



WP1 - Task3 – Usage and Usability

M1.4. Interim Report on Usability Developments in The European Library

Report Georgia Angelaki, The European Library Office,
Author: EDLproject assistant

Report Eric van der Meulen, The European Library Office,
Contributors: Technical Project Manager
Olaf Janssen, The European Library Office, project manager
Sjoerd Siebinga, The European Library Office,
Technical Project Manager

The EDLproject is funded under
the European Commission *eContentplus* Programme.

Executive Summary

The current report presents the interim developments regarding work package 1 (Developing The European Library Network) and in particular Task 3 which aims at enhancing usage and usability of The European Library. This is a workpackage of EDLproject, a project funded by eContentplus from September 2006 to February 2008 and run by The European Library.

The task involves developments in three areas of action: the development of a registration and authentication procedure, the analysis of the log files of The European Library and the promotion of OAI-PMH to the partner national libraries.

Via the analysis of the log files The European Library aims to shed more light into its' user requirements and their information needs. It is a way of deducing who the users are and how they go about interacting with the portal. It is a useful way for identifying particular bottlenecks that the users encounter in the portal so as to improve the interface and the services offered. At a later stage it is hoped that the analysis of the log files will lead to the development of personalized services for the users.

OAI-PMH is an innovative protocol for sharing metadata. It is a solution that is expected to bring a major break-through in the usability of the portal. By harvesting the metadata from the distributed resources of the partner national libraries in one central repository it achieves integrated search and retrieval of materials. Besides this major development the central repository facilitates the implementation of a number of tools and functions that can transform the user experience in The European Library.

The development of a registration system is an additional action that supports the other two: it helps better identify users for the analysis of the log files as well as, it helps, in parallel with the development of the central index, to send to registered users personalised information.

About this report	5
Part A: The registration function	6
Part B: Log files Analysis Project	8
Introduction	8
Issues	9
Project Description	11
About the log files analysis project	11
Background	11
What are the log files?	11
The European Library log files and background work	12
Additional information: The registered users database	13
Individual goals	14
Deliverables	14
Description of tasks and responsibilities	14
Communication	15
Dissemination	15
Project Implementation	15
The University of Padua	15
Individual goals	15
Methodology of parsing	16
Findings	16
General information:	16
Requests	16
Operating systems	17
Reconstruction of sessions	17
Analysis of national provenance of users	20
Human users and crawlers	20
Recurrent users and bouncers	20
Registered users	21
Referrer and country	21
TEL Action logs	22
User study	23
Recommendations	23
Max Planck Institute	24
Individual goals-MPII	24
Methodology	24
Findings	25
General information	25
Search sessions and queries	25
Query formulation:	25
Keywords	26
Query term correlations	26
Users' navigational patterns	26
Collection selection	28

Query-dependent selection of collections	29
Observations about the analysis	29
Conclusions and general recommendations for future work .	30
Future actions regarding the analysis of The European Library log files.....	32
Part C: PROMOTING OAI-PMH	33
Aims of the task.....	33
What is OAI-PMH	33
Why is The European Library promoting OAI-PMH?	33
Benefits of implementing OAI-PMH	35
Issues	37
Actions for the promotion of OAI-PMH	37
The European Library OAI -PMH java installer	39
Further assistance and promotion.....	40
Overall Outcomes:	40
Further steps regarding the promotion of OAI-PMH	41
Part D: Annexes	43
Annex 1: Description of the TEL Action Log Files Schema	43
Annex 2: Final Reports from the University of Padua and Max Planck Institute.....	44

About this report

The current report provides an account of the actions accomplished within the EDLproject Work Package 1 (Extending The European Library) Task 3: Developing Usage and Usability of The European Library portal. Other work package tasks involve: the design and implementation of plans for the following national libraries to become full members of The European Library: Belgium, Lichtenstein, Greece, Spain, Ireland, Iceland, Sweden, Luxembourg and Norway. Task 4 involves the development of a metadata registry for The European Library and the investigation of metadata interoperability among diverse cultural heritage domains like museums, archives and libraries.

WP1 T3 extends from September 2006 (start of EDLproject) until the end of EDLproject (February 2007). The author of the current report is the overall coordinator for any actions within this Task.

The task is divided into two reporting periods. The current report is the interim report covering actions from September 2006 to August 2007. The final report on usage and usability developments is due in February 2007. The project has three separate strands: the development of an opt-in registration system and an authentication procedure for the users of The European Library, the analysis of the log files of The European Library which was realised as external project with the help of two research institutes (the University of Padua and Max Planck Institute for Informatics) and the promotion of OAI-PMH protocol for Metadata Harvesting. These areas are connected in the sense that they contribute to the development of The European Library portal usability.

More in particular, the analysis of the log files aims to shed more light into who the users of The European Library are in order to use this information for the design of personalised services. The adoption of OAI-PMH promises to revolutionise the back-end architecture of The European Library in a way that the portal can host much more user-friendly functionalities. The registration system is a specific functionality that serves a dual role: it helps gather more information about the users, which can then be associated and analysed with the log files of these users and shed more light into user profiles. On the other hand it is a necessary feature which, together with the expansion of the Central Index via the OAI-PMH will enable providing personalised services to the users.

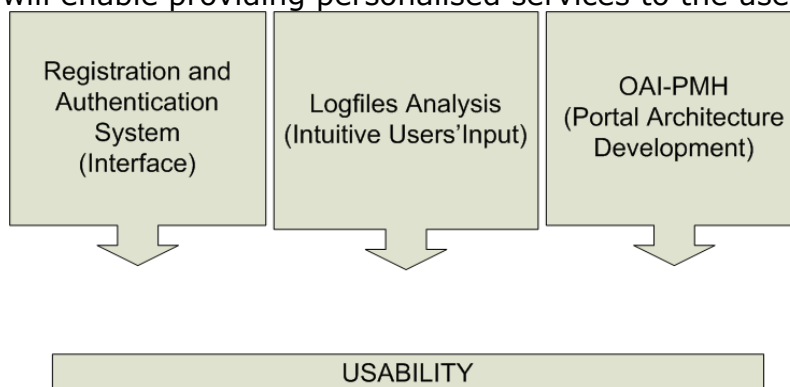


Figure 1: The 3 activities of the current task

The current report is organized in the following way: In the first section we present short information about the registration system and the registered users. In the second part we present the project of the log files analysis as it was

realised with the two partner institutions. We present the project both from a project management point of view as well as regards the outcomes. A more detailed account, though, regarding day-to-day project management actions is provided in the EDLproject Quarterly reports. Recommendations and a plan of further actions is also included in the current report. In the third section we present promotion activities as regards OAI-PMH that The European Library has undertaken to its partners. The two last sections are accompanied by an overall evaluation of the activities and a list of actions that are planned to take place in the second part of the WP1T3 part.

The following people have contributed to the report: Eric van der Meulen, (technical project manager, TEL Office), Olaf Janssen (project manager, TEL office), Sjoerd Siebinga (technical project manager, TEL office) and Julie Verleyen (Technical Technical Team Manager/Developer). Other partners have contributed with their actions in this Task. They are mentioned in the document.

Part A: The registration function

Maristella Agosti from the University of Padua, Department of Information Engineering and Gerhard Weikum from Max Planck Institute for Informatics evaluated The European Library log files and recommended in a paper delivered on the 23rd of June 2006 that The European Library should develop a registration and authentication procedure for the users. It would help gather, with the consent of the users some personal information such as the profession, the country they are based on and particular areas of interest for research.

The registration function was developed and went live in August 2006 (http://www.theeuropeanlibrary.org/portal/user_registration.html). Currently is currently available in the 21 languages of The European Library interface. Log in was until recently possible only from the portal's home page. Realising that this provided an extra difficulty for users to register, the possibility was expanded to all the "Organisation" pages. The information is stored in a MySQL database.

Upon registering, the user is asked explicitly to provide some information about himself¹.

Mandatory information gathered: the user's name, the country where he comes from and their profession.

Additionally, the user is asked to provide at his own will the following information: sex, the place where he/she works (university, library, private individual, etc) , and the areas of his/her interest choosing (up to 5) predefined fields from a menu. He/she is also asked where she learned about us, again choosing from particular options. A radio button selection asks whether he/ she wants to participate in a user survey and to receive the newsletter. The default setting is no. In the figure below you see the registration form.

¹ The European Library is very much concerned with protecting users' personal data. The Privacy policy (http://libraries.theeuropeanlibrary.org/policy_en.html#privacy) discloses The European Library's practice in collecting and analyzing personal data.

HOME	COLLECTIONS	TREASURES	LIBRARIES	ORGANISATION
------	-------------	-----------	-----------	--------------

Register

As a registered user you are able to:

- ◆ save favorites for referencing at a later date
- ◆ save session query history for referencing at a later date

Services will be expanded in the near future to include:

- ◆ save selected collections for use at a later date
- ◆ receive email alerts concerning collections of interest
- ◆ receive email alerts concerning subjects of interest

Username and email

Choose a username: *

Email address: *

Confirm email address: *

Your information

First name: *

Last name: *

Salutation: ---

Country: * Please select...

Your profession: * Please select...

Your place of work: * Please select...

Areas of interest: *
NB You can select more than one item (up to 5) in the list by holding down the Control [Ctrl] key while clicking on consecutive items.

- Computers, information and general reference
- Bibliography
- Library and Information Sciences
- Encyclopedias
- Magazines

Where did you hear about us: * Please select...

I wish to receive the newsletter: yes no

I will participate in user surveys: yes no

Please type in the code you see below: *

253476

RESET SUBMIT

Figure 2: The registration form

In exchange of the information provided the users have access to the following services:

- save favorites for referencing at a later date
- save session query history for referencing at a later date

Information about the users is gathered for two reasons: one is to save their preferences in a database which can be used to send information to them about new additions or changes that might be interesting for them. An automated system is not yet in place but there is planning for developing RSS feeds for notifying the users about new additions. On the other hand, recorded information about them can be combined with The European Library log files that show every activity they have performed on the portal. Thus, more specific information can be gathered about the registered users as there is a record of personal details and preferences and records of their interactions with the portal. Through the analysis these data can be combined in order to extract particular navigational patterns. An automated way can be set up to propose materials, collections, services, etc, that better suit his profile. This is the goal of the project that is presented in the second part of the report.

However, the number of registered users has remained small (in August the number was around 1000) and is growing slowly. That is because The European

Library aims to provide free search and access to all but also because technical architecture restraints have made it difficult to provide any real usability on the portal. The latter is partly addressed by the adoption of OAI-PMH and the expansion of the Central Index. Feedback shows that the users are not finding the services provided so far interesting enough to register and log in again.

However, some functions are only possible if the users register and then log in the portal. In this way, their personal information can be “activated” and they can receive particular information they have asked for. For this reason, there is provision for the expansion of the services for the users that are willing to receive more personalised information. Services will be expanded in the near future to include:

- save selected collections for use at a later date
- receive email alerts concerning collections of interest
- receive email alerts concerning subjects of interest

The data from the registered users’ database was made available to the partners from the University of Padua and Max Planck Institute for the purpose of the research of the log files.

Part B: Log files Analysis Project

Introduction

The importance of taking more into account user needs and requirements into the design of the interface and the services of The European Library is the driver behind the current undertaking of analysing The European Library log files. Log files are small pieces of information that are stored on the servers of an online service and which record information about the users as well as their click-through history: their IP address, the domain name where they come from, the language they use for searching, the operating system, at what day and what time they accessed the service, etc. The intent is to feed information from the analysis of the log files into the design of personalised services that better fit the needs of individual users.

The need to involve more actively the users in the design of the interface and the services of The European Library has been the outcome of the two user studies that The European Library conducted in 2005 and 2006. It was also a recommendation of the final review of the TELMEMOR project. The European Library has done two user surveys in the past to approach the way the portal is perceived by its users. User surveys have been very informative about the way users interact with The European Library portal. Outcomes showed that users do not like to use advanced search options and they don’t understand the necessity of selecting collections for their queries. Both these characteristics are currently pivotal for effectuating a query in The European Library that delivers satisfactory results. User surveys have indicated the way for enhancing the usability of the portal. Much of the feedback received from the user studies has immediately fed

the portal usability developments and the overall product description and planning of The European Library.

After having done two user surveys, The European Library office decided to experiment with analyzing The European Library log files for receiving input regarding the user actions and needs for the following reasons:

- Log files analysis provides intuitive input. User surveys usually provide a controlled environment or controlled questionnaire, even the selection of the users is sometimes controlled. Log files are "impartial" in the way that everything is recorded and can be analysed from all the users of the portal. Logfiles record in an intuitive way all users' interactions with a service.
- Granularity of information provided. User studies can not go down to identifying individual needs. The outcomes that are particularly interesting in user studies are the greater figures in the percentages, that is, what are most of the users doing or believe. In view though, of The European Library moving towards a release compliant with Web 2.0, more information was necessary about the users, information that could be exploited in a way that individual preferences and search habits could be explored in view of better satisfying these needs. It makes it easier to identify different user groups (the recurrent users, the bouncers, the "ethnic" groups, the "academic" groups, etc...) and their characteristics with greater accuracy. It also makes it easier to track individual users and their habits.
- Mass of information. Contrary to the user studies where there is a limited (though representative) sample of users, in logfiles analysis all records form the sample of analysis. The use of a greater amount of data allows for a granularity of analysis from individual users to different groups.
- Evolving information. Another advantage that log files offer is that they record in a continuous way the user activity. This means that analysis does not provide a "frozen" image in time. It can be an ongoing process that helps not only delivering personalized services but also to show evolving user profiles and habits in time.

Therefore, the greater advantage that log files probably offer is that they can provide the basis for delivering personalised services and results to groups of users or individuals. Log files are not "personalised" (no personal data are stored there) but still personal in the way they refer to an individual user. Though difficult technically, log files can be used to understand what an individual user is looking for and how he goes about searching. In an ultimate way and if combined with some information that the user agrees to provide-like preferences of collections or themes or topics and language, he can get back results that are more likely to satisfy his personal information needs.

Issues

Log files provide a very rich pool of data where information like the IP address of the user, where he entered the site from, the time and the day he accessed the

portal, what keywords he entered, etc, are recorded. The information is recorded in a structured, however, not immediately understandable or readable way. Log files information has to be to be accessed, organized and parsed using some method that will arrange them in meaningful chunks of information and interpreted in a way that provides answers to questions set.

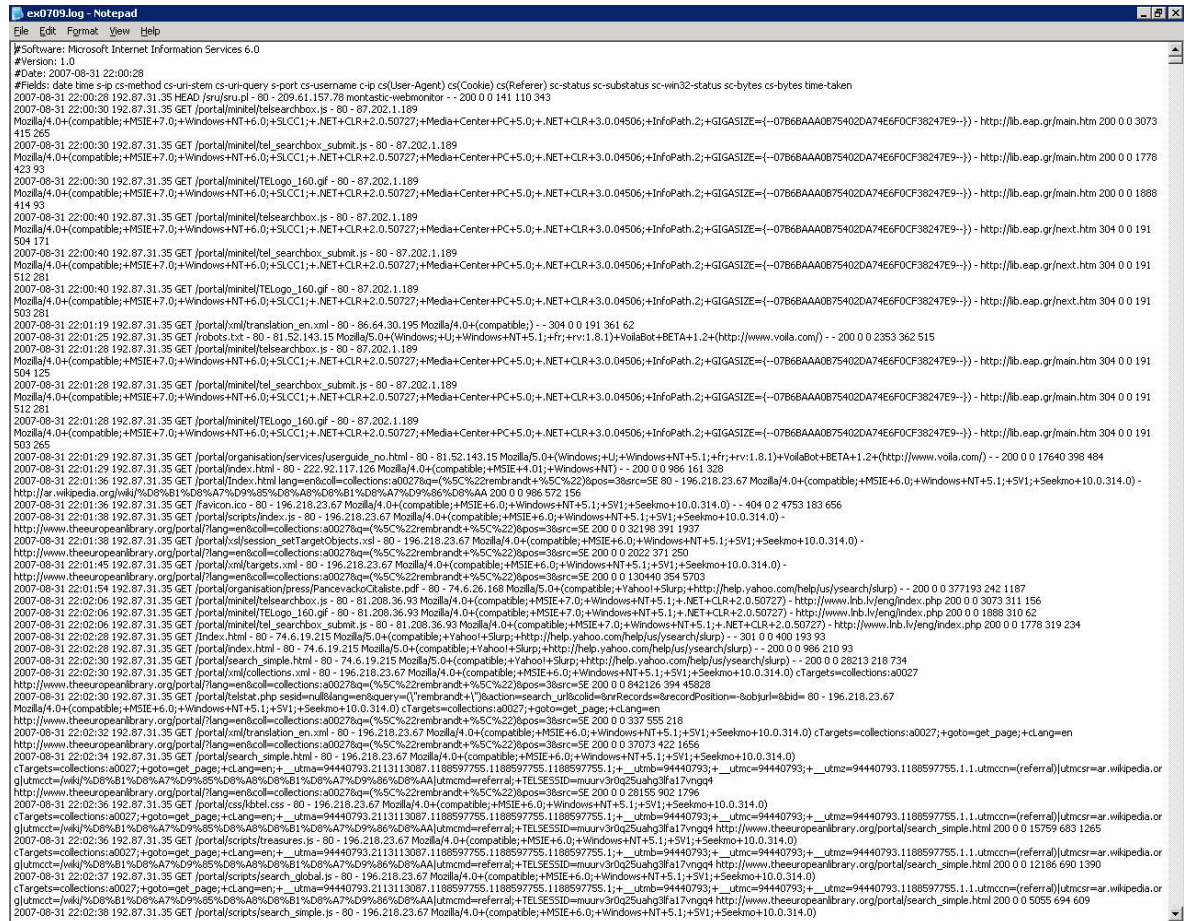


Figure 3: Screenshot of the HTTP log files

In this process, there are some limitations and some shortcomings that have to be overcome. First of all the information is so rich that it can end up being meaningless. There are thousands of requests recorded every day on the TEL servers. These come from real users but also from other websites that link to The European Library, from crawlers and software agents. What is useful to know has to be defined beforehand as it affects the way information is parsed, combined and analysed in order to produce meaningful information about the users.

Technical bottlenecks have to be addressed as well. For example, individual users cannot be identified in the cases as organisations and institutions often use proxy servers that collapse several users in one IP address. In this case the use of heuristics and assumptions are necessary in order to define as accurate as possible users and other variables of interest.

It has to be mentioned here that the current project is not the only way that The European Library has made use of the log files. Since 2006 we have been using commercial software tools that analyse the portal's http logs and give results about the traffic. The first tool we used was Awstats (mid 2006) and since a few

months now we have been using Google analytics. These tools provide statistical reports that have been useful for counting visitors and supplying statistical information about them like the country they come from and other information we extract from the TEL HTTP IIS logs. However, the use of these tools and the results they give is not very transparent. The difference as well in using them lies in the fact that they cannot provide personalised information. Therefore, the scope of the current project is different for the use and the aims of web statistics tools and the aim is not to extract simply statistic information about the users and how they perceive our portal.

Project Description

About the log files analysis project

Background

The current project involves the analysis of The European Library logfiles with a view to understanding more profoundly who The European Library users are and how they interact with The European Library portal. The European Library will use the analysis to help build better services for the users and develop the portal according to user needs.

For the needs of the analysis The European Library addressed two specialised institutions for collaboration: Department of Information Engineering of the University of Padova² (UNIPD) in Italy and Max-Planck-Institut für Informatik³ in Germany. Professor Maristella Agosti⁴ led the research team in the University of Padua and research in MPII was supervised by Gerhard Weikum⁵ Research Director at the Max-Planck Institute. In both cases doctoral students were involved in the project: Tullio Copotelli and Giorgio Maria di Nunzio in UNIPD and Julia Luxenburger in MPII. MPII and UNPID are both members of the DELOS Network of Excellence on Digital Libraries.

After some initial preparation, a project proposal was compiled and signed by all parties. The objective of the project proposal was to split The European Library log files to the two parties and to allocate separate areas of interest that each institute would investigate. The aim of the project is to understand better who the users of The European Library are and to try to draw an as clear picture as possible about them. It is hoped that this information will help us develop parts of the portal, develop or discard services and eventually be able to provide personalized services to the users.

What are the log files?

Logfiles are pieces of information that are being recorded on the server of a system; they record users' interaction with the system (click-through history) such as how they navigate, what they look for and for how long, as well as,

² <http://www.dei.unipd.it/wdyn/?IDsezione=2>

³ <http://www.mpi-inf.mpg.de/>

⁴ <http://ims.dei.unipd.it/members/agosti/>

⁵ <http://www.mpi-inf.mpg.de/~weikum/>

information about their local system (what browser they use, what country they come from, etc). This information is non-personal, and therefore the analysis does not violate individuals' privacy. Initially log files were used to monitor traffic and better administer a website. It was soon realised that the analysis could help track down users' actions, reveal where users stumble, what attracts them better or where they decide to leave. Besides users' studies and questionnaires, or rather, complementary to it, logfile analysis offers a significant tool for understanding users' intuitive search habits.

The European Library log files and background work

Initially The European Library had been recording the following Log files:

- The Verity server logs (action logs, user tracking)
- The IIS - http traffic logs, divided in two⁶: 1)the logs from the static part of The European Library portal - "ABOUT US, LIBRARIES, TREASURES" (<http://libraries.theeuropeanlibrary.org>) and 2) logs from the dynamic part, i.e. the search possibility under SEARCH and COLLECTIONS in the main navigation - <http://www.theeuropeanlibrry.org>. IIS log file format is conformant with W3C specifications⁷.

There has been some initial work on The European Library log files accomplished by DELOS before the start of this project. In response to a request by The European Library, Max Planck Institute and The University of Padua looked into The European Library log files recording formats and proposed some changes for better logging recording in June 2006. Their recommendations resulted in changes in the logging formats but also the setup of a users' authentication system (registration and log in).

The University of Padua analysed the IIS http traffic logs, mainly focusing on the Logs from the dynamic part of the Portal (search) from November 2005 and January 2006 and came up with a report for enhanced http logging on the 26th of June, recommending also that The European Library collects the following information in http logs: Bytes sent, Bytes Received, Time Taken and Cookie values. These Logfile format changes have been applied in The European Library logging system in September. More specifically in September the IIS logging format was changed to record also the referrer and cookies fields.

Prof. Gerhard Weikum and his doctorate student Julia Luxenburger from Max Planck Institute analysed The European Library Verity server logs from March 2006 and came up with recommendations on the 23rd of June 2006 for enhanced logging, proposing to record all actions of users on the portal as well as additional data such as sessionId, timestamp, etc.

The userId would help identify an individual user even he returned after a long time to the portal. SessionId and Timestamp could help in identifying to what extent users rephrase a query for the same information demand or switch to search material for a new topic. These recommendations led to the set up of a

⁶ This has currently changed

⁷ <http://www.w3.org/TR/WD-logfile.html>

new logging system by Eric van der Meulen (The European Library Action Logging System).

Additionally, Max Planck as well as UNIPD proposed the set up of a registration system and an authentication procedure.

Currently there are two types of TEL log files:

- The HTTP IIS log files record the following information: IP (Internet Protocol) address, the user agent (human, crawler, software, etc), referrer field (a Uniform Resource Locator address which communicates the last page viewed by the user) and the cookie⁸ which records in its turn: the language selected by the user during the navigation of the service; the collections that the users have selected and the session identifier assigned by the server to the user. The TEL HTTP logs are conformant to the W3C Extended Log File Format.
- The TEL Action Logs. After the examination of the verity server logs and the reports that the software produced by Max Planck Institute a new system for recording was decided to be put in place. The TEL Action Logging system was developed by Eric van der Meulen. In this system a single entry is recorded for each action. The functionality is realised using the same principle that the search/retrieve functionality uses: client-side http requests. But rather than requesting information for the user from a remote database, this functionality posts information about users actions to a mySQL database. In this case all queries are logged, not just queries to the central index (which are available in the Verity logs). The information stored in the database can be collated in such a way as to give a clear view of user behavior in the portal. This function went live on the 19th of December which means that little data was available at the start of the project. The original Verity logs are now a subset of The European Library Actions logs. The MySQL database schema is analytically presented in the Annex 1.

Additional information: The registered users database

In their examination of The European Library log files, both institutions also recommended that The European Library sets up a users' authentication process which was realised as described in the previous chapter.

The intent behind the setup of the authentication of users was dual. On one hand, users would be asked to proactively provide some information about them which could be later used to send them personalized information. For example, they would be notified about the addition of new collections to their interest. It was also considered that tracking through the log files a registered user would provide a much deeper understanding about his implicit needs and habits, which would again be useful for providing personalised services. The aim was therefore to use both sources of data, the log files and the data from the registered users in order to associate particular users' profiles with certain usage habits.

⁸ Cookies contain data sent by a Web server to a Web client, are stored locally by the client and sent back to the server on subsequent requests

In case a user has registered, her respective userid allows to identify her and relate her with her profile captured in a separate table (tel_user_share). Otherwise the default value for this field is set to "guest" and the only way to possible group actions by users is through the users' IP addresses.

Individual goals

Individual goals were set for each institution. They are described later in the document.

Deliverables

The project run from April to June 2007 and resulted in two reports from each institute incorporating the analysis as well as recommendations for improving usage and usability in the portal. In the project proposal the deadline for the interim report was set for the 4th of June and the final report would be delivered on the 2nd of July. However, due to feedback given from TEL Office, the University of Padua finally presented 2 versions of the interim report and 3 versions of the final one. The last one was delivered on the 27th of August.

Description of tasks and responsibilities

Georgia Angelaki, The European Library, EDL project Assistant

Coordination: allocation of tasks, meetings organisation, follow-up of actions, responsible for final reporting

Eric van der Meulen, The European Library, Technical Project Manager

Technical responsible for log files and for the dispatch of the data to Max Planck Institute and The University of Padua, implementation, technical feedback, input to GA.

Prof. Maristella Agosti, Department of Information Engineering, University of Padua

Supervision of work in The University of Padua and reporting, meetings participation

Giorgio di Nunzio, Department of Information Engineering, UNIPD

Log files analysis, delivery of methodology and tools, reporting, meetings participation

Tullio Copotelli, Department of Information Engineering, UNIPD

Log files analysis, delivery of methodology and tools, reporting, meetings participation

Prof. Gerhard Weikum, MPII, Department 5: Databases and Information Systems

Supervision of work in Max Planck Institute and reporting, meetings participation

Julia Luxenburger, MPII, Department 5: Databases and Information Systems

Log files analysis, delivery of methodology and tools, reporting

Jill Cousins, The European Library, Director

General input

Olaf Janssen, The European Library, Project Manager
General input

Communication

There were three teleconferences held with each institute, together or separately and numerous e-mails were exchanged for the needs of the project. In particular, TEL office provided technical advice and feedback to the partners to support them in their research.

Dissemination

In the course of the current project both institutes delivered papers to conferences in which they presented the methodology they adopted and their initial findings. This was outside of the scope of the current project. However it adds some credibility to the research as well as gives some publicity to The European Library activities in the academic community which is one of our main target audiences. The following articles were written for the relevant conferences:

- Agosti M., Di Nunzio G.M., Niero A., From Web Log Analysis to Web User Profiling, Second DELOS Conference on Digital Libraries, 5-7 December 2007, Tirrenia, Pisa, available at: http://www.delos.info/index.php?Itemid=171&id=526&option=com_content&task=view
- Luxenburger J., van der Meulen E., Weikum G., A User interaction model for The European Library Portal, PersDL 2007, 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries, Paper available at: <http://www.dblab.ntua.gr/persdl2007/papers/22.pdf>
- Agosti M., di Nunzio G.M., Web Log Mining: A Study of User Sessions, PersDL 2007, 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries, Paper available at: <http://www.dblab.ntua.gr/persdl2007/papers/72.pdf>
- Agosti M., Angelaki G, Coppotelli, Di Nunzio G.M., Analysing HTTP Logs of an European DL Initiative to Maximize Usage and Usability, paper in progress for ICADL 2007.

Project Implementation

The University of Padua

Individual goals

The goals for INIPD in the project proposal were set as follows: "The University of Padua will use IIS and The European Library Action logs from January 2007 to April 2007 to perform a deep analysis of the usage sessions: to analyse traffic to the portal and to identify general profiles of The European Library users.

The University of Padua will address the following issues:

- Deep analysis of sessions using the log data (data from The European Library Action Logs, The European Library HTTP logs and data from the registered users database) from October 2006 on; the sessions to be analyzed are going to be those reconstructed by means of heuristics, and those stored in the cookies.
- Analysis of access to The European Library from different geographical areas both for human users and crawlers.
- Preliminary study on what user profiles can be built from the knowledge contained within The European Library registered users' database and The European Library Action logs.”

Methodology of parsing

A methodology had been conceived before the start of the project by the University of Padua: it consisted of a parser to gather information from the log files and a database to store the information extracted. The database enables separation of the different entities recorded and facilitates data-mining and on-demand querying of the data. It makes possible to merge the information from different sources and therefore, to compare and combine more easily information from the log files with the registered users' database data. The specifications of this can be found in [see DELOS Conference on Digital Libraries in Pisa on 13-14 February⁹ ...]

A database schema had been developed to match the logging format for The European Library HTTP logs that the University of Padua gathered from November 2005 to September 2006. The schema was redesigned to match the new logging format after September 2006.

Findings

Based on the final report that UNIPD delivered on August 27, for the analysis they used data from 1st October 2006 to 30 April 2007.

UNIPD used the HTTP logs for its analysis. They made an initial analysis of the TEL Action logs but did not associate the data from the two sources.

One of the main concerns of UNIPD was to identify individual sessions as accurately as possible. Sessions form the basis of any analysis of users' actions. Besides this, the main questions they addressed as well as the findings are presented below.

General information:

Requests

A total of 22,458,350 HTTP requests were recorded during these seven months in the http logs. Thus, The European Library is considered a busy service. A

⁹ http://www.delos.info/index.php?option=com_content&task=view&id=526&Itemid=273

higher activity is observed during the day than in the night, considering that the time recorded by the server refers to the Central European Time. Such an activity corresponds with the higher number of visitors coming from Europe.

Operating systems

Most of the users are using Microsoft products: 74% uses Windows, 1% uses Mac and another 1% uses Linux while the rest use other operating systems. Microsoft Internet Explorer (60%) and Firefox (13%) are the two most widely used web browsers.

Reconstruction of sessions

The aim of the session reconstruction is to identify single users and to separate human users from software agents (crawlers, spiders, banners, etc) that create traffic to the portal too. UNIPD used two approaches to identify individual sessions:

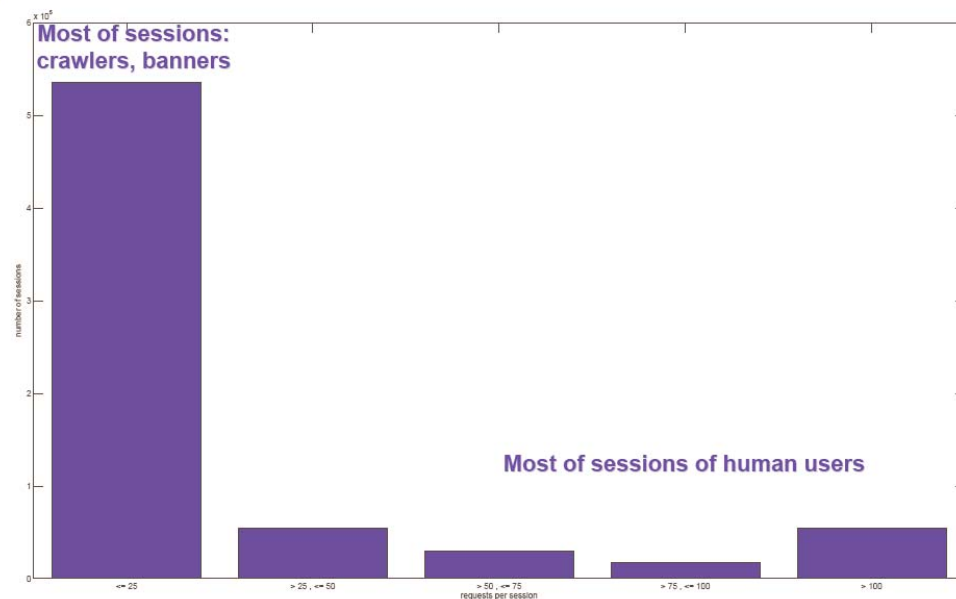
- Using a heuristic that combines the IP (Internet Protocol) address and the user agent allowing a fixed gap of time between two consecutive requests. This method shows 690,879 sessions for these seven months but does not distinguish between human users and crawlers. Software agents create the greater traffic to the portal. For this reason, using this method it appears that the great majority of sessions (77,44%) have a small number of requests(≤ 25). UNIPD states that sessions with more than 100 requests are only 7,90%. These are indicative of human interaction with the portal.



Reconstruction of sessions: use of a heuristic

Number of sessions:

6×10^5



≤ 25: a very small number of requests

figure4: Reconstruction of sessions with the use of a heuristic¹⁰

- Using cookies: This method is more reliable for identifying human users. If a user disables cookies on The European Library portal most of the functionality of the portal is blocked. Therefore this method counts real sessions by users on the portal. Over the period under examination there were 209.900 different sessions on the basis of the cookies content. This point was not very clear to the University of Padua who thought that using this method there is a risk that sessions from users who have disabled cookies are not recorded in the logs and therefore the numbers of sessions appearing in this way are much smaller.

There is a sizable number of sessions, almost 45% of the total number of sessions which last more than 60 seconds.

¹⁰ Tables and figures in this part of the report come from the final report of the University of Padua

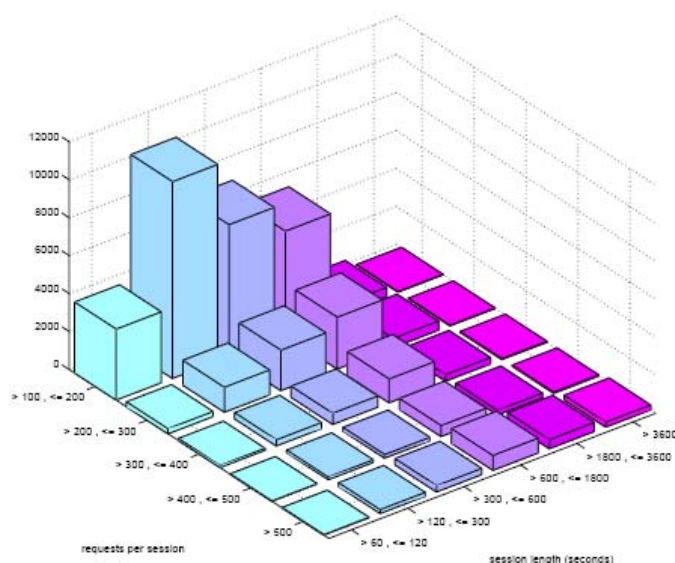


Figure 5: Sessions (cookies) which last more than 60 seconds with a total number of requests per session >100.

The following characteristics are identified about the sessions:

The vast majority of the sessions ¹¹(77, 44%) involves 1 query (1-25 requests). Sessions are short in requests and in duration. The rest of the sessions are interesting because they clearly refer to human activity on the portal and involve more than 1 query. These sessions last mainly between 2-10 minutes and involve 100-200 requests. There is a sizable number of users that spend quite some time interacting with the portal.

The number of sessions that the two methods show varies greatly. The heuristics method is interesting for monitoring overall traffic to the portal and might be useful for separating crawlers' activity from other agents and the requests created by the Mini Search Box¹² of The European Library when it is loaded on a remote site. However, when it comes down to human users the cookies method is more useful. Wherever we can, we will be presenting the results only from the cookies method as they are indicative of human activity on the portal. The heuristics method has revealed the following interesting issue.

Examining the national provenance of users UNIPD observed a significant discrepancy between the figures from the heuristic method (10,78%) and the cookies that showed only 0.99% incoming users from Latvia. This led the University of Padua to investigate and realise that because of the Mini Search Box placed in the portals of some national libraries, every time such a page from a national library loads there are 3 requests recorded on The European Library log files.

¹¹ Using heuristics though, which include software agents.

¹² The Mini Search Box (http://www.theeuropeanlibrary.org/portal/organisation/services/services_en.html#mini) is a promotional version of The European Library search box. Any website owner can request and place it. Users perform a query directly from other websites to The European Library website.

The European Library has developed a Mini Search Box that other website administrators can request and put on their website. The Mini Search Box has been used by libraries, various websites and blogs and appears to be successful (again from the logfiles). UNIPD observed that every time the MSB is loaded from TEL, there are 3 requests recorded on TEL servers regardless of whether a user types in or not a query. This is an interesting observation for TEL that could use log files to monitor the use of the MSB from its partners. It is a very concrete example for the usefulness of log files to estimate the success or failure of a particular service developed by TEL. It is an interesting finding that needs further analysis in combination with other data such as what is the use of the MSB, by which users, from which countries, what kind of queries they enter in the MSB, if they are recurrent, that is if they keep using this pathway to enter TEL, etc.

Analysis of national provenance of users

The next issue that UNIPD tackles is the national provenance of users. UNIPD uses the ISO codes to demonstrate traffic that comes from particular countries.

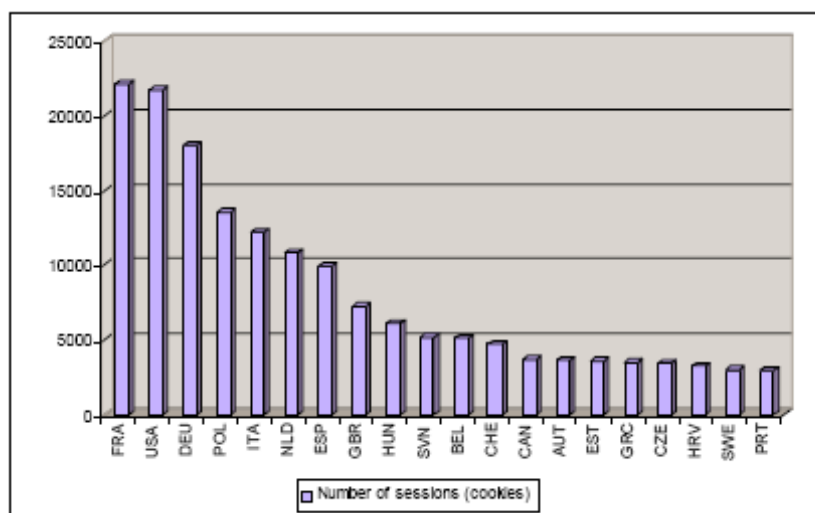


Figure 6: The nations with the highest numbers of sessions (cookies)

Human users and crawlers

75,72% of the total requests to the portal are effectuated by human users while the rest of the requests come from software agents, crawlers, etc.

Recurrent users and bouncers

UNIPD used the __utmoz values set by Google analytics to identify returning users. The results are presented in the table below.

# of visits	# of users	# of sessions	pct of users
1	98,281	98,281	82.35%
2	12,492	24,984	10.47%
3	3,737	11,211	3.13%
4	1,657	6,628	1.39%
5	848	4,240	0.71%
6	519	3,114	0.43%
7	361	2,527	0.30%
8	224	1,792	0.19%
9	217	1,953	0.18%
10	117	1,170	0.10%
11	91	1,001	0.08%
12	89	1,068	0.07%
13	66	858	0.06%
14	51	714	0.04%
15	43	645	0.04%
> 15	555	19,760	0.47%
Total	119,348	179,946	100.00%

Figure 7: returning users

Most users (82,35%) of The European Library are one time users. Among the rest, during these months a very small percentage (0,47%) accessed the site more than 15 times.

Registered users

Among the registered users (614 for this period) 349 had returned at least one time. 365 registered but never logged in. UNIPD observes that more than half of the returning users in fact returned only one time, that is directly after they registered and logged in within 15 minutes. After this they never logged in again. Only 63 of the registered users (18%) returned again a day later or further after their registration. Out of the registered users that performed a query 73 performed only one query session.

Referrer and country

The log referrer field is analysed for the provenance of the users. The University of Padua focused on the data from April 2007.

Domain	# of requests	nation	percentage
www.lnb.lv	46,269	lv	23.05%
www.nlib.ee	31,400	ee	15.64%
zbdigital.blogspot.com	17,597	com	8.77%
www.knihovnabbb.cz	16,817	cz	8.38%
www.bn.org.pl	11,718	pl	5.84%
www.biblioteke.org.yu	10,471	yu	5.22%
www.nsk.hr	7,353	hr	3.66%
www.kodolanyi.hu	4,221	hu	2.10%
librarylingo.web-log.nl	3,293	nl	1.64%
vivl-livad.voi.sch.gr	2,310	gr	1.15%
www.corriere.it	1,791	it	0.89%
hodges-model.blogspot.com	1,723	com	0.86%
neptun.lbi.li	1,699	li	0.85%
www.digital.nbs.bg.ac.yu	1,658	yu	0.83%
www.google.com	1,390	com	0.69%
janmetdekorteachternaam.blogspot.com	1,247	com	0.62%
www.janmetdekorteachternaam.blogspot.com	1,052	com	0.52%
digiemb.blogspot.com	1,031	com	0.51%
others	37,722		18.79%
Total	200,762		100.00%

Figure 8: Top referrer websites

A listing of some of the top referrer websites for April was provided with Latvia being the website where most of the requests originate from. The results are however equivocal; they show the total number of requests to TEL, that is they include the total number of times that the MSB was loaded to other websites even if it was not used for a query. In fact the ratio of actual visits to requests of loading of the MSB for the Latvian national Library was 1 to 800. The ration for Estonia is 1:86. This is an interesting point about the provenance of the users that needs to be better analysed. Separate figures have to be provided for users that actually visited TEL and the number of times the MSB was loaded to other portals as both numbers inform us about different things.

It is interesting to observe that there are a lot of blogs that actually generate a lot of traffic.

TEL Action logs

A first analysis of the TEL Action Logs is attempted, however, without associating this data with the data from the web logs. The TEL Action logs show only user sessions that perform queries on the portal. 53.475 sessions were reconstructed from January to April 2007. The most striking finding is therefore that the majority of visitors to the portal do not perform any query. This information deserves more analysis to understand the motives of the users. Does this mean that they don't understand how to search? Is it related to the language interface? Is it because they are only interested to receive information about The European Library but not search something on The European Library? From this point on UNIPD suggests that user surveys are used to identify the incentives of the users. For this reason, a small scale user study was realised with a group of Masters' students in Information Retrieval at the University of Padua. The users were asked to freely interact with the portal and then fill in a questionnaire. The study is discussed further down in the document.

# of sessions	# of users	percentage
never log in	316	41.20%
log in without quering	351	45.76%
1	73	9.52%
2	15	1.96%
3	2	0.26%
4	6	0.78%
> 4	4	0.52%
Total	767	100.00%

Figure 9: Number of Query sessions and the registered users- April 2007

User study

The University of Padua requested from TEL to use a questionnaire that had been used in 2006 for an online users' survey. UNIPD asked to use it to perform a small scale survey with students from the university. The reason was to cross reference the data from the logs with the explicit information supplied by a known and defined user group.

The goals shifted in the final report. As it is stated "The final aim of the study is to gain insights on a specific group of data, and to use them in a more general way, possibly generalizing them". However, in the final report the characteristics of the survey that UNIPD performed do not appear. UNIPD goes on to present some results but we are hesitant as regards the validity of this survey. The only reference we are given is that a "group of people" was involved.

For this "group of people" we learn that they do not understand the use of the collections selection and to a large extent the results duplicate what we already had known from the very extensive and elaborate user survey of 2006. The connection between the survey and the log files analysis is by no means provided. We had provided our objection to the user survey just be mentioned in the report without further explanation. It is an item that does not fit in the initial project proposal and is not clear as to what it is useful. For this reason we will not expand any further here.

Recommendations

UNIPD provided the following recommendations:

- Encourage registration and logging in of users by:
 - Offering extra functions and services that are only available to them
 - Making logging in function more obvious by adding it the username and password in the front page
 - Assigning a unique identifier to each new user. Immediately recognise a registered user when he visits the site by the cookie and log the same identifier
 - Inserting in the authentication process the possibility to be automatically recognized without the request of user name and password. This can be done with the use of the cookie.

- Consider the possibility of introducing language parameters in the search depending on where a user comes from; e.g. prioritise on results given from the national library of his country
- Consider combining user studies with log files analysis to get a deeper understanding on users
- Consider providing more advanced search options for particular user groups
- Add an option to the user to look only for digitized items

Max Planck Institute

Individual goals-MPII

As mentioned earlier, the trigger for the analysis of the TEL log files is to define with as much precision as possible individual users and user groups so as to develop personalized services for them. In the framework of the current project Max Planck Institute decided to focus on what constitutes the core service of the TEL portal: the search and retrieval of documents and investigate query-related issues.

Max Planck Institute proposed to analyse the query and result-click history of individual users and user communities, in order to develop a statistical model of user interaction behavior. In particular MPII proposed to focus in the following questions for the analysis of the logs:

- How do users go about rephrasing their queries when the first results do not match their information needs?
- Which are the collections that an individual user typically prefers (e.g., by frequently clicking results)? How do queries and preferred collections statistically relate?
- How many result clicks (and interleaved query reformulations) does a user typically need to find the desired information?
- When is a user satisfied with the final results, and when does he give up?

- At the start of the project Max Planck Institute estimated that the model would allow recommendations to be made for the development of The European Library portal such as:
 - selecting preferred collections on a per user basis
 - ranking or re-ranking results based on user's previous interaction behavior
 - automatically rephrasing or expanding a user's queries when initial results are not satisfactory

For the needs of the analysis Max Planck Institute used the available data from The European Library Action Logs from 19 December 2006 to 31st of May 2007 together with data from The European Library registered users' database for these months.

Methodology

The main methodology that MPII developed involves the statistical model of the navigational patterns of users.

Additionally, the following assumptions are made for the needs of the research:

- Sessions are identified by the PHP sessionID and the additional requirement of no more than 5 min of inactivity between subsequent action within the same session. As mentioned before, session reconstruction forms the basis for any user identification. This is not always straightforward due to the fact that the time a session ends is not defined in the logs, therefore, assumptions have to be made. MPII identified individual sessions based on contiguous (semantically related) queries.
- The sessions are reconstructed either by identifying registered users or using the sesid in combination with the date fields (contiguous sequences of user actions).
- Individual users were identified by registration or via the cookies.

Findings

General information

Statistical data are easier to extract and give some information about the users. MPII supplied the following information that can be extracted from the TEL Action Logs:

- Language: The language selection in the interface of the portal gives only hints about the background of users. The majority of users (84%), leaves the default interface language English, despite of whether they are native speakers or just capable enough with the English language.
- Registered users. There was a very small amount of registered users (827) by the end of May. There is some equivalence between the countries with the largest amount of visits and the countries where most of the registered users come from (France, Poland, Italy). However, registered users mainly experiment with the portal but don't log in again even if they return to the portal.

Search sessions and queries

MPII focus has been on queries (re)formulation, navigational patterns and collection selection. These issues are interconnected and will hopefully, as more data accumulate, give a good picture of the search habits of the users, their satisfaction level and will lead to recommendations for improvement of what forms the basic service of the portal.

Session length: The average number of actions is 9,16 and the average number of queries on the portal is 2,55. According to MPII these figures are conformant to what is known for users from other web studies. Queries in The European Library however differ from what is standard in the web in general.

Query formulation:

By far the majority of queries in The European Library are plain keyword queries. The table below depicts the use of the advanced search elements.

Syntax	#	Syntax	#
AND	6557	Type	209
OR	305	Language	1,065
NOT	265	ISBN	1,124
Title	10,102	ISSN	267
Creator	8,223	Plain	94,234
Subject	1,813	Total	112,708

Figure 10: frequency of usage of query syntax¹³

Keywords

MPII also provided a table for the most common queries encountered in the timeframe of the research. From the top20 most frequent queries we observe a great frequency of keywords relating to a nameplace: Poland appears in 126 queries, France in 73, polska in 70, Zagreb in 67, Europe in 59, Italy in 58 and London in 51.

There is also great frequency of keywords relating to an author: Shakespeare (121), Kochanowski (91), Dante (89), Goethe (79), Mickiewicz, Hugo, Cervantes (53). In the 20 first places we only encounter 2 books titles ("Harry Potter" and the "Bible") and two general topics appear ("energy", "history"). It is interesting as well that the most frequent keywords relate to a Europe-related placename or subject. However, the reoccurrence of keywords is rather small.

Query term correlations:

Query Term Correlations are investigated in view of making recommendations about automatic query expansion or query reformulation suggestions.

The logs contain queries such as ("title all lenz and creator all büchner") in the simple search. Most probably such queries are the result of users trying to use advanced search features when posing queries through the simple search interface.

A first attempt has been made to identify correlations of terms appearing in the search. 3 methods were used for these in order to investigate both simple and advanced search. This is an interesting effort although the results are poor yet, most probably because little statistical data is issued from the logs and more interesting data will occur as more months of logfiles accumulate. It is interesting however to see that there are repeated queries that show semantic correlation which also justifies the attempt of TEL to build such a component of automatic query expansion¹⁴.

Users' navigational patterns

¹³ Tables and figures at this part of the report coem from the final report of Max Planck Institute on the Analysis of The Euroeapan Libray Log Files: <http://www.edlproject.eu/membersonly/downloads/tel-report-less-technical.pdf> (password-protected area)

¹⁴ TEL is prototyping a MACS (Multilingual Access to Subject Headings) tool

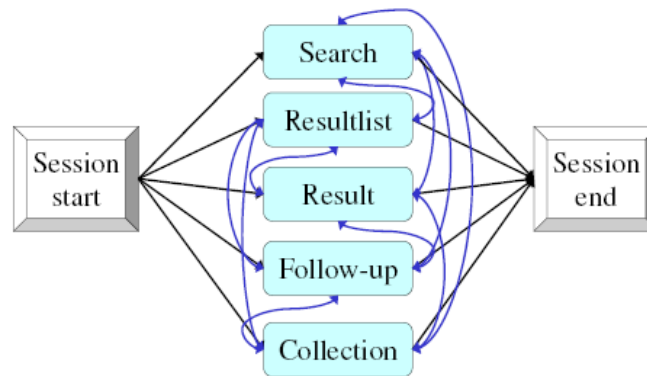


Figure 11: Markov model of the users' search behaviour

MPII devised a Markov model that allows to group users' actions within the portal and to identify usage patterns. The model allows to draw some conclusions about users' satisfaction as well as to see which collections are preferably queried and propose methods for automatic collection suggestion. In the Markov Model user actions constitute states and are grouped into action types. For example all search actions (simple search, advanced search, search again starting from a results page, etc) are grouped under the search action type. Aggregating all user sessions, MPII estimated the transition probabilities in users' actions. Using this model, MPII asked particular questions about the users and observed the following:

1. How users behave after they pose a query and see the first result list:
 - a. In 38% of the cases they rephrase their query
 - b. 25% view another result list (15% requests the next 10 results, 85% switch to the result list of another collection)
 - c. 4% switches to a completely different collection set.
 - d. 16% end the session
 - e. 17% views a detailed record information

A, b and c show that the user doesn't seem to have found what he is looking for but insists on finding it. D shows that he abandons the search. It is interesting to see that only in 17% of the cases the user seems to be really interested in the query results.

2. What is the last state of a session?
 - a. From questions 1 and 2 MPII is trying to infer the success or failure of a search query.
 - b. Around 18% of the sessions end in result or follow-up action states (successful cases).
 - c. Sessions ending in search states are not successful (22%) as they result in viewing of very few records with a few queries.
 - d. Sessions ending in result list are so-so (46%)
 - e. Sessions ending in Collection Selection (4%) are mostly exploratory as they are short and contain very few queries and almost none results are viewed.

Collection selection

In The European Library, the user has several ways he can select the collections that are of interest to him. The table below shows the order of their preference in choosing method for selecting collections. The great majority of sessions rely on the default collection list without ever changing this setting.

Collection Selection	# Sessions
Thematic	4902
Default	516
Country	127
Subject	33
Description	6

Figure 12: Popularity of method of selecting collections

Comparing the frequency with which collections are queried (view figure 13) and records are viewed (figure 14) as well as their ranks in the default and thematic collection lists, MPII concludes that users most often view the results in the default order presented in the TEL interface but are more thoroughly looking at collections that are not always ranked higher in the default list. In this observation MPII sees a great potential for better ranking of collections.

Collection	#	Rank in default list
Online books, images, maps, music...	103,069	1
British Library integrated catalogue	29,005	2
Online catalogue of the German National Library	10,980	3
SBN OPAC (Italian)	10,000	8
ARSBNI 1	8,906	Top-1 in "digitized books"
BN-OPALE PLUS (France)	6,618	9
General Catalogue Koninklijke Bibliotheek	5,856	5
Serials and periodicals	4,774	Top-1 in "newsreports and periodicals"
Science, technology and business	4,680	Top-1 in "scientific articles"
Cartography catalogue	4,481	Top-1 in "maps & atlases, cartography"
The Danish National Collections	3,935	20
Printed music catalogue	3,589	Top-1 in "music collections"
BNPOL (Poland)	3,451	23
Collections from the National Library of Portugal	3,325	4
HELKA (Helsinki University)	3,215	6

Figure 13: Top 15 collections (viewed resultlists)

Collection	#	Rank in default list
British Library integrated catalogue	3,990	2
Online catalogue of the German National Library	751	3
SBN OPAC (Italian)	629	8
Online books, images, maps, music...	616	1
The Danish National Collections	392	20
BN-OPALE PLUS (France)	382	9
Cartography catalogue	354	Top-1 in "maps & atlases, cartography"
Serials and periodicals	325	Top-1 in "newsreports and periodicals"
General Catalogue Koninklijke Bibliotheek	307	5
Science, technology and business	258	Top-1 in "scientific articles"
Printed music catalogue	173	Top-1 in "music collections"
HELKA (Helsinki University)	171	6
National Library of the Czech Republic	169	16
Collections from the National Library of Portugal	162	4
Amicus	159	14

Figure 14: Top 15 records (viewed records)

The analysis reveals the key role of the collections in the search process. Even if a user doesn't understand the use of collections in the first place and the need to select a small amount of them for his search, it seems that he is fiddling with the collections at a later stage, once he receives the results from the default list and

then selects results from other collections than the one that appear first. Additionally, 4% of users end their session in collection selection which means that probably they experiment with it or that they are deterred from moving on to see a detailed record. There lies indeed a potential for better collection selection or automatic collection selection based on the kind of queries.

Query-dependent selection of collections

While the above provide general information about collections ranking, MPII aims at a later stage to investigate collections dependent on query context in order to propose automatic selection of collections based on the query. In this direction, MPII examines correlations between queries and collections whenever the result record for a given (query, collection) pair is viewed or a follow-up action is undertaken. They focused particularly on the collection of Atlases and Religion (see tables 11 and 12 of MPII report). Information is still very scarce though to indicate any potential for query-driven collection suggestion.

Furthermore a model¹⁵ was considered to learn query –specific rankings of collections. The model considers user-related and query-related parameters such as collection preferences, country of provenance, language, the query associated with a collection preference, etc. The model will prove its usefulness as more data accumulate. The above constitute mainly areas of future work.

Observations about the analysis

This is the first time that The European Library engages in the analysis of the logs. It has been therefore an exploratory endeavour. It has not been very clear to us since the start what was to be expected from the analysis. This is also why we decided to collaborate with two experts on the field.

UNIPD has met most of the initial aims set in the project proposal and there has been some interesting information that arose from the analysis of the HTTP logs but this information still remains at statistical level. The information provided is useful in the sense that it forms the basis for constructing the user profiles. However, the individual pieces of information at this stage do not add up to forming coherent individual or group profiles. As mentioned before, information like what UNIPD has provided, we are capable of monitoring on a daily basis via web statistical tools like Google Analytics and Awstats, while what we had been looking for was how this information can be combined to form user and group profiles.

That has been a risk in a way in using only HTTP log files that mainly record traffic to the portal. Only after our persistent efforts, UNIPD included in the third version of the final report it provided some analysis of the TEL Action logs. The aim of the project as set in the project proposal was not however fulfilled, i.e. to track the same users in the HTTP and in the TEL Action logs which provide much richer information, and if available, combine the data with data from the registered users database and form an as much as possible complete idea about their actions and about who they are.

¹⁵ MPII delivered two versions of their report. In the first one which was more technical the RankingSVM model was presented while in the second, more simplified report only the way it can be used is mentioned.

It appears that UNIPD starts from the data that can be immediately extracted from the log files while specific questions should be put to provide the context for the analysis. A more clear methodology that will combine information from all three sources of data (http logs, TEL Action logs and data from the registered users' database) has to be developed that will reply to the following issues:

- How can individual user profiles be identified? How should the available information be combined to provide profiles?
- Which user groups can be identified? (e.g. national, academic and research, bouncers and recurrent users, users that spend much time on the portal, etc). What are the common characteristics they have? Are there usage patterns and particular behaviours that are observed for these groups? - How can these individual profiles and group profiles be associated with particular usability developments in the portal?

Such questions should form the basis for parsing, combining and analyzing the information of the logs. Regarding the latter, UNIPD should differentiate between the methodology of parsing and the methodology used in the analysis of the logs.

MPII came up with a very interesting methodology as regards user interaction with the portal and in particular query formulation analysis. The setting of specific aims and questions made it probably easier to filter the data in order to get specific results to the questions set. This led to the proposal of two models that can be used, the first one for automatic query expansion and the second for automatic collection selection.

There is a clear logical procedure followed in the analysis that made it easier for us to understand how the individual elements of analysis interrelate. For example, after investigating query syntax UNIPD concluded that queries are mainly keyword-based. Then they analysed query term correlations to identify semantic relations of particular terms in order to feed this into a process of automatic query expansion. Seeing that collections selection does not work for most of the users and on the other hand the default order of appearance on the portal is not the optimum one, they investigated a way for automatic selection of collections based on the query that would run on the background.

However, we acknowledge that data are still sparse to provide statistically valid observations and therefore findings are preliminary and any recommendations are quite premature.

Conclusions and general recommendations for future work

Some interesting and some unexpected results arose from the analysis like the possibility that log files analysis offers to monitor the use of the Mini Search Box remotely loaded in other websites. An interesting though disappointing observation was that only 17% of the users of the portal that perform a query appear to be satisfied with the results as they go on to viewing a full record. Another interesting result is that more than half of the human visitors to the portal do not effectuate a search.

However, the analysis of the two institutions left us with more questions than answers. Why do users perform only one query? Are they looking for this

particular thing? Are they satisfied with the results? Are they just browsing so they only experiment with the search? Why do the users leave the interface language in English? They don't see the possibility to change language or are they capable enough with the English language? Same for the advanced search. They don't use it because they don't spot this possibility, because they don't understand how it works or just because they choose to ignore it?

The current project helped us explore what we can and what we cannot get from the analysis of the log files as well as indicated the way for systematizing the analysis of the log files from now on. At the beginning of the project our expectations from the analysis were probably too high for the current state of affairs. We did not receive concrete user profiles and specific recommendations for improving the usability of the portal. The project revealed the following reasons which should also be kept in mind in any continuation of the project:

- More data are needed both from the log files and from the registered users. Changes were introduced in the TEL HTTP logs last September and the TEL Action Logs only started to be recorded on the 19th of December 2006. Therefore log files data are too sparse still to produce statistically viable observations and recommendations. More users should be encouraged to log in as well so that more detailed data for their profiles can be gathered.
- Log files cannot possibly reply to all questions about the users. A combination of user surveys and log files analysis is probably the best way to form a clear picture of the motives of users for doing things. In this respect, UNIPD's attempt might be interesting however incomplete.
- A more systematic methodology should be adopted, in particular from UNIPD's side, so as to look beyond statistical data into identifying all the different flavours of the individual users and user groups. Efforts should initially focus in combining as much as possible available information about users. The two institutes should collaborate closer in order to avoid duplication of effort and to combine methodologies. For example the Markov model of user interaction would be interesting to be examined for each particular user group. On the other hand, UNIPD attempted to identify user satisfaction about collections and collection selection while this issue was covered via the logs analysis and the transition probabilities by MPII.
- The most rich and useful information is yet to come from the combination of the HTTP logs and the TEL Action logs together with the information from the registered users database. This will allow us to track individual users, as they return to the portal and over a period of time track their characteristics, their actions in order to be able to provide real personalized services.
- There can be many user groups (country, language, educational background, recurrent users, etc) and users may belong to more than one. Priority for the analysis should be given to the ones that are more important for increasing traffic to the portal such as bouncers and users from countries that recently joined The European Library.

Future actions regarding the analysis of The European Library log files

UNIPD and MPII focused at this stage of the analysis on the web logs and the TEL action logs respectively. These two sets of data allow us to observe different aspects of the users' interaction with the portal. The first ones focus mainly (with some error margins) on the way users approach the portal (traffic), that is which country they come from, from which website they access our portal, what they type in order to access the portal, what language their interface is. From the TEL Action logs we are more informed on the particular actions they perform on the portal: what keywords they enter, which collections they select, which collections they select to see results from, how far they go about viewing results, etc.

TELplus

The effort to analyse The European Library log files is continued via a new project that will most probably start at the beginning of September 2007:TELplus. In this eContentplus funded project The European Library collaborates again with the University of Padua and Max Planck Institute to further developments in the investigation and use of the log files to develop personalization in the portal. In particular, Work Package 5, *User personalisation services – log file analysis and use of annotations*, is led by UNIPD and will focus in the following tasks:

- Analysis of user requirements will explore the requirements of the partner libraries as regards user personalization services and will provide the necessary background knowledge for the other work package tasks
- Log files analysis builds on the current project of the TEL Action Logs and HTTP logs and inserts a new parameter: the analysis of the http logs for search engine optimization purposes.
- Profile building of particular user groups, specification of user models, knowledge extraction techniques
- Personalised search, context-aware query expansion together with the
- Personalised alerting and notification are two specific personalised services that will be developed.

Most of the research as regards the analysis of the log files is taken over by the two partner institutes in TELplus WP5. As regards continuation of work within edlproject and until February 2008 that the current task ends the following actions will be realized:

- It will be proposed that EDLproject countries provide the basis for analysis of user groups in TELplus and in particular for the analysis of "national" provenance for the development of localisation services
- Special attention will be given so that the results from TELplus WP5 developments are communicated as best possible to the EDLproject partners.
- A different ranking of the collections will be considered to match the observations of MPII.
- Recommendations provided particularly by UNIPD will be implemented to attract more registered users, like the development of new services and the addition of the username and password on the main pages of the portal.

Part C: PROMOTING OAI-PMH

Aims of the task

The aim of this action is to outline the benefits of OAI-PMH to EDLproject as well as to the other The European Library partners in order to persuade more of them to make their metadata available for harvest via this protocol.

What is OAI-PMH

OAI-Protocol for Metadata Harvesting¹⁶ (OAI-PMH) defines a mechanism for harvesting records containing metadata from repositories. It gives a simple technical option for data providers to make their metadata available to services, based on the open standards HTTP (Hypertext Transport Protocol) and XML (Extensible Markup Language). The metadata that is harvested may be in any format that is agreed by a community (or by any discrete set of data and service providers), although unqualified Dublin Core is specified to provide a basic level of interoperability. Thus, metadata from many sources can be gathered together in one database, and services can be provided based on this centrally harvested, or "aggregated" data.

The OAI protocol does not define the link between this metadata and the related content. OAI-PMH does not provide a search across this data, it simply makes it possible to bring the data together in one place. In order to provide services, the harvesting approach must be combined with other mechanisms.

Much promise is seen for the use of the protocol within an open archives approach. Support for a new pattern for scholarly communication is the most publicised potential benefit. Perhaps most readily achievable are the goals of surfacing 'hidden resources' and low cost interoperability. Although the OAI-PMH is technically very simple, building coherent services that meet user requirements remains complex. The OAI-PMH protocol could become part of the infrastructure of the Web, as taken-for-granted as the HTTP protocol now is, if a combination of its relative simplicity and proven success by early implementers in a service context leads to widespread uptake by research organisations, publishers, and "memory organisations".

Why is The European Library promoting OAI-PMH?

The European Library Technical Architecture bottleneck

The European Library is promoting the adoption of OAI-PMH to its partners for the important usability functions that the protocol facilitates. The adoption of OAI-PMH will help circumvent technical restraints that impose limits to any serious usability developments in the portal.

¹⁶ Information in this section is collated from the official website of OAI-PMH: <http://www.openarchives.org/pmh/>

The European Library is a portal that offers distributed access to the collections that are locally hosted by its partner libraries. The partner libraries are currently 30 and provide access to 290 collections out of which 171 are searchable while the rest are browse-only, that is, the user follows a URL from the TEL portal to the native interface of the collection.

For querying the distributed resources held remotely in the national libraries, The European Library is currently structured around a hybrid combination of communication protocols (mainly Z39:50 and SRU).

SRU is used for querying the distributed record repositories. In the case of Z39:50 querying is intermediated by two kinds of gateways: The Central Gateway (hosted in the British Library) or a gateway located locally at the libraries' servers.

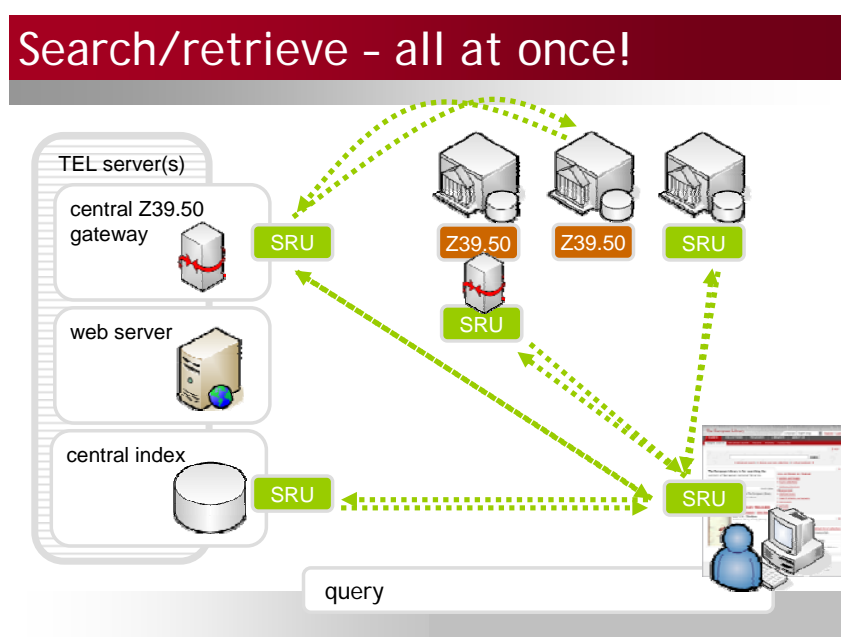


Figure 15: Graphical representation of communication protocols in TEL

This architecture was devised in order to keep a low barrier for access to the collections of the partner libraries using the different communication protocols that partner libraries already had in place. However, its efficiency lies on one precondition: the pre-selection of collections by the user. That is because it is impossible to query simultaneously more than a dozen collections and receive results within a reasonable time lapse. Therefore, the user has to narrow his selection into a handful of collections if he wants to receive quick and satisfactory results.

There are some other shortcomings with the federated model. Because the results are transmitted from distributed resources, there can be practically very little control over them, that is, there are very limited possibilities for manipulating the information before its display. For example, the results are displayed by the order of the response of the servers of the libraries. Currently it is not possible to apply any method that eliminates duplicate results, enables ranking based on the relevance of the results with the search or makes it

possible to translate results as they come in. Due to these constraints, usability regarding search and display of results is currently very limited.

The above mentioned protocols are meant for access to metadata. Access to objects and related services is done via the URL in the identifier field or via an intermediate Open URL service. Services developed and added to the portal complement the usability of the search results and are a separate area of research and development within The European Library.

The European Library is systematically investigating user interaction with the portal to understand users' search habits and preferences and identify any problems they might encounter. It has been clearly pointed from past user surveys of 2005 and 2006 that the users do not understand the need for selecting collections and they do not want or do that. The European Library is investigating ways to overcome this problem.

The initial solution that was implemented was the supply of a default list of collections for search. Every library has selected one collection which forms part of the default list of collections. This is usually the main national library catalogue. If settings are not changed by the user, then the query is realised among these default collections. It is obvious again from user studies that the majority of the users do not change the selection of collections. As Julia Luxenburger pointed out in her report, queries are keyword-based and they are performed on the default list of collections. Users' expectation is to receive the most relevant results at one go much as they do on Google and other powerful search engines.

Acknowledging the need to build fast and integrated query to meet the search requirements of its users, The European Library is building and expanding the Central Index, which in combination with a repository harvests and indexes metadata from the records of the partner libraries. OAI-PMH is used to harvest the metadata.

Hosting the metadata centrally and enabling queries to be performed on the Central Index offers significant advantages compared to performing distributed search for a number of reasons that we will analyse further down. In short, it makes search much faster, it gives much more relevant results and facilitates the manipulation of metadata, that is, it makes it possible to combine multilingual tools, ranking and clustering options that enhance user experience.

Benefits of implementing OAI-PMH

The protocol separates data providers from service providers. Data providers deploy an OAI-PMH compliant repository (also called archive). Service providers deploy an OAI-PMH harvester. The protocol supports multiple metadata formats: MARC, DC records in XML, DC Qualified, MODS TEL Application profile. Records should be exported in XML.

In our case, data providers are the libraries that hold the metadata and that make them available for harvesting and search by search engines and other service providers. The European Library acts both as a service provider and as a

data provider as it provides the central index for harvest by external search engines in order to enhance the paths to the content for the users.

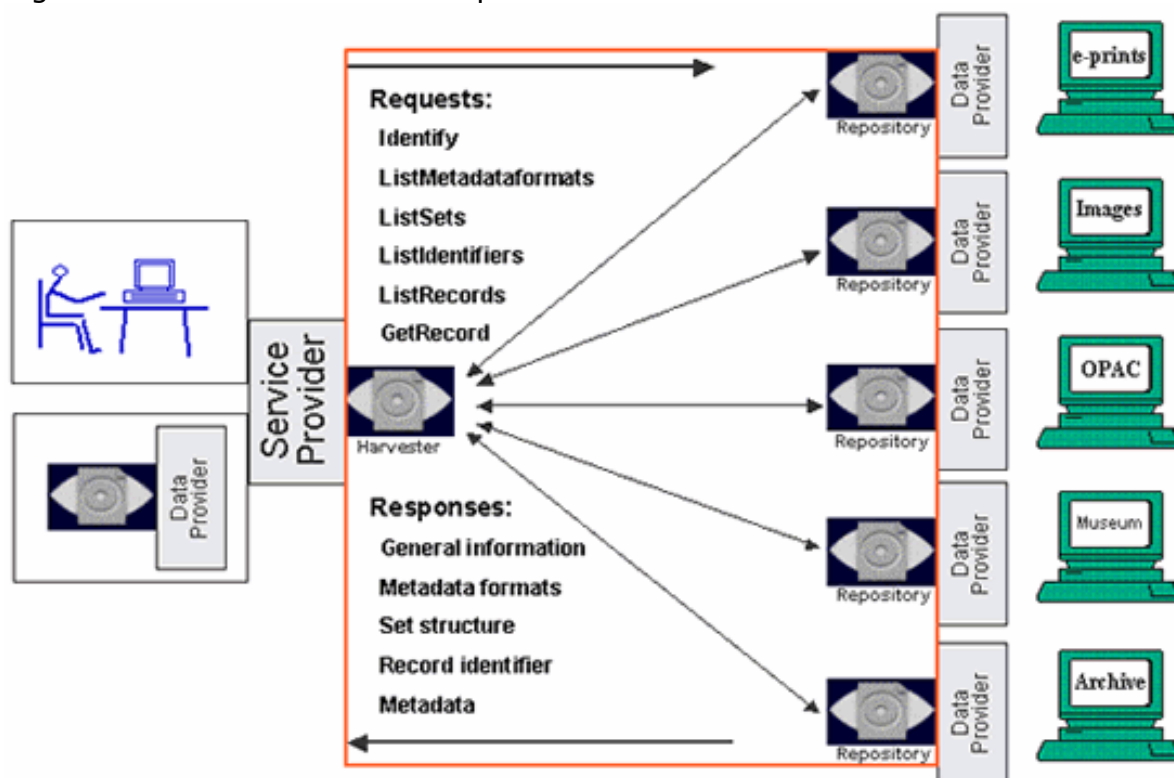


Figure 16: OAI: Overview and structure Model

Collecting and indexing the metadata in one central repository provides tremendous benefits for the search of materials and for enhancing the overall search experience of the user:

- Search becomes much faster as queries only address one target.
- No longer necessary to select collections as all harvested records are individual items under one "super-collection" hosted in the central index. The central Index reduces the number of clicks for the user to get to the content.
- Results are available 24/7. Potential unavailability or remote system malfunction that spoils the results is avoided since metadata are harvested in advance at regular times.
- The database in combination with the central index makes it possible to apply on-the-fly manipulation of results. It is possible to search everything at once and sort results by relevance, popularity, or page ranking, e.g. google experience in TEL on better data.
- It also makes it possible to do back-end manipulation of data with the use of multilingual tools such as name and subject authority records and thesauri, apply controlled vocabularies and experiment with advanced filtering and clustering with CIDOC-CRM and FRBR
- Such a rich pool of data makes it possible to experiment as well with new services. For example, it makes it easier to extract in an automatic or semiautomatic way data from particular fields in order to create virtual collections (e.g. Slovenian 19th century poetry).
- It facilitates user personalization services such as sending automatic feeds for new materials. Using the capabilities of the Verity search engine, the TEL portal can do on-the-fly background searches in the Central Index and

suggest to the end-user potentially interesting objects or related collections the user did not actively search for.

Besides the faster and more quality results it offers to the users, OAI-PMH and the expansion of the Central Index can be very useful for the partner libraries:

- It helps better monitor users' search habits. As we analysed in the previous part of the report, integrated search makes log files analysis more precise which in turn can give essential feedback on usage behaviour. This can help the libraries to design and provide better user services as well as prioritise on what to digitise.
- It facilitates the collaboration between libraries: Metadata integration on a central index may speed up and facilitate a range of activities across libraries e.g. locating materials of interest, harmonising, evaluating and normalising cataloguing entries, coordinating preservation and advancing research on issues such as name authorities mappings, subject translations, etc
- Additionally, The European Library Central Index can be integrated as an individual target in the local portal systems of each library. Thus, the contents of the national libraries of Europe become more integrated in the everyday working environments of thousands of users.

The Central Index is a major asset and a unique selling point for The European Library and provides added value for the partner libraries. It facilitates services that were not possible before for the individual libraries.

Issues

OAI-PMH has been slow in adoption, mainly because it is a relatively new protocol while Z39:50 is a long and well established protocol in the library world. Z39:50 supports all the rich metadata formats of the libraries which is another reason it is so much appreciated by libraries. Additionally, it allows them to keep full control of the metadata which has been a sensitive issue for the libraries. Sometimes, like in the case of the British Library, metadata records are considered an asset worth a fortune which has also been an object for sale. There has been explicit fear that exposing the metadata to external providers, libraries might lose the intellectual and editorial control that has traditionally maintained the quality standards of the records high.

The answer to these fears is that exposing the metadata is a trade-off that can bring additional benefits to the libraries for all the reasons stated above. It has to be understood that the metadata have to be exposed in order to enhance the possibility for more people to reach the so far hidden resources and of course to be able to track and eventually get to the original object.

Actions for the promotion of OAI-PMH

The European Library has tried to persuade the partner libraries to adopt OAI-PMH and enable the harvest of their metadata. The issue has been addressed with the following actions: explaining the use and implementation of the protocol and presenting the possible implementation scenarios, advocating the benefits,

developing an installer to facilitate the implementation of the protocol and addressing any issues regarding implementation on an individual basis.

A. Promotional actions

There has been a systematic effort to inform partners about the benefits of implementing OAI-PMH in order to persuade them to adopt the protocol. Below are listed some of these actions:

- The importance of implementing OAI-PMH has been highlighted in the Requirements Analysis Questionnaire, the document that partner libraries fill in at the start of the project where they supply information about the collections they will provide access to, the communication protocols via which they will provide access, etc. The Requirements Analysis Questionnaires were sent to the partners in September 2007.
- The author of the current report also prepared a sales document entitled Open Archives Initiative: Harvesting More Satisfied Users Internationally! The document outlines in an A4 page the benefits & promotes the use of OAI-PMH among partners. The document was created in November 2006, distributed to the partners and uploaded on the EDLproject website.
- Aubery Escande of the TEL Office created a page on the Handbook¹⁷ presenting the OAI installer that was developed by the TEL Office. An e-mail with information about the installer was sent out to all the partners on the 18th of May.

B. Implementation and Support Actions

Besides the promotional activities highlighting the benefits of OAI-PMH, specific actions were dedicated to helping partners implement the protocol.

On the 11 and 12 of January 2007, the joint Technical Working Group and EDLproject partners Knowledge Sharing Workshop was organised in The Hague. During the two days meeting several presentations regarding the technical and organisational as well as issues specific to the project were organised. 17 participants from 9 libraries participated in the meeting. During the WP1 Knowledge Sharing Workshop on 11+12 January 2007 in the KB, OAI-PMH has been promoted to the participants. The main aim during the KSW was to demonstrate participants that setting up OAI can be quite easy in some cases.

A particular workshop led by Nuno Freire and Sjoerd Siebinga was devoted to promoting OAI-PMH focusing on the implementation requirements by the partners. The general feeling was that OAI-PMH was easier to implement than initially thought and some of the libraries were willing to implement it in the course of the project.

Three scenarios of implementation were presented: A Library Management System (LMS) OAI- module provided by the vendor, In-house OAI-PMH server development and Standalone OAI server.

The following recommendations were produced at the workshop:

- Every library needs to implement OAI-PMH
- Form a TEL-lobby group to ask all vendors to include OAI-PMH modules

¹⁷ <http://www.theeuropeanlibrary.org/handbook/form.php?ref=handbook.php> (enter as guest)

- Ask vendor Aleph (Ex Libris) for TEL Application Profile cross-walk in OAI-PMH module
- Check if the library management systems (LMS) support deleted records
- Libraries should check if they actually own the metadata
- Lowest 'barrier of entry' is with regular full harvest instead of incremental harvest with deleted records and timestamps
- Libraries should make sure that their LMS has persistent identifiers
- Use OAI identifiers as persistent identifiers for deep-linking
- Consult ICT-department about how a server can be set-up

The materials of the workshop and the outcomes were made available on line in the EDLproject website in the partners' password protected area.

The European Library OAI-PMH java installer

During the workshop the main requirements for each implementation scenario were presented and partner libraries are encouraged to adopt any implementation method. Partners can choose any open source standalone OAI-PMH server, install and configure the server. Partners should then develop metadata crosswalk specifications for their format to the TEL Application Profile as well as implement a procedure in the library to transfer the metadata records from the LMS to the OAI-PMH server. However, with a competent programmer even this scenario should take a few hours.

Based also on the outcomes of the the EDLproject KSW, Sjoerd Siebinga from TEL office developed a platform independent installation package for a (zero configuration) stand-alone OAI server. This significantly lowers the implementation barriers of OAI for some collections, provided that the records are in MARC(XML) format. A weekly dump from the Library's Content Management system can be converted to MARCXML and be instantly made available via OAI. The OAI server comes with a full web-based intuitive graphical user interface that allows for easy definition of multiple collections/sets and metadata-formats.

To deploy the installer, partners must be able to export the metadata records from the Library Management System to separate files (ideally with the identifier as file name). If they have more than one collection, each collection needs to be exported to a different folder. They must be able to export only updated, created, and deleted records periodically. Finally, records must be exported to xml.

To further assist partners and facilitate the process of implementation, Sjoerd Sienbinga from the TEL Office has undertaken to configure the installer the support the particular metadata formats of the partners that meet the requirements set above. Partners are encouraged to send between 100 and 1000 metadata records (both in the native and the xml format) at The European Library Office that then makes a crosswalk from the metadata xml format to TEL application profile and oai_dc, and includes it in the OAI-PMH installer.

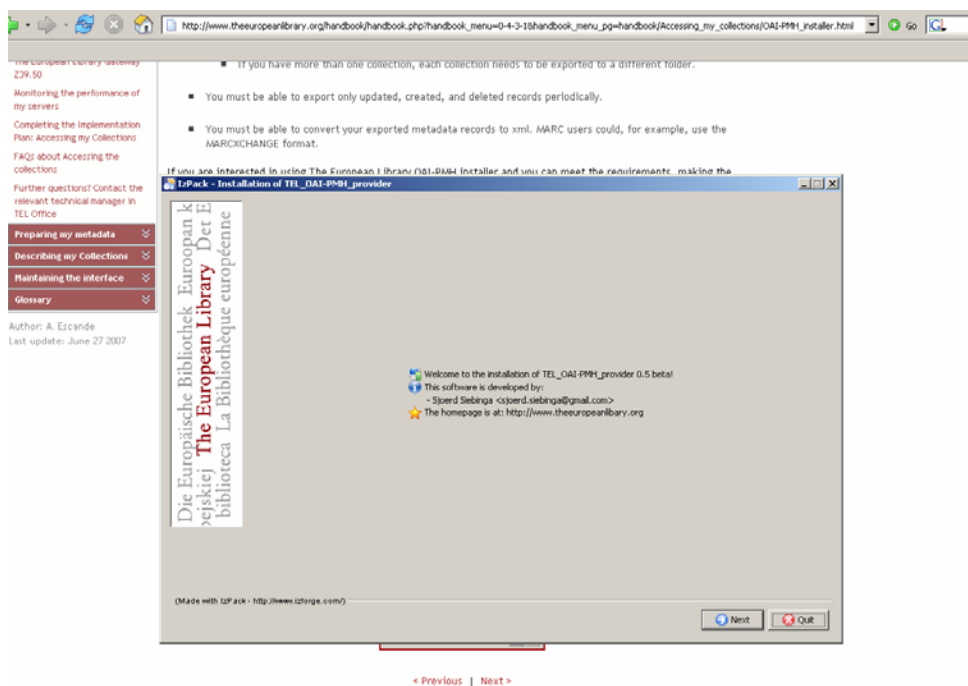


Figure 17: Pict: interface of the TEL OAI-PMH java installer

Further assistance and promotion

During the regular quarterly teleconferences with all TEL Full Partners, Olaf Janssen has been promoting the use of the stand-alone OAI-server TEL Office is offering to its Full Partners for free. There has been quite some interest in this software, although none of the partners has actually implemented it yet. Some relevant notes in relation to this.

- Estonia tried to install the OAI-server but has experienced some problem with the installation. The error is currently being investigated.
- Slovenia has shown interest to test it and, if successful, use the product
- Spain did not have the correct version of java yet installed and couldn't run the installer but they are in the process of obtaining it.
- Liechtenstein would first favour the implementation of the Aleph OAI-module. This can happen once development work on this has been finished by the national library of Austria. If this takes too long or if implementing it turns out to be problematic, Liechtenstein will be happy to install the stand-alone OAI-server provided by TEL Office.
- The communication with all these partners has been interesting for TEL Office to identify any bugs in the OAI-server.

Overall Outcomes:

# Searchable collections by protocol	Number	Proportion
Total	171	100%
SRU	77	46%
Z39.50	36	22%

OAI-PMH	58	34%
---------	----	-----

Figure 18: Number of searchable collections: by protocol (as of 13.08.2007)

Countries that implement OAI-PMH: Currently only Portugal and Czech Republic use OAI-PMH for all their collections. Poland uses it for 11 out of 12 of its' collections. Italy, Slovenia, BnF (Gallica), Serbia, Hungary and Estonia use it for some of their collections.

Of the 50 collections that are expected to be added in The European Library portal until February 2008, 44 will be OAI-PMH compatible and only 6 will be available via Z39:50. This is a very good ratio. Among the 9 EDLproject libraries Belgium, Iceland, Lichtenstein, Norway and Spain will first make their contents available via another access protocol and at a later point will switch to OAI-PMH. We consider that the TWG and EDLproject knowledge sharing workshop had direct effect on this.

Following the successful efforts of the Austrian National Library Ex Libris (<http://www.exlibrisgroup.com/>) has created an OAI-module for Version 16 of Aleph. They are willing to include TEL Application Profile as a default profile in this module. Olaf Janssen is carrying out the coordination between ONB, Aleph and the 8 other European national libraries that are currently using the Aleph software (Luxembourg, Liechtenstein, Sweden, Iceland, Latvia, Czech Republic, Denmark, UK).

Further steps regarding the promotion of OAI-PMH

- TEL Office will continue to promote to Full Partners the use of the stand-alone OAI-server or any other OAI method via the teleconferences held regularly with them. Any outstanding issues regarding implementation will be addressed on an individual basis.
- During the next Technical Working Group of The European Library a session will be devoted to OAI developments, including the stand-alone OAI-server. Additionally, the proposal from the BnF to test a combined method of metadata harvesting via OAI with a link included to the larger and richer Z39:50 record will be investigated. The idea for this initiative is to leverage the advantages of both protocols, that is, the simplicity of OAI and the easiness in sharing metadata as well as the possibility to offer richer information by linking to the original record via Z39:50. The intent during the workshop is to keep the interest of partners alive and to push further the implementation of OAI-PMH.
- Within the eContentplus funded TELplus project which is likely to start in September 2007, WP2 will focus entirely on the implementation of OAI-PMH for the following libraries: Estonia, Slovenia, Latvia, Hungary, Portugal, France, Poland, Slovakia and Iceland. The work package also includes a workshop on best practices, guidelines and tools for the partner libraries. In particular the work package aims at the development of alternative solutions for the libraries to implement what most easily and best suits their needs:
 - Development of a system that will make it possible to install local metadata repositories comprising a local OAI-PMH server and

components necessary to manage and interface it with the local databases

- Development of an OAI-PMH central service that will manage all the synchronizations in a network of OAI-PMH servers, and will provide a central metadata repository service to be used by other functionalities of the TEL portal.
- Analysis of the ways SRU and Z39:50 targets can be harvested and production of guidelines for implementers

Part D: Annexes

Annex 1: Description of the TEL Action Log Files Schema

The European Library - user tracking/logging capabilities

Eric van der Meulen
Created: Nov. 6, 2006
Last revision: April 26, 2007

The following tables describe the elements being logged with the TEL custom logging functionality.

Table 1.1 - Logged values

field name	value	description
userid	string	“guest”: user not logged in
	integer	registered user id if user is registered and is logged in
userip	string	User i.p. address
sesid	string	Php session id created on session-start
lang	string	Two letter language code – based on language view in portal
query	string	Query text
action	string	Depending on the action the user has performed several action phrases are logged: See table 1.2
colid	string	Collection id that action is performed upon
nrRecords	integer	Total number of retrieved records per collection
recordPosition	integer	Position of viewed item in the total record list
sboxid	string	identifier for remote searchboxes which query the portal via url
objurl	string	URL of object being viewed. Relevant actions: available_at, see_online (a url to the remote object), view_full (url of the record within the TEL portal)
date	datetime	Timestamp yy-mm-dd hh:mm:ss

Table1.2 - Action vocabulary

action value	description
search_sim	Search initiated simple search form
search_adv	Search initiated from advanced search form
search_res	Search initiated from search form on the results page
search_res_rec_any search_res_rec_all	Search initiated from within a full record view by clicking on search(magnifying glass) icon in the record’s available fields
search_url	Search initiated from url query string This string may also have a domain name attached to it (search_url_www.domain.org) if it is coming from a remote tel search – minitel (a marketing tool)
view_brief	Short title display – list
view_full	Long title display – individual record Activated when a user clicks on a title link in the list of brief records displayed (20 per page), or when a user clicks on the <i>previous</i> or <i>next</i> link when already viewing a full record
jump_to_page	In brief title display user can enter a numerical value for

	skipping several pages of records
available_at	“Available at Library” link clicked to view record in native interface
see_online	“See online” link clicked to see object in native interface
page_brief	“next” or “previous” links clicked to page through brief record lists (20 per page)
col_set_theme	Collections chosen from theme list
col_set_theme_country	Collections chosen from country list on homepage or resultspage
col_set_country	Collections chosen from all collections tab (collections listed by country)
col_set_subj	Collections chosen from subject list
col_set_desc	Collections chosen by searching by description
col_set_default	Collections default list reinstated
option_save_session_favorite	Session favorite saved
option_send_mail	Record sent by email
options_save_reference	Record saved for reference manager use
service_denmark	full record service link used
service_hungary	full record service link used
service_netherlands	full record service link used
service_uk	full record service link used
service_all	full record service link used
show_help_helpfilename	“help” link clicked

Annex 2: Final Reports from the University of Padua and Max Planck Institute

Julia Luxenburger, Gerhard Weikum, A User-Interaction Model for The European Library Portal, Final Report (<http://www.edlproject.eu/membersonly/downloads/tel-report-less-technical.pdf>)

Maristella Agosti, Giorgio Maria Di Nunzio, Tullio Coppotelli, The European Library Project on “Analysis of The European Library Logfiles” (<http://www.edlproject.eu/membersonly/downloads/UNIPD-final-report-with-recommendations.pdf>)

Both reports are available on the EDLproject website (www.edlproject.eu) in the password-protected area intended for access by the partners of the project. You can request for the password and username of the site from Georgia.angelaki@kb.nl.