. It starts with a general overview over the aspects of multilingual access in digital libraries. Building on previous experiences, the study focuses on a combination of log file analysis and an usability test on user needs and desired features for multilingual access based on a functional digital library with multilingual requirements (Europeana). I present the Europeana Clickstream Logger, which logs and gathers extended information on user behavior, and show first examples of the data collection possibilities. The outcome of the analysis is a description of user requirements. The dissertation concludes with the development of a possible approach for the design of multilingual information systems.

Keywords: CLIR, MLIA, user study, Log file analysis

1 Introduction

Most of the world´s people have a native tongue other than English. In contrast, more than 70% of the public web sites are expressed in English[1]. More and more users need support to retrieve relevant information across languages boundaries [16]. Especially digital libraries, such as Europeana[2], need to provide methods and tools that enable people to access multilingual information more effectively.

Increasingly, research is concerned with requirements for and the development of multilingual information systems. Multilingual information access (MLIA), as it is used in the dissertation, includes all issues of accessing, searching and retrieving data irrespective of the language in which information objects are expressed [17][18]. Cross-language information retrieval (CLIR) technologies are targeted on answering

[1] http://www.oclc.org/research/activities/past/orprojects/wcp/stats/intnl.htm
[2] http://www.europeana.eu

queries in one language with a list of objects in other languages. This can be achieved either by query or/and document translation, whereas the query translation usually is favored.

The research project reported in this paper focuses on the user side and aims to understand search processes, including users' interaction with multilingual digital libraries. The study addresses the issue of how users behave and interact and what the consequences for system development in a multilingual information environment are.

The paper is structured as follows: Chapter 2 gives an overview of the different levels and functions of MLIA systems. Chapter 3 discusses related work and the main findings of different studies, which will be summarized and included in the evaluation. The proposed research, focused on transaction log analysis (TLA) and an usability test is described in Chapter 4. Following in Chapter 5, the connection between user requirements and system requirements is shown. I conclude with a short summary and open questions to be discussed on the consortium.

2 Aspects of Multilingual Information Access in Digital Libraries

There are different levels of multilinguality the user is confronted with. This chapter gives an overview of varied levels and functions of search based systems such as Europeana. The European Digital Library will provide a multilingual common access to Europe´s cultural heritage.
Following, several aspects of MLIA regarding interface issues, the input of search terms and the display of the results are presented.

2.1 Multilingual User Interface

The most elementary level of multilinguality is the user interface. The translation of all static content elements on the information system's publicly viewable web sites and a systematic administration of language information for all content elements is called "language-skinning".
Currently two different options for language determination are available:

1. The user selects the interface language by a drop-down-menu or logos (e.g. flag images)
2. The language interface is selected automatically based on the language settings of the user agent (i.e. browser) or the geographic location of the user determined via IP-address.

2.2 Multilingual Search

The most essential component of a truly multilingual information system is the multilingual search function. Interactive MLIA systems provide an additional challenge to designers, because users may not have the necessary language skills to find and interact with objects written in multiple languages. To provide effective access to multilingual document collections, users require search assistance. Three approaches for multilingual search capabilities exist today:

1. Query translation: the original query is translated into additional languages that the document collection contains
2. Document translation: the documents in the collection are translated into the query language
3. Interlingua: both queries and documents are translated into a single language, which transforms the multilingual information retrieval process to a monolingual one.

The query translation process includes several stages such as query formulation and reformulation, language detection, and translation which posit particularly challenges for MLIA systems. The disambiguation of terms is even more problematic when more than one language is used.

Regarding multilingual search functionalities it needs to be clarified what kind of interaction is desired and useful to achieve optimal query translation and how systems can help the user select the most appropriate translations, especially with ambiguous terms?

2.3 Multilingual Result Representation and Filtering

The multilingual result representation can be performed at two levels: at the metadata or the digital objects level. . For textual documents, it needs to be determined whether result translation happens at the metadata level or the original document level. Within metadata records, the most appropriate translation candidates are titles and subject keywords.

The possibility to filter a result set by language determining can usually be implemented in two ways:

- Advanced Search: a user can determine the desired language of the documents in the result set by choosing from a list of available languages.
- Refinement filter for result set: the user can filter a result set by language after the first search has been processed.

How to present results in different languages is still an open question and needs further research [17].

3 User Preferences for Multilingual Information Access in Digital Libraries

In line with their efforts on establishing multilingual access to their content, several digital libraries have conducted studies on user needs and requirements, but few have paid specific attention to multilingual issues. The studies use a variety of research methods, including observation, surveys, interviews, experiments, and transaction log analysis. Although similar methods were used, a comparison or generalization of these findings is difficult because of the very different user groups involved. Some researchers selected their participants with regard to their information needs, others with regard to their language skills. The number of participants varies considerably between the different studies.

Since 2000 the Cross Language Evaluation Forum[3] (CLEF) has carried out several experiments with cross-language search tasks. Especially the interactive track, iCLEF, focused on problems of multilingual search assistance and the LogCLEF track 2009 provided interesting approaches [2][9][15]. One of the most basic outcomes of the experiments is that support for user-assisted translation of the query improves search results [24]. Additionally the TrebleCLEF Coordination Action[4] organized a "Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective" [25] to identify the essential features that MLIA systems should offer. The report presents some general requirements and best practice recommendations which were collected from experiments in iCLEF [24]. iCLEF 2008 as well as studies with the Google web search engine and the Google Translate service analyzed the behavior of users when facing strictly multilingual information access task in order to identify the differences in the search behavior according to the language skills [1] [20].

The European Digital Library (TEL)[5] and the EDL project also prepared user surveys and log file analysis to gather user requirements. They analyzed weblogs, the Gabriel guestbook and the Gabriel search engine queries [12], with the result, that multilingualism is one of the biggest problems in accessing portals. The full translation of documents is not required, only subject translation seems to be useful. Most users are satisfied to have the possibility to decide whether a document is relevant or not. [25].

The University of Padua analyzed the IIS http traffic logs of The European Library portal. One main characteristic about the sessions was that 77.44% involve only 1 query. Another finding was that the majority of visitors to the portal do not perform any query [7]. Additionally the Max Planck Institute for Informatics analyzed the verity server logs (action logs, user tracking) to research the user interaction behavior. In particular, they focused on the query and result-click history. Concerning the interface language selection, they found that the majority of users (84%) leaves the

[3] http://www.clef-campaign.org/
[4] http://www.trebleclef.eu/about.php
[5] http://search.theeuropeanlibrary.org/portal/en/index.html

default interface language English [7]. Another finding was that the most frequent keywords relate to European place names or subjects.

Through a study of library catalog search logs, the CACAO Project[6] found, that in a library operating in a multicultural context, about 20% of the queries are written in three languages, namely Italian, German and English [4].

The Europeana online survey conducted by the independent research agency IRN Research determined that over 50% of all respondents or 69.6% of those reaching the search results page refined the search by language [10].

The Eurovision system and the services of Tate Online were evaluated by multilingual users [5] [14]. The key finding was that many users are more likely to visit the collection site if it were translated into their preferred language and most of the participants were willing to accept a text which they could understand but was not perfectly translated. Within Multimatch[7] two extensive user studies were organized [15] [19]. The Clarity prototype was used to perform different tasks to explore interaction issues. Among other things, they suggest that users should have the possibility to choose the language they want to search in, depending on the individual skills and the task.

The previous findings show a growing interest in MLIA issues but also demonstrate that there are still a few open questions left. Especially inconsistent statements require further research. For example have all studies in common, that users are more likely to visit a Web site if it is translated into their preferred language. However, the majority of users still leave the default English interface.

4 Proposed Research approach

4.1 Methodologies

I am going to use the combination of quantitative and qualitative research methods, which provides different insight and therefore a more complete picture of users and their behavior. Transaction log analysis (TLA) and in-depth usability tests as complementary tools can allow a deeper understanding of users' interaction with information systems. The major advantage of TLA is that it automatically and passively captures real users in their own daily environment. It is also an effective way to detect discrepancies between what users say they do (for example in a survey or interview) and what they actually do when they use an online system or web site [6]. However, this method of analysis has a number of challenges and limitations. It is nearly impossible to identify individual users with absolute accuracy. The same user may use several IP addresses or several users can share one IP address. Using hostnames to group or locate users geographically can also be misleading. Aside from that, difficulties arise, when an attempt is made to answer questions concerning the

[6] http://www.cacaoproject.eu/
[7] http://www.multimatch.org/

users´ motivations. Log entries are limited to the users' interaction and do not reveal backgrounds or preferences [23].

Questions that arise from the TLA and those which cannot be answered by it will be addressed by the following usability test. The test design, including the tasks and questions for the participants will be based on previous observations in general. Like all research methods, there are a few limitations and pitfalls with qualitative user tests as well. The obtained data cannot be generalized.

4.2 Effective Log file Analysis for Multilingual usage of search based web sites – Europeana ClickStream Logger (CSL)

In the web, a transaction log is an electronic record of interactions during a search session between the information system and the users searching for information [11]. Common log file entries are general and therefore contain limited information concerning multilingual issues. Clickstream logging is a logging approach, which enables to mine complex data in order to analyze user paths. The term "clickstream" describes the path a user takes through a website. A clickstream is a series of actions or requests on the web site accompanied by information on the activity being performed [13]. It allows to track application state changes and therefore traces user behavior in a way that a traditional http transaction log is unable to. For the Europeana Clickstream Logs (CSL), different activity types or states with a particular focus on multilingual access aspects are logged. Table shows an abbreviated log entry for an interface language change:

Table 1. Abbreviated example from Europeana CSL with action: language change


```
["action" : "LANGUAGE_CHANGE", "agent" : "Mozilla/4.0
(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR
1.1.4322; InfoPath.2; .NET CLR 2.0.50727)", "date" :
"2010-10-27T20:50:58.226+02:00", "invoked_at" : { "d" :
"2010-10-27", "t" : "20:50:58" }, "ip" : "ES", "isBot"
: false, "lang" : "ES", "oldLang" : "EN",…]
```


The log entry shows a user from Spain, who changes the interface language from the default English to the Spanish translation. The URL of the requested page (where the interface language could in this case also be reconstructed from), the referrer page, the session and user id as well as page numbers are noted but not shown in this example. Different user (search) patterns can be discovered which will be used to better understand user behavior in a multilingual environment. Beside the general analysis of user behavior the study will mainly focus on language information from the log data such as interface language, country information from the IP address, query language as well as language of the results viewed.

4.3 Understanding users` behavior

The second part of the analysis consists of usability tests which will answer questions about the observed behaviors, dealing with the motivation and background of different users. For Europeana seven Personas have been identified and characterized according to their search behavior and literacy [8]. Currently the descriptions do not contain any language information in order to be adaptable to all European Countries. Through a usability test with subjects fitting to one Persona group from different countries we want to determine whether different language skills and/or cultural backgrounds influence their behavior.
Depending on the results from the usability test observations it could be necessary to interview a small number of users from different countries to provide detailed context data [3].

4.4 From User Requirements to System Requirements

The outcome of the analysis will be a catalogue of user requirements in a multilingual environment. The challenge is to find a balance between user and system requirements. Not everything users want is also feasible. Through the prioritization of requirements, one possible design approach for interactive MLIA systems will be presented.

Table 3 shows one example for the translation of user needs to system features. Assumed users want to control the query translation by choosing different translation candidates, the necessary conclusion would be an interactive MLIA system that supports user assisted query translation, at least as an opportunity if things go wrong.

Table 2. Relation between User Behavior and System Requirements

User Requirement	System Requirement
Transparent query translation	Include user-assisted query translation facilities. Support of translation candidates/suggestions from the user

5 Summary

Figure 1 summarizes the outline of the research project as described above. Starting from a state of the art overview of user studies in a multilingual environment a combined research approach has been presented: Log file analysis supported by in-depth usability tests. In the next few months the ClickStream Logs will be analyzed, so it will be possible to compare data from a longer period of time. The interpretation of the results will be used for the validation of the research question.

6 Discussion

I would especially benefit from presenting and discussing the Clickstream Logger results, since this will influence the qualitative analysis significantly. It would be helpful to receive feedback on the current log structure as well as the results and their interpretation concerning multilingual issues. Furthermore the selection of participants for the qualitative analysis and the criteria for prioritization of results could be discussed with the mentors.

Acknowledgement

This research project is partly funded by EuropeanaConnect. Especially, I want to thank Sjoerd Siebinga for developing the ClickStreamLogger.

References

1. Aula, A., Kellar, M.: Multilingual Search Strategies. In: CHI EA '09: Proceedings of the 27th international conference extended abstracts on human factors in computing systems. pp. 3865-3870 New York,: ACM (2009)
2. Bosca, A. and Dini, L.: CACAO Project at the LogCLEF Track. In: Working notes of the Cross Language Evaluation Forum (CLEF) (Corfu, Greece, 30 September -2 October 2009) (2009)
3. Boyce, C., Neale, P.: Conducting In-Depth Interviews. A Guide for Designing and Conducting In-Depth Interviews for Evaluation Input. (Pathfinder International Tool Series, Monitoring and Evaluation – 2) (2006) http://www.pathfind.org/site/DocServer/m_e_tool_series_indepth_interviews.pdf?docID=6301
4. CACAO: D7.4 User Requirements for Advanced Features (2009) http://www.cacaoproject.eu/fileadmin/media/Deliverables/CACAO_D7.4.pdf
5. Clough, P., Sanderson, M.: User Experiments with the Eurovision Cross-Language Image Retrieval System. In: Journal of the American Society for Information Science and Technology, 57(5), pp. 697 – 708 (2006)
6. Covey, D. T.: Usage and Usability Assessment. Library Practices and Concerns. Washington, DC: Digital Library Federation (2002)

7. EDLproject: M1.4, Interim Report on Usability Developments in The European Library (2007)
http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive/edlproject/downloads/M1.4_Interim%20Report%20on%20Usage%20and%20Usability.pdf
8. EuropeanaConnect: Personas Catalogue V.2 (2010)
9. Hofmann, K., Rijke, M. de, B. Huurnink, B., Meij, E. J.: A Semantic Perspective on Query Log Analysis. In: Working Notes for the CLEF 2009 Workshop (2009)
10. IRN Research: Europeana Online Visitor Survey Research Report Version 3 (2009)
http://version1.europeana.eu/c/document_library/get_file?uuid=e165f7f8-981a-436b-8179-d27ec952b8aa&groupId=10602
11. Jansen, B. J.: Search log Analysis: What it is, what's been done, how to do it. In: Library & Information Science Research, 28, pp. 407–432 (2006)
12. Janssen, Olaf: Gabriel 1997-2003 & Gabriel/TEL user survey (2003)
13. Joachims, T.: Optimizing Search Engines using Clickthrough Data. In: KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 133–142, New York, NY, USA, ACM Press (2002)
14. Minelli, S. H., Marlow, J., Clough, P., Cigarran Recuero, J.M., Gonzalo, J., Oomen, J. and Loschiavo, D.: Gathering Requirements for Multilingual Search of Audiovisual Material in Cultural Heritage. In: Proceedings of Workshop on User Centricity – state of the art (16th IST Mobile and Wireless Communications Summit (2007)
15. MultiMatch: D1.2, User Requirements Analysis (2006)
http://www.multimatch.org/docs/publicdels/D1.2Final.pdf
16. Oakes, M., Xu, Y.: Search Log Analysis at the University of Sunderland. Paper presented on the 10th Workshop of the Cross-Language Evaluation Forum (2009)
17. Oard, D.W.: Multilingual Information Access. In: Encyclopedia of Library and Information Sciences, 3rd Ed. (2009)
18. Peters, C., Sheridan, P.: Multilingual Information Access. In: Agosti, M, Crestani, F, Pasi, G (Eds.): Lectures on information RetrievalSpringer Lecture Notes In: Computer Science Series, vol. 1980. Springer New York, New York, NY, pp. 51-80 (2001)
19. Petrelli, D., Beaulieu, M., Sanderson, M.: User Requirement Elicitation for Cross-language Information Retrieval. In: The New Review of Information Behaviour Research, 3, pp. 17-35 (2002)
20. Srinivasarao, V.: Mining the Behavior of Users in a Multilingual Information Access Task. Cross Language Information Forum. In: Evaluation of Multilingual and Multi-modal Information Retrieval: 9th Workshop of the Cross-Language Evaluation Forum (2008)
21. TELplus: D3.2, Improving Full-text Search in printed Digital Libraries' Collections through Semantic and Multilingual Functionalities - Technologies Assessment & User Requirements (2009)
22. TELplus: D5.1, Report on User Requirements of the Target Library Services (2008)
23. Tenopir, C.: Use and Users of Electronic Library Resources: An Overview and Analysis of Recent Research Studies (2003)
24. TrebleCLEF: D3.3, Best Practices in System-oriented and User-oriented Multilingual Information Access (2009)
www.trebleclef.eu/getfile.php?id=249
25. TrebleCLEF: D3.2, Workshop on Best Practices for the Development of Multilingual Information Access Systems: the User Perspective (2008)
http://www.trebleclef.eu/getfile.php?id=