Which Log for which Information? Gathering Multilinguality Data from Different Log File Types

Maria Gäde, Vivien Petras, and Juliane Stiller

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

Abstract. In this paper, a comparative analysis of different log file types and their potential for gathering information about user behavior in a multilingual information system is presented. It starts with a discussion of potential questions to be answered in order to form an appropriate view of user needs and requirements in a multilingual information environment and the possibilities of gaining this information from log files. Based on actual examples from the Europeana portal, we compare and contrast different types of log files and the information gleaned from them. We then present the Europeana Clickstream Logger, which logs and gathers extended information on user behavior, and show first examples of the data collection possibilities.

1 Introduction

In order to provide a useful and effective search and retrieval service to users, information systems need to understand who their users are, what they are looking for and how they expect to find it. Adaptive information systems change their interfaces and functionality features according to new user requirements or observed user behavior. User expectations and behavior can be gleaned from extensive questioning (interviews, surveys) or from observation of user interactions with the system (log file analysis). While qualitative methods like interviews can gather a lot of interpretive information (why do user behave that way), they can only survey a small part of the user population and can be misleading depending on the subjective notions of the questioned users. Quantitative data like log files are able to give an overview of all users active in a system. However, one might not be able to explain every observed event. Log files can be helpful in providing insights in trends and common activities as well as problems within an information system. This is valuable in exploratory analyses and can be extended and intensified with qualitative studies after a basis for questions has been established. A deep and careful evaluation of log file data can reveal more insights into user actions and interpretations of pathways through an information system than is visible at first glance, especially if the different log resources are combined with domain and system knowledge.

In this paper, we explore what kind of questions can be answered by analyzing data from log files and what data should be logged in order to understand user behavior. Especially in Europe, information systems increasingly deal with multilingual content and multilingual users alike so that questions of multilingual search capabilities and multilingual representation (starting with how many language versions of a static homepage should be created) need to be answered. We are looking at digital library applications, and in particular the search process, in order to elaborate the system components that are affected by multilingual issues and identify those aspects where data from log files can provide insight into user needs and user expectations. Our use case is Europeana[1], the European digital library, archive and museum portal with the particular charter to serve all European citizens (and their respective languages) effectively.

The remainder of the paper is organized as follows: section 2 briefly introduces Europeana, whereas section 3 reviews related work. Section 4 develops a list of questions concerning multilingual access categorized by their occurrence in the search process. Section 5 reviews common logging approaches and their outcomes for a digital library (http access logs by the web server, the search engine logs and Google Analytics log data), which are also used in Europeana, and their effectiveness in answering the questions laid out in section 4. Section 6 introduces a compound approach for logging developed for the Europeana search portal, the Europeana Clickstream Logger (CSL), which provides improved logging capabilities for answering some of the multilingual questions that could not be answered before. We conclude the paper by showing some examples of the improved logging and discussing some of the benefits of this compound approach.

2 Europeana

The Europeana portal integrates digital objects from Europe's cultural institutions such as libraries, audio-visual archives, museums and archives. The aim is to provide universal access to Europe's cultural heritage in a single access point. The material aggregated in Europeana ranges from videos and images to all kind of texts and even sounds. By the end of 2010, Europeana wants to give access to 10 million objects from all over Europe.

Europeana wants to offer access to cultural and scientific heritage material in a multilingual environment [1]. That means to overcome the language diversity in Europe and to offer access to objects in different languages regardless of the users' language skills.

Generally, multilingual access to content in digital libraries can be achieved on different levels:

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

1. Interface - providing parallel interfaces in different languages,
2. Browsing capabilities - providing parallel browsing structures, e.g. for a subject classification, in different languages,
3. Search capabilities - enabling the user to find documents in a language different from the query language, and
4. Result representation - offering the possibility to translate the results in the users' preferred languages.

Europeana already displays all static interface pages in 26 European languages and provides language-sensitive search suggestions ("People are currently thinking about") from its homepage. Search results can be filtered by the language of the objects in the result list. The EuropeanaConnect project[2] is currently working on developing multilingual search (query translation) and browsing capabilities for up to 10 European languages. The development of these features is accompanied by user studies and evaluation efforts targeted towards providing further insights into user behavior (partly language-dependent) and user requirements for multilingual features within the system. Log file analyses play an important role in gathering this data.

3 Log File Studies in a Multilingual Environment

Various studies have used information from query logs to learn about the search process. Different approaches have been applied to log file analyses [2]. Previous research on query logs from search engines such as Altavista or Excite looked on general statistics, including average number of words per query, time length of query sessions, reformulation and topics [3][4]. Several studies discuss the use of clickthrough data for adequate and continuous analysis of user behavior and search engine optimizing [5]. However, very few studies seem to have targeted multilingual aspects in particular.

The Cross Language Evaluation Forum's (CLEF) iCLEF track 2008 proposed a task searching images in a naturally multilingual database (Flickr). The search logs from the Flickling search interface were analyzed with regard to user search behavior, strategies and success according to language skills [6]. In 2009, the LogCLEF track was launched with the aim to analyze and classify user queries in order to understand search behavior in multilingual contexts and to improve search systems. It consisted of two tasks: LAGI (Log Analysis and Geographic Query Identification) and LADS (Log Analysis for Digital Societies)[7]. To understand the search process and the user interaction with the system, data from The European Library (TEL) and Tumba! was evaluated. Lamm et al. [8] investigate user search performance and interaction with the TEL interface. They discovered different search behaviors of users from different countries. Bosca and Dini [9] present an approach for translation improvement by analyzing user input. The European Digital Library (TEL) and the EDLproject also conducted

[2] http://www.europeanaconnect.eu/

several log file analyses to gather user requirements. Although the full translation of documents is not required, subject keyword translation already seems to be useful for the evaluation of result relevance [10].

The University of Padua analyzed the http traffic logs of The European Library portal finding that the majority of visitors to the portal did not perform any query. The search sessions mostly (77.44%) involved only 1 query, which are short and not reformulated, one of the main challenges regarding language detection [11]. The Max Planck Institute for Informatics in Saarbrücken analyzed server logs to research the user interaction behavior. In particular, they focused on the query and result-click history. They found that the majority of users (84%) leaves the default interface language in English [11]. The evaluation of queries showed that the most frequent searches relate to European place names or subjects.Through a study of library catalog search logs, the CACAO project found that in a library operating in a multicultural context, about 20% of the queries were written in three languages, namely Italian, German and English [12].

Previous log file studies either incorporated multilingual aspects only tangentially or in an exploratory way. This paper attempts to define a set of criteria that could be analyzed within a multilingual information environment and describes a clickstream logging approach to gather extended information on user interaction by using the example of the Europeana Clickstream Logger.

4 Logging Multilingual Information

A standard http access log file tracks every activity performed by a web server, usually page requests. From very few lines of logged data, a lot of information can be gleaned, for example, what kind of questions are asked or how long users stay on a site. Below, a set of criteria is listed, which can provide context for multilingual aspects of user behavior and system features, that could potentially be collected from observing activities logged by the system in some form.

The criteria are grouped in order of possible occurrence of the logged activity in a "classic" search process, where a user approaches a system, launches one or several queries and reviews the results.

User background information (e.g. country of access system; system language, referrer site): Looking at static system parameters like the language of the user's operating system can provide insight into the native or preferred language of the user or other routine language uses, which can be used to infer default languages for the information system's static pages and dynamic features.

Change of interface language: The change of the interface language of an information system is an active intervention of the user, possibly indicating even stronger preferences for favored languages, the incapability of understanding the default language or a stronger engagement with the system (enough interest to adapt the system to one's own uses) than the background information.

Query language: The query language indicates the language a user feels most comfortable searching in. Identifying the language of the query is crucial for

the search process: language-specific query processing and indexing can lead to better search results (both in terms of ranking and in terms of serving the user with documents they can understand). If the query language is not indicated by the user, automatically detecting the query language is necessary. This can be done by analyzing the query terms, the languages of the documents found by the query or by inference from the background information like system or interface language. Sometimes, however, users change their query language in order to adapt to what they assume the language of the documents in the collection to be. Differences between the query language and the system or interface language can indicate such a switch, to which an information system can react accordingly (by suggesting different search strategies, for example). Verifying the query language through user intervention is a possible way of integrating a learning mechanism in the system although this requires careful interaction design so that the usability of the system is not hampered.

Query type (e.g. simple or default query, advanced or fielded query, related items, pre-selected categories for browsing): Different query types can indicate different user preferences or pathways through the system. The most obvious one is when a user selects a language filter for the requested results from the advanced query interface. A related item search can point towards similar documents both thematically but also in the same language. Searchers using pre-selected categories or suggested searches (like the "People are currently thinking about" feature in Europeana) might approach the system differently, not with a fixed information need in mind but serendipitous interest, to which the system can exhibit different behaviors.

Query content (e.g. named entities, dates & numbers, themes or topics): Identifying the content of the query or the object of the user request can have a profound impact on multilingual search quality. Named entities, dates and numbers, for example, should be treated differently in query translation and indexing (sometimes different language versions exist, sometimes they do not). Thematic or topical questions could be expanded or translated with the support of multilingual subject schemes like thesauri or classifications. Context- and language-sensitive recommendation features can suggest new search paths or strategies based on the initial query content.

Query translation: Query translation can be done automatically by the information system or could be offered as an additional service to the user. Logging the languages a query is translated into can provide insights into user preferences and needed feature improvements (e.g. add more languages to the translation feature). Log file data could also help in detecting different language versions of the same query or even adding unknown terms to a dictionary. If the information system offers a manual query translation feature, users can be induced to implicitly improve the translation process by suggesting more precise translations if the original and new query languages are known.

Search results: The language of the search results - if it is identified - cannot only help in query language detection but can also guide the user interaction towards restricting or expanding the result set by language or by offering value-added

services like result translation if results language and other languages (interface, query) differ. Log files can give insights into the language of the particular fields within a metadata record (records themselves can contain multilingual information) that were found by a particular search or the distribution of languages over the whole results set or just the top number of search results.

Result set views: Result set views include any action the searcher performs on a result list, like opening up a particular record for a full view, paging through a result list (if it spans more than one page), saved items or even saved searches. These actions usually indicate a stronger interest of the searcher and lead to valuable information about preferred documents for a search or preferred post-search interaction scenarios.

Result translation (e.g. subjects, key metadata fields, full document): Most information systems do not offer ready-made translation capabilities for search results yet. However, user studies have shown that even translation of small but content-heavy parts of a document (e.g. the title) can support the user in deciding whether a document is relevant for their information need. Logging translation requests, the requested languages as well as views of the translations and subsequent actions can provide valuable insight into feature improvement for future multilingual development (for example, which languages are most often requested, which parts of a document should be offered for translation etc.).

User-generated content (e.g. tags): In certain information systems, users can add content to an existing metadata record; in other systems, users create a wholly new record themselves. The language of user-generated content can play an important role in observing user communities (do users describe objects in the same way) or can even be used for translation or multilingual search (especially if user-generated content in other languages than the original language of the existing metadata is added).

Query reformulation / query expansion / query refinement: After an initial search, users might replace, expand or refine the original query. If they switch languages within a session, this could indicate an attempted translation of queries to increase recall. Query reformulation can indicate language- or context-specific equivalents. Of particular interest are filtering activities, a direct intervention by the user that restricts a search by language or content provider - both indicators of language-dependent user preferences or behavior. A related item search ("more like this") is also a query reformulation since the original query is now changed based on the targeted result object. Both the path through the system and the new results can give new language-specific information.

5 Comparing Log Data

The content of a web server log depends on the server and its settings. The log entries can be presented in different forms. We briefly compare 3 types of logging data gathered at different places: the http access logs from a web server, the search engine logs and the data provided by Google Analytics. The examples

are drawn from the particular case of Europeana, but could be applied to other similar digital libraries as well.

5.1 Transaction Log - Web Server

Each web server keeps log files on the transactions between server and user. Commonly, two files are written. The error log file lists the reported errors and cannot be customized; the access log can be modified to adapt to information needed. Transaction log analysis (TLA) focuses on system performance, information structure or user interaction [13].

For Europeana, Apache web server logs[3] are used. Figure 1 is an example showing a request for an image file from the image cache from a results list for the query "italy".

```
123.123.123.123 - - [11/Mar/2010:09:42:06 +0100]
"GET /cache/image/?uri=http://images.scran.ac.uk/rb/images/
thumb/0098/00980252.jpg&size=BRIEF_DOC&type=IMAGE HTTP/1.0"
200 2843 "http://www.europeana.eu/portal/brief-doc.html?
start=1&view=table&query=italy" "Mozilla/5.0 (Windows; U;
Windows NT 5.1; it; rv:1.9.2) Gecko/20100115 Firefox/3.6
(.NET CLR 3.5.30729)"
```

Figure 1. Example line from the Europeana Apache log

The IP-address and the system language indicate the country of origin or the language (Italian) of the user. Keeping in mind that proxy use or default settings might be misleading, final conclusions about the user's language skills cannot be drawn. The IP-address, however, also implies the reach of the service and usage in different countries. The user agent requesting pages from the server is also stored in the log files. This can be used to differentiate between search engine bots crawling the site and real user sessions.

From the logs, one can also observe the interaction of the user with the system. In this example, the page requested is introduced by the "GET"-request method which is followed by the desired file, an image. Subsequent to the http status code and the number of bytes transferred is the referrer URL which shows the page visited before requesting the present page. In this example, the referrer URL shows the inital query "italy". The query can be analyzed in terms of content, language and type (advanced or simple).

Generally, it is possible to reconstruct individual user sessions, assuming the same IP address is used, and identify query reformulations and search results clicked. This requires an analysis of several log entries and the relationships between them. Multilingual information can be gleaned from the IP address, system language indicators, and the query (although only after one would go through a language detection process). However, the interface language cannot be logged this way nor the search results.

[3] http://httpd.apache.org/

The use of Web 2.0 techniques for dynamic functionality, which is prevalent on the web today also causes some problems. These techniques make use of client-side technologies that do not necessarily communicate with the server to process user requests. For example, actions executed through javascript (AJAX) and Flash are often not logged in the transaction logs. For every access request, including all the style (CSS), javascript and image files are logged as well. Considering that an average html web page consist of at least 10 requests, the crucial information is easily obscured. In addition, the application state (what happens on the server between when the request arrives and the response sent back to the user) is not saved either.

5.2 Query Log - Search Engine

The query log of the search engine might only log a subset of actions the user performs on the system, namely the searches. These can also be reconstructed from the web server access logs as shown in the previous example. In the Europeana portal, the search engine used is Solr[4]. Europeana uses a stateless design, where queries are transformed into Solr syntax and then sent from the front-end servers. Because of this, session and user information is lost and cannot be reconstructed, even though the query content and query fields are contained.

The Solr search engine logs - at least for Europeana - can not add any new information that the web server log does not contain already.

5.3 Page Tagging - Google Analytics

Page tagging (adding JavaScript to a web page to gather information about access) generates detailed statistics about the use of a Web site without web server access. A prominent example, which is also used for Europeana, is Google Analytics [5]. It offers general data analysis but also provides customized reports on access statistics and strategies implemented to direct traffic streams.

Dynamic tracking allows to identify and classify the site visitors with respect to individual regions, states/provinces and cities. It is very simple to generate reports showing distributions by language (the user system language) and country or referrer sites. However, for search-based systems like Europeana, Google Analytics cannot compile information regarding interface languages (application states) or the search process (queries, results) itself.

5.4 Log File Trade-offs

The analysis of log files comes with 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. Furthermore, it is possible to hide the true location by using proxies. The IP address of users or

[4] http://lucene.apache.org/solr/
[5] http://www.google.com/analytics/

any other background information can only give indications about the language preferences of users but might not be reliable in every case.

Google Analytics uses cookies in order to track visitor activity. Users who delete their cookies will still be tracked by Google Analytics, but they will be identified as a new visitor to the site leading to incorrect session results. Using host names to group or locate users geographically can also be misleading. It is difficult to interpret user intentions from queries and search activities alone. Log entries are limited to transactions of the web server and do not reveal backgrounds or preferences [14]. They do not give any clues about the reasons behind certain actions the users executed [15]. Motives for certain actions can be very complex and dependent on the users' perception on the performance of the system.

Common log file entries are general and therefore contain limited information concerning multilingual issues. The logs cannot be used for statements about the frequency of interface language change or relation between interface language and query language. The reconstruction of individual user sessions is difficult and time-consuming. The calculation of the distribution of query or result languages cannot be easily done. Information about application states is also obscured. Information about the user and the results presented or certain client-side technology-based activities is not retrievable. For example, whether the user is logged in, how many results are returned, and statistics from which languages the results came from.

6 Europeana Clickstream Logging

Clickstream logging is a logging approach, which enables to mine complex data in order to analyze user behavior. 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. Clickstream logging allows to track application state changes and therefore traces user behavior in a way that a traditional http transaction log is unable to. In this section, we describe a clickstream logger developed for Europeana, which pays particular attention to multilingual aspects of the portal and its users. Logging data from the Europeana Clickstream Logger (CSL) allows to reconstruct user sessions easily and provides a more complete picture on language issues.

6.1 Logging Multilingual Information in Europeana

For the Europeana clickstream logs, six different activity types or states with a particular focus on multilingual access aspects are logged. These actions indicate a stream of user activities which can be categorized as follows:

Interface language-specific actions: The interface language or the change of the interface language is logged for each transaction. This would otherwise not appear in the http access log.

Search-related actions: All search-related activities including information about query terms, result numbers and distribution of results by language and country are logged. Filtering (e.g restricting by language, provider, date) and related searches (from an initial result list) are also logged.

Browse-related actions: For the http transaction log, a browsing activity (e.g. clicking on one of the images cycling across the Europeana homepage that are suggested as search entries) is the same as a search via term entry: in both cases requests are sent to the search engine. However, from a user interaction perspective, browsing and searching might point to different user intentions. The clickstream logger logs all browsing activities and their initial starting points (e.g. did the search originate from the cycling images, the suggested searches, the time line, the saved searches links or saved user tags).

Ajax-related actions: Client-side activities (post actions) that might not appear in the transaction log, which involve interactive features of Europeana, are logged here. This includes saving or removing social tags, searches or objects.

Navigation-related actions: User paths through search results are logged here, for example, when a user moves away from Europeana by following a link from a detailed results page to the original object or when the user returns to the results list.

User management-related actions: This involves actions connected to user account creation, logging in and out and changing passwords.

Additionally, errors and requests on static pages are logged.

6.2 Europeana Clickstream Logger (CSL) Data

The comparison of a web server access log entry and the associated clickstream log shows that the clickstream approach includes much more information especially on multilingual usage. Figures 2 and 3 represent a single user action: restricting a result set by language ("fr" - french) following an initial one-word query ("treasure").

The http access log contains the query and the language facet:

```
http://www.europeana.eu/portal/brief-doc.html?query=treasure
&qf=LANGUAGE:fr&view=table
```

Figure 2. Abbreviated example line from the Europeana http access log

The clickstream log shows a logged-in user (ID: 12345), who performed a search ("treasure") in the simple search interface, refined it by the language facet "fr" and now looks at the general result list. Also logged is the position in the result set (page 1), the number of returned objects (3), the result distribution by the top 5 languages and by country of the provider (top 5) and the interface language (Norwegian).

```
[action=BRIEF_RESULT, view=brief-doc-window, query=treasure,
queryType=simple, queryConstraints="LANGUAGE:"fr"", page=1,
numFound=3, langFacet=en (1,126), es (28), fr (3), mul (3),
de (2), countryFacet=france (3), userId=12345, lang=NO,
req=http://www.europeana.eu/portal/brief-doc.html?
query=treasure&qf=LANGUAGE:fr&view=table
```

Figure 3. Abbreviated example from the Europenana CSL with action: language filter

The URL of the requested page (where the query could in this case also be reconstructed from), the referrer page, the date, time and IP address and system information for the user are noted in both log variants and are not shown in the examples.

Figure 4 shows the clickstream log entry for an interface language change - an action that cannot be logged with the http transaction log.

```
[action=LANGUAGE_CHANGE, oldLang=EN, userId=, lang=FR,
req=http://europeana.eu:80/portal/aboutus.html?lang=fr
```

Figure 4. Abbreviated example from the Europeana CSL with action: language change

7 Conclusion

In this paper, we compare logging methods with respect to their ability to answer questions about user behavior and user intentions in a multilingual information environment. The Europeana Clickstream Logger presented above alleviates some of the information gaps that occur in traditional http access logs. The Europeana Clickstream Logger logs actions that are impossible to reconstruct from transaction logs (e.g. language changes or differences between browse and search requests), but also logs actions and their results in a way that is simpler and less error-prone to analyze. It therefore makes knowledge explicit that was previously obscured by parameters that could only be interpreted by aggregating data from different resources or that require extensive system expertise. This detailed logging enables us to gather a more complete view of the multilingual issues facing a system and to glean user intentions more explicitly. Other important multilingual aspects that cannot be studied from log files - the automatic detection of the query language, for example - still need to be tackled. However, with the clickstream logging in place, we can track how changes from the baseline (e.g. rerouting requests by language) will affect user behavior giving us an important evaluation tool for further interaction research.

Acknowledgment Work in this paper was partially funded by the eContentplus project EuropeanaConnect (ECP-2008-DILI-528001). We would like to thank Sjoerd Siebinga from the Europeana Foundation for implementing the Europeana Clickstream Logger, his continuous support and great patience in answering our questions.

References

1. Purday, J.: Think culture: Europeana.eu from concept to construction. The Electronic Library **6** (2009) 919–937
2. Jansen, B.J.: Search log analysis: What it is, what's been done, how to do it. Library & Information Science Research **28** (2006) 407–432
3. Silverstein, C., Marais, H., Henzinger, M., Moricz, M.: Analysis of a very large web search engine query log. SIGIR Forum **33** (1999) 6–12
4. Jansen, B.J., Spink, A., Bateman, J., Saracevic, T.: Real life information retrieval: a study of user queries on the web. SIGIR Forum **32** (1998) 5–17
5. 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, New York, NY, USA, ACM (2002) 133–142
6. Gonzalo, J., Clough, P., Karlgren, J.: Overview of iclef 2008: Search log analysis for multilingual image retrieval. In Peters, C., Deselaers, T., Ferro, N., Gonzalo, J., Jones, G.J.F., Kurimo, M., Mandl, T., Peñas, A., Petras, V.(eds) CLEF 2007. LNCS, vol. 5706, pp. 227–235. Springer, Heidelberg (2008)
7. Mandl, T., Agosti, M., Nunzio Di, G., Yeh, E., Mani, I., Doran, C., Schulz, J.M.: Logclef 2009: the clef 2009 multilingual logfile analysis track overview. In: Working Notes of the Cross Language Evaluation Forum (CLEF). (2009)
8. Lamm, K., Mandl, T., Kölle, R.: Search path visualization and session performance evaluation with log files from the european library. In: Working Notes of the Cross Language Evaluation Forum (CLEF). (2009)
9. Bosca, A., L. Dini: Cacao project at the logclef track. In: Working Notes of the Cross Language Evaluation Forum (CLEF). (2009)
10. Mane, L.: D3.2 improving full-text search in printed digital libraries' collections through semantic and multilingual functionalities - technologies assessment & user requirements. Technical report, TELplus (2009)
11. Angelaki, G.: M1.4 interim report on usability developments in the european library. Technical report, EDLproject (2007)
12. Trojahn, C., Siciliano, L.: D7.4 user requirements for advanced features. Technical report, CACAO project (2009)
13. Jansen, B.J.: The methodology of search log analysis. In Jansen, B.J., Spink, A., Taksa, I., eds.: Handbook of Research on Web Log Analysis. Information Science Reference (2009) 100–123
14. Booth, D.: A review of methodologies for analyzing websites. In Jansen, B.J., Spink, A., Taksa, I., eds.: Handbook of Research on Web Log Analysis. Information Science Reference (2009) 143–164
15. Hulshof, C.: Log file analysis. In: Encyclopedia of Social Measurement. Volume 2. Elsevier (2004) 577–583