

# LogCLEF 2010: the CLEF 2010 Multilingual Logfile Analysis Track Overview

Thomas Mandl<sup>1</sup>, Giorgio Maria Di Nunzio<sup>2</sup>,  
Julia Maria Schulz<sup>1</sup>

<sup>1</sup> Information Science, University of Hildesheim, Germany  
 [{mandl,schulzju}@uni-hildesheim.de](mailto:{mandl,schulzju}@uni-hildesheim.de)

<sup>2</sup> Department of Information Engineering, University of Padua, Italy  
[dinunzio@dei.unipd.it](mailto:dinunzio@dei.unipd.it)

**Abstract.** Log data constitutes a relevant aspect in the evaluation process of multilingual search services. Activity logs allow to study the usage of search engines and to better adapt them to the needs of their users. The study of multilingual log analysis was promoted by the Cross Language Evaluation Forum (CLEF). For the second time, the track LogCLEF was conducted. As in 2009, large log files were obtained from information providers. One log covers 30 months of activities on the website of The European Library (TEL) and the second log shows user activities of users on the German EduServer. Seven groups explored the data using a variety of approaches. They analyzed languages of queries, activities within sessions and success of searches. The data for the track, the evaluation methodology and results are presented and discussed.

## 1 Introduction

Web Search Engines deal with the representation, storage, organization of, and access to information items which are essentially Web pages. The characterization of the user information need is not simple, and this problem can roughly be divided into three aspects: how the user poses his request to the search engine, how the user interacts with the search engine, and how the search engine organizes the results.

Log data constitute a relevant aspect in the evaluation process of the quality of a search engine and the quality of a multilingual search service; log data can be used to study the usage of a search engine, and to better adapt it to the objectives the users were expecting to reach [1]. The log data can be used to study the usage of a specific application, and to better adapt it to the objectives the users were expecting to reach. The analysis of transaction logs for studying automatic information access systems has a long history, much earlier than the World WideWeb as we know it today.

The interest in multilingual log analysis was promoted by the Cross Language Evaluation Forum (CLEF)<sup>1</sup> in the track LogCLEF<sup>2</sup> which was conducted for the first

---

<sup>1</sup> <http://www.clef-campaign.org/>

<sup>2</sup> <http://www.uni-hildesheim.de/logclef/>

time in 2009 [2] and for the second time in 2010. LogCLEF is an evaluation initiative for the analysis of queries and other logged activities as expression of user behavior. The main goal of LogCLEF is the analysis and classification of queries in order to understand search behavior in multilingual contexts and ultimately to improve search systems. Another important long-term aim is to stimulate research on user behavior in multilingual environments and promote standard evaluation collections of log data.

LogCLEF differs from other evaluation tracks since its goal is not the production of a gold standard for a specific task, but to create a forum for the creative exploration of user behavior based on logs.

The data sets used in 2010 were activity logs derived from the The European Library (TEL) Web site<sup>3</sup> and the German EduServer<sup>4</sup> -- Deutscher Bildungsserver (DBS) -- maintained by the DIPF, the Leibniz Institute for Educational Research and Educational Information. The task definition, the data for the track, the evaluation methodology and some results of submitted experiments are presented in this overview paper.

## 2 Task Definition

The main question behind the task definition comes from search service providers who wonder how they can improve their services. Ultimately, researchers need to better understand user behavior in order to reach that high level goal. Two objectives of the analysis of the logs are proposed, one for each set of log files:

**TEL:** Investigate language of queries with respect to successful search sessions. A successful search could be defined as one of the following actions listed in the right hand box of the TEL interface when an item of the result clicked is listed.

- + Services:
  - Availability at the library,
  - Link to other services,
  - collection homepage
- + Options:
  - Save in favorites,
  - Send by email.

Potential research issues for TEL:

1. language identification for the queries
2. initial language vs country IP address
3. subsequent languages used on same search
4. country of the library vs language of the query vs language of the interface

**DBS:** The objective of the analysis of the DBS logs is the exploration of the relation between query and viewed content. The analysis can explore formal issues of

---

<sup>3</sup> <http://www.theeuropeanlibrary.org/>

<sup>4</sup> <http://www.eduserver.de/>

the query and content as well as the distribution of words within both.

Potential research issues for DBS:

1. Are query terms related to the content viewed and/or paths taken within the system?
2. Can query modifications be explained by the content viewed?
3. Develop metrics to identify successful searches

### 3 Data Description

The data for LogCLEF 2010 collection consists of two large log files from information providers:

- The European Library (TEL) logs: As in 2009, a large log of activities from The European Library are provided. This service provides access to several national libraries of Europe. Users and content come from many languages.
- German EduServer (Deutscher Bildungsserver, DBS) logs: The "Deutscher Bildungsserver" is a quality controlled internet directory for educational resources. A raw server log representing three months of activities on the portal is made available. The size of all files is 5 GB.

The following table gives an overview on the log resources which were been made available at CLEF over the last years.

**Table 1:** Log file resources at CLEF

Year	Origin	Size	Type
2007	MSN	800.000 queries	Query log
2009	Tumba!	350.000 queries	Query log
2009	TEL	1.870.000 records	Query and activity log
2010	TEL	2.600.000 records	Query and activity log
2010	TEL	1.5 GB (zipped)	Web server log
2010	DIPF.de	5 GB	Web server log

#### 3.1 TEL

TEL is a free service that offers access to the resources of 48 national libraries of Europe in 35 languages, it aims to provide a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage. Resources can be both digital (e.g. books, posters, maps, sound recordings, videos) and bibliographical and the quality and reliability of the documents are guaranteed by the 48 collaborating national libraries of Europe.

The data used for this task are search logs and Web server logs of The European Library portal.

### 3.1.1 TEL Action Logs

Search logs are usually named “action logs” in the context of TEL activities. In TEL portal’s home page, a user can initiate a simple keyword search with a default predefined collection list presenting catalogues from national libraries. From the same page, a user may perform an advanced search with Boolean operators and/or limit search to specific fields like author, language, and ISBN. It is also possible to change the searched collection by checking the theme categories below the search box. After the search button is clicked, the result page appears, where results are classified by collections and the results of the top collection in the list are presented with brief descriptions. Subsequently, a user may choose to see result lists of other collections or move to the next page of records of current collection’s results. While viewing a result list page a user may also click on a specific record to see detailed information about the specific record. Additional services may be available according to the record selected.

All these type of actions and choices are logged and stored by TEL in a relational table, where each record represents a user action [3]. The most significant columns of the table are:

- A numeric id, for identifying registered users or “guest” otherwise;
- User’s IP address;
- An automatically generated alphanumeric, identifying sequential actions of the same user (sessions) ;
- Query contents;
- Name of the action that a user performed;
- The corresponding collection’s alphanumeric id;
- Date and time of the action’s occurrence.

**Table 3:** Examples from the TELlog (date has been deleted for readability)

id;userid;userip;sesid;lang;query;action;colid;nrecords;recordposition;sboxid;objurl;date 892989;guest;62.121.xxx.xxx;btprfui7keanuelu0nanhte5j0;en;("plastics mould");view_brief;a0037;31;;; 893209;guest;213.149.xxx.xxx;o270cev7upbblmqja30rdeo3p4;en;("penser leurope");search_sim;;0;-;; 893261;guest;194.171.xxx.xxx;null;en;("magna carta");search_url;;0;-;; 893487;guest;81.179.xxx.xxx;9rrtrdp2kqrd706pha470486;en;("spengemann");view_brief;a0067;1;-;; 893488;guest;81.179.xxx.xxx;9rrtrdp2kqrd706pha470486;en;("spengemann");view_brief;a0000;0;-;; 893533;guest;85.192.xxx.xxx;ckujekqff2et6r9p27h8r89le6;fr;("egypt france britain");search_sim;;0;-;;
---

Action logs distributed to the participants of the task cover the period from January 2007 until June 2008 and from January 2009 until December 2009. The log file contains user activities and queries entered at the search site of TEL. Examples for entries in the log file are shown in Table 3.

### 3.1.2 TEL Web Server Logs

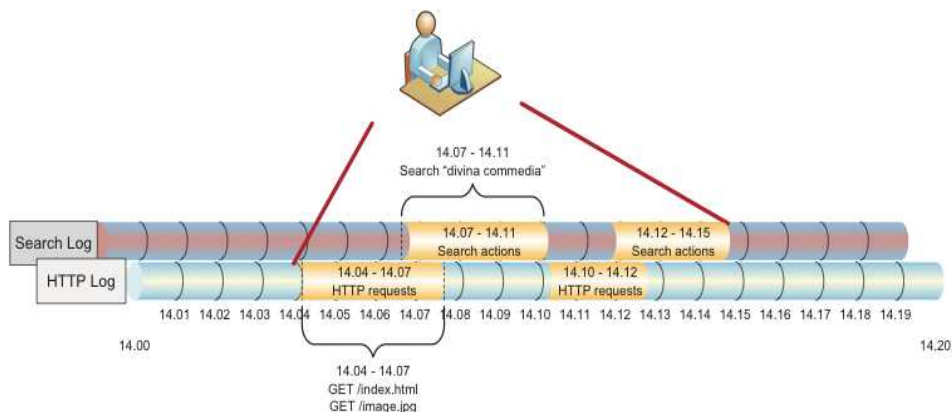
The Web server log files of TEL cover the same period of the first data set of action logs, from January 2007 until June 2008. These log files are saved in 18 text files

(zipped), one for each month of the year, and each record contains the following fields:

- date: year-month-day.
- time: hour:minute:second.
- HTTP method: for example GET, HEAD, POST, etc.
- URI stem: the path of the requested file.
- URI query: the string of the query in the URL, if any.
- IP address: the address of the client, (obfuscated, e.g. 127.0).
- User agent: the user agent of the client.
- Cookie: the cookie sent to/by the client.
- Referrer: the URL of the resource which linked the client to TEL.

The Cookie field is divided into subfields by semi-colons “;”. Some of the subfields are (some of them are ignored for this task):

- cTargets: the identifiers of the collections selected by the user;
- TELSESSID: the identifier of the session. It is the same identifier recorded in the action logs under the name “sesid”. This is an important field to cross-analyze action logs to Web server logs. Figure 1, shows an example of how a user session may be stored in the two different logs.



**Figure 1.** An example of how actions of the user are recorded in the two log data sets. Searching and browsing activities of the same computer are uniquely identified by the TELSESSID field which is stored both in the action logs and in the cookie field in the HTTP logs.

### 3.2 EduServer

The quality controlled "Deutscher Bildungsserver" is a clearinghouse for educational resources on the Web<sup>5</sup>. It also contains content provided by the DIPF as well as

<sup>5</sup> [http://www.bildungsserver.de/start\\_e.html](http://www.bildungsserver.de/start_e.html)

descriptions and reviews on Web sites on education. The Internet resources (web sites) are described, checked for their quality, manually indexed and classified. The logs were collected in the time between September and November of 2009. The logs are server logs in standards format in which the searches and the results viewed can be observed. An excerpt is shown in table 2. The logs have been anonymized by partially obscuring the IP addresses of users.

The two upper levels of server names or IP addresses have been hashed. This allows the reconstruction of sessions within the data. Note that accesses by search engine bots are still within the logs. The logs allow to observe two types of user queries:

- queries in search engines (in the referrer when DBS files were found using a search engine)
- queries within the DBS (see query parameters in metasuche/qsuche)

**Table 2:** Examples from the DBS log (some data has been modified for readability)

```
f64.alicedsl.de - - [09/Nov/2009:00:23:09 +0100] "GET /zeigen.html?seite=5892
HTTP/1.1" 200 22436 http://www.bildungsserver.de/zeigen.html?seite=2521
"Mozilla/5.0 (Windows; U; Windows NT 5.1; en-US; rv:1.9.1.5) Gecko/20091102
Firefox/3.5.5"
80d.superkabel.de - - [09/Nov/2009:00:26:28 +0100] "GET
/db/fwulesen.html?Id=200006289 HTTP/1.1" 200 16301
http://www.google.de/search?hl=de&source=hp&q=+landes+filmstelle&btnG=Google-
Suche&meta=&aq=f&oq =&fp=6013614429992176 "Mozilla/5.0 (Windows; U; Windows NT
6.0; de; rv:1.9.1.4) Gecko/20091016 Firefox/3.5.4 (.NET CLR 3.5.30729)"
937.googlebot.com - - [09/Nov/2009:00:27:09 +0100] "GET
/db/ffach2.html?fach=2&Rnum=12&Snum=3 HTTP/1.1" 200 16019 - "Mozilla/5.0
(compatible; Googlebot/2.1; +http://www.google.com/bot.html)" 5bd.ono.com - -
[09/Nov/2009:00:30:46 +0100] "GET /db/mlesen.html?Id=42021 HTTP/1.1" 200 180746
- "Java/1.6.0_13"
8f4.primacom.net - - [09/Nov/2009:00:30:45 +0100] "GET /zeigen.html?seite=771
HTTP/1.1" 200 45871
http://www.bildungsserver.de/metasuche/qsuche.html?feldinhalt1=aktive+medienarbe
it&bool1=AND&finden=finden&searchall=
ja&datenbanken%5B%5D=db_s_seiten&DBS=1&art=einfach "Mozilla/5.0 (Windows; U;
Windows NT 6.0; de; rv:1.9.1.4) Gecko/20091016 Firefox/3.5.4 (.NET CLR
3.5.30729)"
```

The logs also allow so observe the browsing behavior within the DBS server structure. The following pages are of most interest:

- the descriptions of the educational web sites within DBS (mlesen)
- thematic lists of educational web sites (zeigen, anzeigen, fachlist, listen)
- a newspaper documentation on articles about education (zeitdok)

The logs allow to access two types of content and compare them to the queries.

- the descriptions of the educational web sites within DBS
- the content of the educational web sites themselves (which might have changed since the logs have been collected) in those cases where the user might have accessed them

## 4 Participants and Results

The two following sections shows the participants of LogCLEF 2010 and presents some results. For more detailed results, the reader is referred to the papers by the participants which describe the approaches and findings in more detail.

### 4.1 Participants

As shown in Table 4, a total of 7 groups submitted results for LogCLEF. Of the 15 registered groups, only less than 50% managed to obtain results. The results of the participating groups are reported in the following section and elaborated in the papers of the participants. All groups analyzed the TEL logs and none worked with the DBS logs. This might be due to the nature of a raw web server log which requires much pre-processing. LogCLEF could not provide a pre-processed version due to the lack of funding for LogCLEF.

**Table 4.** LogCLEF 2010 participants

<b>Participant</b>	<b>Institution</b>	<b>Country</b>
DAEDALUS	Universidad Politécnica de Madrid & Universidad Carlos III de Madrid	Spain
SINAI	University of Jaen	Spain
TCD-DCU	Trinity College Dublin & Dublin City University	Ireland
NII	National Institute of Informatics & other institutions	Japan
Info Foraging Lab	Radboud University Nijmegen & Maastricht University	The Netherlands
Info Science	Humboldt University Berlin	Germany
CELI s.r.l	CELI Research, Torino	Italy

### 4.2 Results

A large variety of approaches was taken to analyze the TEL log files. This can be considered as a success of the open definition of the task which encouraged creative exploration of the data.

Two groups contrasted user behavior at a quality search service like TEL to common Web search behavior. A group from The Netherlands under the leadership of the University of Nijmegen contrasted frequent queries and number of queries per session in the TEL log with data from an MSN log [8]. Verberne et al. also created a network of actions within TEL visualizing the frequency of actions as well as transition probabilities. It can be observed that view actions are more frequent than search actions and that the full view of a result is selected more often than the brief view.

The NII from Tokio also compared the TEL logs to web search logs and theories developed by exploiting web search [9]. Takaku et al. analyzed the two TEL logs separately and observed few differences between the two time spans. They also integrated the length of a session into their work. Generally a high correlation between the number of actions and the length can be seen, but there are many exceptions which might be interesting for further exploration. Takaku et al. extracted the ranks of the documents clicked by the users and compared the result from Web search experiments.

The DAEDALUS group formally defined success for sessions and queries. They calculated that only 6% of the queries and 10% of the sessions could be labeled as successful.

Three groups focused on language issues. The SINAI group showed that most of the sessions are in English. They also conclude that 50.000 of the sessions exhibit only one action. More than 80% of the sessions have 10 or fewer actions.

The CELI research institute tried to identify the language of search queries [4]. They manually labeled 100 queries and their system managed to correctly identify over 70%. CELI concludes that the integration of named entity recognition needs is necessary.

The difficulties of language identification were elaborated by a group from Berlin [10]. They manually checked 510 queries for their detailed analysis. It showed that over 50% of the queries consisted of only a named entity and an additional 8% included named entities together with another term. Obviously, this complicates language identification and even in the manual analysis 38% of the queries could not be classified as being of one language. Another 31% were English. Stiller et al. also showed that the interface language selected and the origin of the IP are only weak indicators for the query language in their sub set [10].

A group from Dublin [6] also conducted research on the interface language and the origin of the user. Leveling et al. related these factors to the collection selected by the user and managed to develop a scoring function which can rerank the result documents in a way that improves the result quality for the user based on the clicks as observed in the log file. Leveling et al. managed to analyze the content of the queries in order to develop query performance estimators. They implemented IDF and clarity score [6].

## **5 Conclusion and Future Work**

Studies on log files are limited by privacy issues. For the first time, LogCLEF provided evaluation resources for log file analysis which can be used for comparative system evaluation. The second year of LogCLEF obtained more attention by participants. It is intended to encourage and facilitate the exchange of resources and tools generated within the participation at LogCLEF.

In the future, log analysis should be the basis for other evaluation tasks. Logs can show how users behave and what they need. One example could be the selection of topics for retrieval evaluation or for questions answering systems [10].



## Acknowledgments

The organization of LogCLEF was mainly volunteer work. We want to thank The European Library (TEL) and DIPF, the Leibniz Institute for Educational Research and Educational Information, Frankfurt, Germany for providing the log files.

At the University of Padua, the work has been partially supported by TELplus Targeted Project for digital libraries, as part of the eContent*plus* Program of the European Commission (Contract ECP-2006-DILI-510003).

## References

1. Jansen, B.; Spink, A. & Taksa, I. (eds.) Handbook of Research on Web Log Analysis. Idea Group Reference: Hershey et al. 2009
2. Mandl, T; Agosti, M.; Di Nunzio, G.; Yeh, A., Mani, I.; Doran, C. & Schulz, J. LogCLEF 2009: the CLEF 2009 Cross-Language Logfile Analysis Track Overview. In: Multilingual Information Access Evaluation I: Text Retrieval Experiments: Proc. 10<sup>th</sup> Workshop of the Cross-Language Evaluation Forum, CLEF 2009, Corfu, Greece. Revised Selected Papers. Berlin et al.: Springer [LNCS 6241] Preprint in Working Notes: [http://www.clef-campaign.org/2009/working\\_notes/LogCLEF-2009-Overview-Working-Notes-2009-09-14.pdf](http://www.clef-campaign.org/2009/working_notes/LogCLEF-2009-Overview-Working-Notes-2009-09-14.pdf)
3. Di Nunzio, G.M.: LogCLEF 2009 2009/03/02 v 1.0 Description of the The European Library (TEL) Search Action Log Files. [http://www.uni-hildesheim.de/logclef/Daten/LogCLEF2009\\_file\\_description.pdf](http://www.uni-hildesheim.de/logclef/Daten/LogCLEF2009_file_description.pdf) 2009
4. Bosca, A. & Dini, L.: Language Identification Strategies for Cross Language Information Retrieval. *In this volume* (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
5. Perea-Ortega, J.; Montejo Ráez, A.; Garcia Cumbreiras, M. & Ureña-López, L.A.. SINAI at LogCLEF 2010 *In this volume*. (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
6. Leveling, J.; Ghorab, M.R.; Magdy, W.; Jones, G. & Wade, V.: DCU-TCD@LogCLEF 2010: Re-ranking Document Collections and Query Performance Estimation. *In this volume*. (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
7. Lana-Serrano, S.; Villena-Román, J. & González-Cristóbal, J-C. DAEDALUS at LogCLEF 2010: Analyzing the Success of Search Queries. *In this volume* (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
8. Verberne, S; Hinne, M.; van der Heijden, M; Hoenkamp, E.; Kraaij, W. & van der Weide, T. How does the Library Searcher behave? *In this volume* (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
9. Takaku, M.; Egusa, Y.; Saito, H.; Kando, N.; Teraki, H.; Miwa, M.. CRES at LogCLEF 2010: Towards Understanding the User Behaviors through an Analysis of Search Sessions, Search Units and Click Ranks. *In this volume* (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
10. Stiller, J.; Gaede, M. & Petras V. Ambiguity of Queries and the Challenges for Query Language Detection. *In this volume* (LogCLEF 2010 Working Notes, <http://clef2010.org/>)
11. Sutcliffe, R.; Kruschwitz, U. & Mandl, T. Web Logs and Question Answering. In: Proc. Web Logs and Question Answering (WLQA2010) Workshop at the Seventh International Conference on Language Resources and Evaluation (LREC) Malta, 22nd May. S. 1-7. <http://www.csis.ul.ie/wlqa2010/proceedings.htm>