



Title: Research and Innovation in The European Library; the TELplus Project

Georgia Angelaki

The European Library Office, The Hague, The Netherlands

Email address: Georgia.angelaki@kb.nl

Date: October 2008

Abstract

The paper presents the progress of the TELplus project, a primarily research and development project of The European Library which is the shared online service of the 48 National Libraries of the Council of Europe. The project is composed of several areas of work each of which aims at either significantly increasing access to digital content or at enhancing the usability of The European Library portal. The main work areas are: Optical Character Recognition (OCR) of scanned texts; full text search with the use of semantic web techniques; building a centrally managed infrastructure of distributed repositories with the use of the Open Archives Initiative-Protocol for Metadata Harvesting (OAI-PMH); setting up a metadata repository based on the Functional Requirements for Bibliographic Records (FRBR) concept model; building a services infrastructure and investigating user profiles for providing personalised services.

Key subject words: The European Library, OCR, services, OAI-PMH, semantic web, FRBR, personalisation, national libraries

Introduction

The European Library is a service of the Conference of the European National Librarians¹ (CENL). It was launched in 2005 to offer a common access point to the distributed collections of the 48 national libraries of the Council of Europe² and of the State of the Vatican City. It evolved from a project providing general information about the libraries (project Gabriel³) to

providing an integrated point of access to their online resources. In 2005 there were 9 founding members⁴ and at the beginning of 2009 there will be 46 national libraries who are full participants in The European Library providing online access to more than 330 collections.

The mission of The European Library is described as follows: “*The European Library exists to open up the universe of knowledge, information and cultures of all Europe's national libraries*”. With the integration of the remaining European national libraries, an important goal of The European Library’s mission is reached. In the near future, therefore, the weight of The European Library strategy will lie on investing resources to improve the service through innovation. This will be achieved by improving the multilingual search and retrieval of results, providing more digital content and services to users that better fit their particular interests, and improving the overall search and browsing experience of the portal.

TELplus⁵ is a project of CENL that is being taken forward by a consortium of 26 partners, national libraries and research centres to form a pool of research and innovation to provide value-adding services and products for The European Library. It is coordinated by the national library of Estonia⁶ and Eremo s.r.l.⁷ and it is partly funded by the European Commission Programme eContentplus⁸. The project started in October 2007 and will run for 27 months.

The project provides a major research and innovation hub for The European Library. Each work package aims at either significantly raising the amount of digital content in The European Library, or improving access to content and the overall usability of the service. The current paper presents the developments achieved within the project’s main technical work packages during its first year and the expected impact on the service.

The structure of the paper follows the structure of the project’s Work Packages (WP1-WP5)⁹. Each of the following sections describes a work package in progress in the project and its effects on The European Library. The first work package concerns turning into searchable previously scanned images using OCR (Optical Character Recognition¹⁰) from 14 national libraries and the related pilot of The European Library that initiates a long-term full-text indexing strategy of such content. The second work package describes a system supporting a robust and scalable harvesting infrastructure made up of centrally managed distributed OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting¹¹) repositories. Research described in the third work package involves a prototype for a semantic search engine that meets the need for full text query and visualisation, research on techniques to apply semi-automatic subject heading alignments, and the initial investigations for setting up an FRBR (Functional Requirements for Bibliographic Records¹²) repository with the harvested metadata records from the national libraries. There are three sub-areas of work in this work package which are connected with the overall goal of improving access to content putting a special emphasis on multilingual and semantic functionalities. Work Package 4 is dedicated to service infrastructure and the registry being built which is based on the concept of standardising service descriptions. The final work package described talks about the research conducted on user profiles for the provision of personalised services.

Work Package 1: OCRing previously digitised material

While digitisation is not one of the core activities of The European Library itself, existing user satisfaction studies indicate that the success of the service is closely connected with the

supply of digital content to users. The first work package is important for the great amount of digital content it helps the national libraries to produce and bring to The European Library, and for raising awareness about best practices in OCRing - a very costly enterprise for the national libraries and one that is still largely undocumented.

In TELplus Work Package 1 14¹³ national libraries use OCR to turn into searchable full text a total of more than 20 million pages that previously existed only as digital images. The produced full text will be used for detailed searching in the partners' digital libraries and the content will also be accessible via The European Library.

The first survey of the work package revealed the kind of material that was going to be OCRed in the context of the project. More than 50% of the content is newspapers and the rest is mostly books, journals and pamphlets. The national library of Slovenia contributes the oldest book, which dates back from 1500, while the youngest Icelandic newspaper is only 6 years old. The majority of the material comes from the period between the early 19th to the mid 20th century. The geographical coverage of the content extends from the Mediterranean to Greenland and to Latin America. There are documents in more than 30 languages, including Faroese, Greenlandic, Yiddish and Latin and in some cases there is content published in more than one language in the same document. All these raise challenges for the OCRing tools which are not yet developed enough to support minor languages, let alone historical fonts in these languages.

The survey also revealed that national libraries have very different experiences with OCRing: some partners had well established and in some cases mass digitisation routines in place while others were starting to OCR for the first time. At a meeting held at the Bibliothèque Nationale de France¹⁴ in January 2008 national library representatives shared experiences regarding content selection criteria for digitisation and OCRing, quality of OCRing and how to measure it, the problems of in-house OCRing, tendering procedures and prices, software tools and output formats suitable for various uses. This information sharing exercise proved valuable for partners as such information is scarce and there is hardly any literature on best practices in the topic in English¹⁵. The continuation of the work package helps raise awareness about different practices, helps address the gaps in language aids and software tools for minority languages and historical fonts¹⁶, and also helps promote the adoption of standards in the area that will ensure quality results for long term access to digitised content.

At the same time, The European Library conducted a pilot between March and August 2008 for retrieving and indexing some of the partners' sample OCRed materials in order to explore the issues that arise regarding provision of full text search functionality through The European Library. Though born digital material is not expected to raise serious difficulties for indexing, quality of OCRed material is definitely a concern for The European Library and future user studies will need to address the expectations and the tolerance of users in regards to errors in order to shape an appropriate indexing approach. With respect to building services on top of full text, initial efforts to map this issue will be presented in Work Package 3 that is in particular investigating improving full text search and browsing.

In the near future The European Library will work towards setting up a standardised, robust and scalable process for retrieving and indexing mass digitised and born-digital content from national library partners. The pilot showed that the different OCRing routines of the partners result in a variety of output formats and accessibility complications that cannot, in the long term, be addressed on an individual library-by-library basis. Building on the benefits of a

standardised process of providing access to the collections' metadata so far, The European Library will seek to work with the national library partners on an agreement regarding the standardisation of the process of retrieving full-text and on the selection of a limited number of full-text formats that best fit the quality of, and long-term access to, full-text digital content.

Work Package 2: Improving the usability of The European Library through OAI-PMH compliancy

The second work package is aimed at supporting the efforts of The European Library to create a central repository of harvested metadata from the national libraries. This is crucial for providing a fast search and also allows for a number of valuable functionalities and services to be built on top. The work package helps, first, to significantly lower the barrier for national libraries to adopt the OAI-PMH protocol for exposing their metadata and, second, to create a centrally managed infrastructure of distributed metadata repositories specifically adapted to the requirements of The European Library.

The European Library currently uses SRU¹⁷ to query both SRU and Z39.50¹⁸ remote targets and a central repository of indexed metadata from harvested partners' OAI-PMH repositories¹⁹. This federated searching imposes a number of limitations in the search and retrieval of results such as slow response time and a restriction on manipulation that would return more relevant results to the users (e.g. ranking).

The national libraries are gradually realising the importance of moving away from Z39.50 and of (partially) exposing their metadata to various search engines and to The European Library. Metadata harvesting libraries means The European Library can create a central pool of valuable multilingual information that can be manipulated in ways which are expected to provide true added value functionality to users: data can be semantically enriched with thesauri, authority files and other language resources; translation efforts are facilitated; de-duplication of results is enabled, as well as clustering of results; and the possibility to implement FRBR for searching and browsing the records. The European Library aggregated metadata content can be exposed better on web2.0²⁰ applications such as the mini search widget developed by The European Library²¹ which can be placed in other websites.

In this context, the Instituto Superior Técnico (IST)²² in Lisbon, Portugal, is setting up an integrated harvesting infrastructure for The European Library made of two components: one that is client-based intended for the national libraries and one that is installed centrally in The European Library. The system is called REPOX²³ and is designed to meet the particular requirements of The European Library. The local module is a stand-alone java server that includes an easy to use interface for managing the data and the functions associated with the repository. The module is further designed to do character set conversion and to map the partners' metadata formats to The European Library's common metadata format, the TEL Application Profile²⁴.

For The European Library the central harvesting component will be designed to meet the specific needs of the partner libraries' repositories. The system will make it possible to schedule harvesting according to the update rhythm of a particular collection, and also enable notifications to be sent in case a server is down. In case of this, the system will resume harvesting when the server is running again. Work on mappings and character set conversion will also be integrated in the central component. The work package is expected to harvest

more than 30 million records before the end of the project and make them available to The European Library.

Another important expected development is the ability to harvest Z39.50-enabled servers. This will make it possible to work around the restrictions imposed by some library systems that restrict control over the metadata and prohibit its export out of the local databases into OAI-PMH repositories.

Work Package 3: Improving accessibility

Work Package 3 is a multifaceted work package combining innovative research for The European Library in various areas of work under the common goal of improving access to content and enhancing its semantic and multilingual potential.

The subtasks of Work Package 3 are the following:

Task 1- Improving access to full text

Going a step further than Work Package 1, which creates mass full text content for The European Library, this task investigates the user requirements with respect to searching and browsing full-text and to create a small scale prototype of a search engine specifically adapted to the particularities of OCR'd full text as described in Work Package 1 and able to provide enhanced semantic and multilingual functionalities to the users.

The team from the French national library, who are leading the work package, first investigated the international and French market of commercial and open source tools and developed criteria to evaluate their semantic and multilingual functionalities. Some of the functionalities²⁵ that were analysed included the translation of query and result lists, disambiguation of requests, text analysis, autosuggestions and references extraction.

Subsequently, focus groups brought up the priorities for the prototype development and fed into the public call for tenders for the search engine that was published in October 2008. The prototype will be developed from November 2008 to June 2009 and will focus on extraction of names and keywords and on query disambiguation.

An interesting observation that arose from the focus groups was that although users found translation functions very interesting and liked the possibility to perform a query in their native language and retrieve translated results from all other possible languages, they were suspicious to what extent this was possible. Additional user tests and investigations are necessary to shed further light on this area in order to better understand user expectations and what is possible with existing translation tools.

Task 2- Improving multilingual access by investigating automatic subject heading alignments

Alignments of subject headings and controlled vocabularies can be used to improve multilingual access to content. However, manual mappings are very time and labour intensive²⁶. The current task investigates the feasibility of applying semi-automatic alignment techniques for mapping between vocabularies. The research is being conducted by the STITCH team²⁷ that is investigating semantic interoperability of controlled vocabularies in the cultural heritage sector and ontology mappings to enhance the search results. STITCH is participating in the formulation of the SKOS model²⁸ for the formal representation of such

vocabularies in the context of the Semantic Web Activity of the World Wide Web Consortium²⁹.

So far, the Rameau³⁰ subject headings from the French national library, SWD³¹ from the German national library & LCSH³² subject headings from the Library of Congress were converted into SKOS, the standard semantic web representation language for the purpose of the investigations. This development facilitates their use in various contexts such as the prototype of the semantic search engine in the context of Task 1 described previously.

Task 3- Setting up a FRBR repository for The European Library

Using the harvested bibliographic records from partner libraries, the next task aims at improving search and browsing of content through the implementation of Functional Requirements for Bibliographic Records (FRBR). The FRBR conceptual model uses as its basis the four user tasks: “find”, “identify”, “select”, and “obtain”, and tries to combine information from records that refer to different translations, editions or different media type variations of the same work. This user-centered perspective on cataloguing can have a tremendous impact on usability as it can significantly narrow down search results by clustering similar records for the user. The vast amount of multilingual metadata records that come from the various national libraries is very challenging in this respect, especially because there is a lot of duplication of information in the incoming records. FRBR is expected also to be beneficial for the libraries in terms of sharing information and improving existing catalogues.

However, implementation is not easy and there are not many tools supporting it - consequently the model is not widely adopted yet. The first objective of the project, therefore, was to review existing implementation instances³³ and to invite experts in the field to consult with The European Library and to advise on potential implementation scenarios.

The international FRBR workshop³⁴ that was held in the national library of Portugal in October laid the foundations for a prototype experimentation starting with some “super works” such as the Bible or books by Nobel Prize winners, records of which are likely to exist in various libraries and in many languages. The REPOX system that is being built within Work Package 2 will be used, wherever possible, to harvest MARC records from national libraries as these records are much more suitable for FRBRisation than Dublin Core. The international group of experts from libraries and the commercial sector formed at the meeting agreed to follow up on the project. A prototype FRBR repository is expected to be ready in June 2009.

Section 4: Integrating services with the European Library portal

Work Package 4 aims at creating a modular service infrastructure for The European Library and a service registry that would be based on standardised service descriptions. The added value of the standardisation of service descriptions is that services would not need to be hard coded in the portal and the user should be able to easily invoke and use functionality that exists on the web, irrespective of where it resides, on the condition that legal and access restrictions are overcome. Seemingly, creating context-sensitive services means that specific fields in the result of a query could trigger new functions for the user relevant to his query.

Additionally, in the context of the project a number of new services are being developed by the partners and some existing ones are being analysed as use cases in order to build the

services description registry and data-model. The new services to be built are: image, video and text annotation tools; a tool for displaying old maps on Google maps³⁵ using a timeline; thesauri and name authority services; image zoom tool; an asynchronous proxy service; and a text analyser tool.

A test version of The European Library portal has been used to investigate the conditions that need to apply in order to trigger the invoking of the services, to testbed the services and to instruct the portal how to use them. This has given significant insight into the service model. An article about the first version of the service model schema will be published at the January/ February issue of D-Lib Magazine³⁶ by Theo van Veen from the national library of the Netherlands who is leading the work package.

Over the coming year the service model and the services under development are expected to be finalised, and potential integration in The European Library portal will be planned. The research realised in the service descriptions standardisation area might have a great impact not only for The European Library but for any service-providing site. A thorough analysis of the legal aspects that might arise from the use of services from third parties is outside the scope of the current project. However, should the service description model gain broad acknowledgement and uptake the legal framework surrounding such developments will need to be addressed as well in due time.

Work Package 5: User personalisation services – log file analysis and use of annotations

Research realised in the last technical work package of the project on the profiles, needs and usage habits of The European Library users is of primary importance for The European Library. The depth of this analysis can provide a rich source of information that can feed into improving the overall usability of the portal, creating particular services for specific user groups, offering true personalised search and retrieval according to a user's individual profile, and even informing the overall policy of The European Library and the national libraries in areas of strategic importance such as multilinguality, collections' development policy and digitisation.

The research team led by the University of Padua³⁷ analyses the various sources of information from and about The European Library's users in order to build a structured framework for the systematic recording and analysis of the data. There are various sources of information about the users of The European Library, both implicit and explicit. The main ones are log files³⁸, registered users' data and the results of user studies.

HTTP server log files are a source of implicit information about the traffic to the portal and about some general characteristics of the users such as their geographic location (based on their IP address), the site they last visited and the keywords they used to search in The European Library. Additionally, The European Library has built a custom logging system to record in depth the activities related to a search performed by a user - such as the language he uses for the interface, the collections he selects to view, the particular records he chooses and the additional services he might use (i.e. the "see online" function to view the digital object or the "print" option). These sources of information capture the intuitive search practices of users.

Additionally, there is personal information that the users agree to provide themselves upon registering³⁹ with The European Library such as profession, organisation and areas of interest. The research team of the work package is analysing registered users' data in combination with their recorded clickstream⁴⁰ history with the aim of providing true personalised search and services customised to the needs of this particular user.

Furthermore, analysis of the search habits of a targeted national, professional or age group can inform the design of functionalities to meet the needs of these particular user groups. Existing statistical data show, for example, that there is a connection between the country of provenance of users, the language used for the interface and the selection of particular collections. Thus, the data justifies the display of the user interface in French if the user has a French IP address as well as the showing of first results from collections that are likely to have entries in French such as collections from France, Belgium, Luxembourg or Switzerland. This doesn't apply only to users coming from a European country. Users coming from Lebanon for example also switch to the French interface as soon as they enter the portal (since Arabic is not offered yet as an option) and we can assume that they would also wish to first view records in French.

User studies are finally needed to confirm the informed decisions taken based on the analysis of the user logs over the design of particular services or functionalities of the portal. The European Library in the past has conducted two broad user studies that have further helped improve the usability of the portal.

In the framework of the TELplus project more focused user tests aim at investigating user requirements for the particular services to be developed in the project and it is intended that the expertise gathered becomes part of standard practice for developing new services in the future. The national libraries of the Netherlands, of France and of Italy (in Florence), which are partners in the work package, are contributing expertise in user studies and will set up focus groups to participate in the user tests. The first user tests will take place in spring 2009 and will investigate user feedback with respect to using image, text and video annotation tools.

The second year of the project will largely concentrate on continuing research and development in the various areas of work described, significantly increasing the digital content provided to the user and the usability of the portal. Major benefits include 20 million OCRed pages being made available for search by the partners through The European Library, and an increase in the number of centrally harvested metadata by at least 30 million records. The European Library will switch almost entirely from the inadequate distributed federated search to searching a centrally harvested pool of metadata from the partner libraries. This will enhance speed of search and retrieval of results and will make it possible to offer value-adding functionalities to users such as ranking and clustering of results. Semantic enhancement of records will be made possible on the centrally aggregated records with the use of the SKOSified subject headings from Work Package 3 and authority files from Work Package 4. Furthermore, The European Library will be able to showcase and assess the results of using the FRBR paradigm to search and browse records from the partner libraries.

The European Library will have made a significant contribution in the area of standardisation of service descriptions, the impact of which is too early to estimate. Nevertheless, The European Library portal will be able to profit from a more service-oriented modular

architecture. A number of important services will be developed, improving the overall usability and the search experience of the portal. These include annotation tools, as well as geolocation-relevant services and tools for searching full text (e.g. zoom in and highlight of terms-service). Regarding the latter, The European Library will have investigated via a prototype a new semantic search engine specifically adapted to the particularities of the OCR'd full-text from the partner libraries. Finally, a structured framework for analysing logged information from users' actions on the portal will have been developed which, in conjunction with results from user studies, will provide a significant insight into what users are looking for, where they are stumbling and leaving the site and how satisfied they are with the offered functionalities. These in turn will dictate user requirements for the implementation of the services under investigation in the project and will feed into significant improvements in the portal and into the provision of personalised services to the users.

In conclusion it should be mentioned that the important research and development being realised in the context of the TELplus project such as the REPOX system, along with the service infrastructure and the mass digitisation of content, will not only help innovate The European Library portal but will also feed into the development of a ground-breaking new service - Europeana⁴¹. Europeana is the cross-domain pan-European service offering access to the digital holdings of museums, archives, audiovisual archives and libraries. The prototype service is currently being developed and will be launched at the Council of Ministers of Culture in Brussels in November 2008. The European Library is not only one of the most important library content aggregators in Europe, but is also the cornerstone of this major initiative that has been raised by the European Commission⁴² and has been warmly embraced by all the member states for promoting and exploiting the rich European cultural heritage.

Endnotes

All websites mentioned in the endnotes were accessed on 2008-10-15

¹ CENL website: <http://www.cenl.org/>

² Council of Europe: <http://www.coe.int/>

³ Project Gabriel archive website:

http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive_en.html#gabriel

⁴ http://www.theeuropeanlibrary.org/portal/organisation/cooperation/archive_en.html#tel

⁵ TELplus project website: <http://www.theeuropeanlibrary.org/telplus/index.php>

⁶ National Library of Estonia: http://www.nlib.ee/?set_lang_id=2

⁷ EREMO s.r.l.: <http://www.eremo.net/en/home.php>

⁸ eContentplus Programme: http://ec.europa.eu/information_society/activities/econtentplus/index_en.htm

⁹ The TELplus workplan: <http://www.theeuropeanlibrary.org/telplus/workplan.php>. The other work packages that are not presented in the current paper involve the integration of the collections of the national libraries of Bulgaria and Romania (WP6), Dissemination (WP7) and Management and Coordination (WP8).

¹⁰ About Optical Character Recognition: http://en.wikipedia.org/wiki/Optical_character_recognition

¹¹ The Open Archives Initiative Protocol for Metadata Harvesting: <http://www.openarchives.org/pmh/>

¹² The Functional Requirements for Bibliographic Records model produced by IFLA:

<http://www.ifla.org/VII/s13/frbr/frbr.pdf>

¹³ Partners in the WP1 are: Austrian National Library (ONB), National Library of the Czech Republic (NLP), National Library of Estonia (RR), French National Library (BnF), National Széchényi Library of Hungary (OSZK), National and University Library of Iceland (LBS-HBS), National Library of Latvia (LNB), Martynas Mazvydas National Library of Lithuania (LNM), Norwegian National Library (NLN), National Library of Poland (BN), Slovak National Library (SNK), National and University Library of Slovenia (NUK), National Library of Spain (BNE), National Library of Sweden (NLS)

¹⁴ Bibliothèque Nationale de France: <http://www.bnf.fr/>

¹⁵ Part of the D1.2: "A survey of existing OCR practices and recommendations for more efficient work" will be publicly published on the TELplus website in November 2008 and this constitutes a significant contribution to relevant literature on the topic.

-
- ¹⁶ The information gathered through the project will also be exploited in the context of another European Project-IMPACT (<http://www.impact-project.eu/>) that is doing more in-depth research into innovating OCRing technology and tools.
- ¹⁷ SRU: <http://www.loc.gov/standards/sru/>
- ¹⁸ Z39.50: <http://www.loc.gov/z3950/agency/>
- ¹⁹ The European Library architecture is explained in The European Library Handbook: http://www.theeuropeanlibrary.org/handbook/handbook.php?handbook_menu=0-4-1&handbook_menu_pg=handbook/Accessing_my_collections/european_library_portal.html
- ²⁰ About the web 2.0: <http://www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html>
- ²¹ The European Library Mini Search Box: http://www.theeuropeanlibrary.org/portal/organisation/services/minisearchbox_en.html
- ²² Instituto Superior Technico: <http://www.ist.utl.pt/>
- ²³ The webpage of REPOX is: <http://rebox.ist.utl.pt/>
- ²⁴ The European Library Application Profile for objects: http://www.theeuropeanlibrary.org/handbook/Metadata/tel_ap.html
- ²⁵ The mock-up that was produced showcasing the multilingual and semantic functionalities under investigation for the user tests can be explored on this page: <http://maquette.bnf.fr/labs/telplusonline/>
- ²⁶ The European Library in the framework of EDLproject developed a prototype integrating the results of the MACS project: <https://macs.vub.ac.be/pub/> enabling access to bibliographic data from the national libraries of France, Germany, Switzerland and Great Britain using aligned English, French and German subject headings. The results are included in the following report: <http://www.edlproject.eu/downloads/D2%201final.pdf>.
- ²⁷ <http://www.cs.vu.nl/STITCH>
- ²⁸ <http://www.w3.org/2004/02/skos/>
- ²⁹ <http://www.w3.org/2001/sw/>
- ³⁰ The Rameau Subject Heading Language: <http://www.w3.org/2006/07/SWD/wiki/EucRameau>
- ³¹ SWD Subject Headings: <http://www.d-nb.de/eng/standardisierung/normdateien/swd.htm>
- ³² Library of Congress Subject Headings: <http://www.loc.gov/cds/lcsh.html#lcsh20>
- ³³ The TELplus report on the review of existing FRBR examples: [D3.6 Report on FRBR experiments \(29 July 2008\)](#) will be made publicly available on the TELplus website in November 2008.
- ³⁴ The agenda and the presentations of the workshop are available on the website of the national library of Portugal: http://www.bn.pt/workshop_telplus/index.html
- ³⁵ <http://maps.google.com/>
- ³⁶ <http://www.dlib.org/>
- ³⁷ Department of Information Engineering, University of Padova, Italy: <http://www.dei.unipd.it/wdyn/?IDsezione=1>
- ³⁸ About HTTP server logs: http://en.wikipedia.org/wiki/Server_log
- ³⁹ The European Library registration page: https://search.theeuropeanlibrary.org/portal/en/user_registration.html
- ⁴⁰ <http://en.wikipedia.org/wiki/Clickstream>
- ⁴¹ <http://www.europeana.eu/>
- ⁴² http://ec.europa.eu/information_society/activities/digital_libraries/europeana/index_en.htm