

ECP-2006-DILI-510003

TELplus

Provision of access to newly OCR-ed material through The European Library

Deliverable number	<i>1.6</i>
Dissemination level	<i>Public</i>
Delivery date	<i>19 January 2010</i>
Status	<i>Final</i>
Author(s)	<i>Georgia Angelaki, Nuno Freire, Michael Kranewitter, Sally Chambers</i>



eContentplus

This project is funded under the *eContentplus* programme¹,
a multiannual Community programme to make digital content in Europe more accessible, usable and exploitable.

¹ OJ L 79, 24.3.2005, p. 1.

1. Introduction

The present document is intended as a coversheet to Milestones 1.5 and Deliverable 1.6. The deliverable is classified as "other".

The result of WP1T5 as of December 31, 2009 is that more than 24 million pages of OCRed text are searchable from The European Library on the test version of the portal. The content comes from the national libraries of the following countries: Austria, Czech Republic, Estonia, France, Hungary, Iceland, Latvia, Lithuania, Norway, Poland, Slovakia, Slovenia, Spain and Sweden. In total, there were 1,6 million additional pages than the initial amount envisaged OCRed and delivered for indexing.

The pages have been delivered and indexed by IST and made searchable for TEL.

Search is possible at this moment via SRU query of the index that lies in IST. Currently, all full-text content is searchable at one go as one single collection of content. In the future, as the content expands for each library and we gain more experience with users querying the full-text, we might consider splitting the content into individual collections.

The European Library worked on the XSLT transformations that are necessary for the integration of the full-text results in the results page and on the user interface for the integration of the full-text search results. The European Library provided as well input to IST for the improvement of the SRU queries to the text.

2. Content harvested and indexed

Country	Initial amount of pages planned	Actual amount of pages delivered	Other files
Austria	500.000	534.000	
Czech Republic	3.400.000	2.579.511	
Estonia	200.000	713.933	597
France	7.000.000	8.242.908	
Hungary	200.000	237.914	
Iceland	2.800.000	5.727.149	
Latvia	100.000	195.075	
Lithuania	100.000	125.477	
Norway	1.600.000	1.600.000	

Poland	420.000	436.198	
Slovakia	200.000	185.000	
Slovenia	320.000	328.502	
Spain	5.500.000	3.033.525	
Sweden	200.000	253.653	
Total	22.540.000	24.192.845	597

On the 31st of December 2009, more than 24 million pages were made searchable from The European Library test portal. At the end of the project some content was still not harvested from Spain, Slovakia and Czech Republic. Spain had all the pages OCRed, however, the content was still being loaded into their digital library system at the end of December and it was not possible to make the remaining pages available for harvesting on time. With regards to Czech Republic, the Monograph collection was fully harvested but the Periodicals were harvested only partially due to problems encountered in linking metadata with the fulltext. Slovakia also was experiencing delays with the OCRing process and was unable to make available the remaining 15.000 pages before the end of December. These issues are closely monitored and the libraries are expected to make the content available, or fix problems in the coming weeks. This does not alter the fact that TELplus reached its targets both for the numbers of pages OCRed and for the number of pages harvested.

3. Search Scenarios

The current search scenarios are enabled:

- Simple Search from Homepage

The user searches on TEL portal homepage ticking the full-text search collection available under "digital collections". He searches, for example, for "Austerlitz".

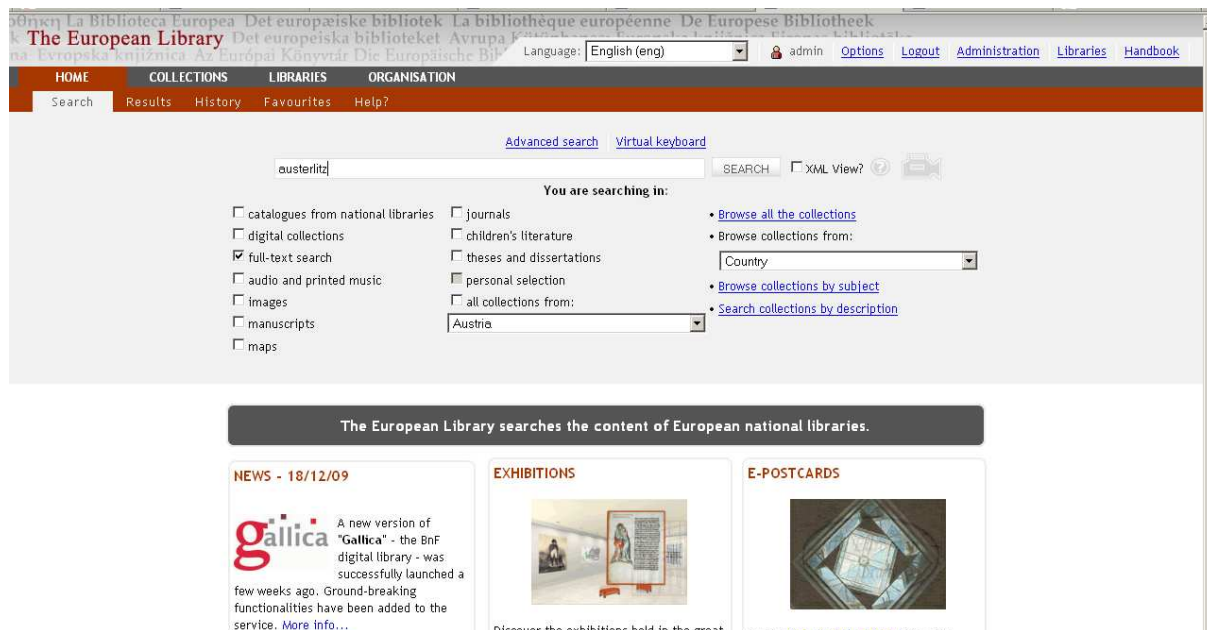


Figure 1: full-text search theme in The European Library Homepage

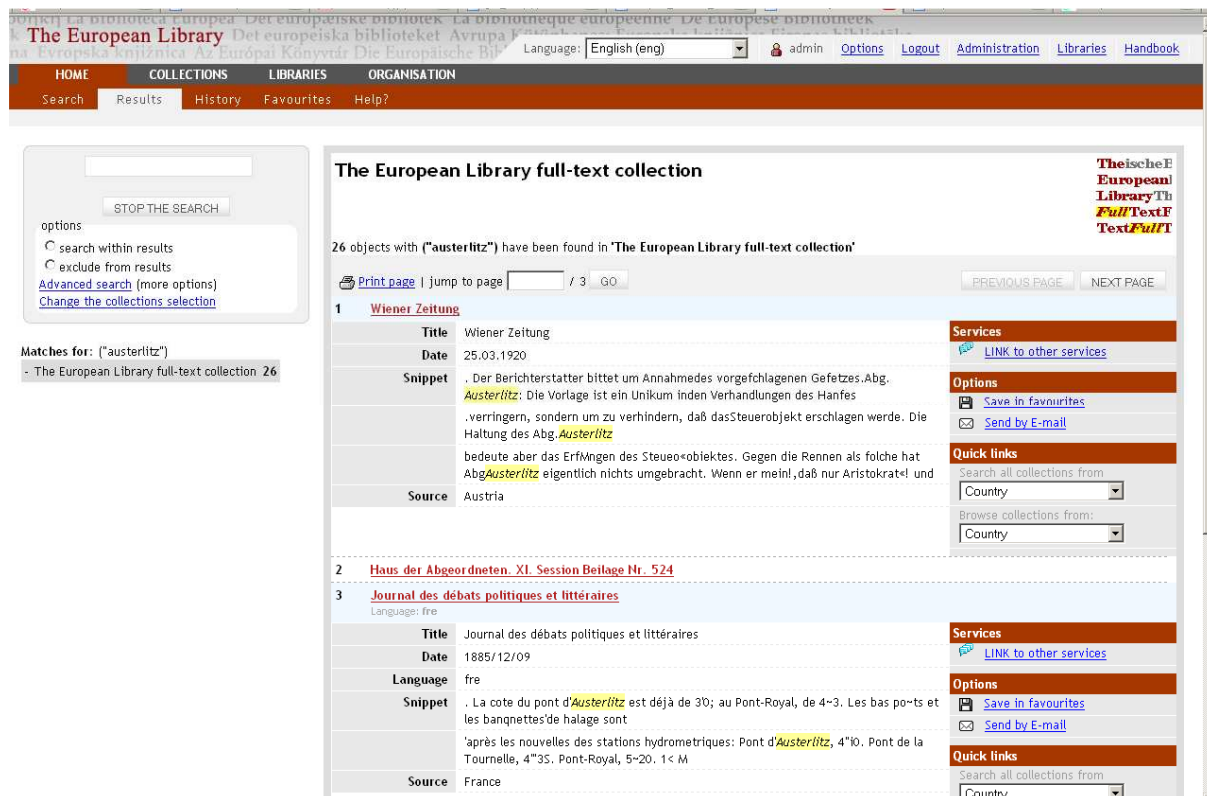


Figure 2: full-text search results (list of results and detailed record display)

- Search from “Collections” page

User performs a query in the full-text aggregated collection from the "Collections" page ticking the collection description appearing under The European Library Harvest.

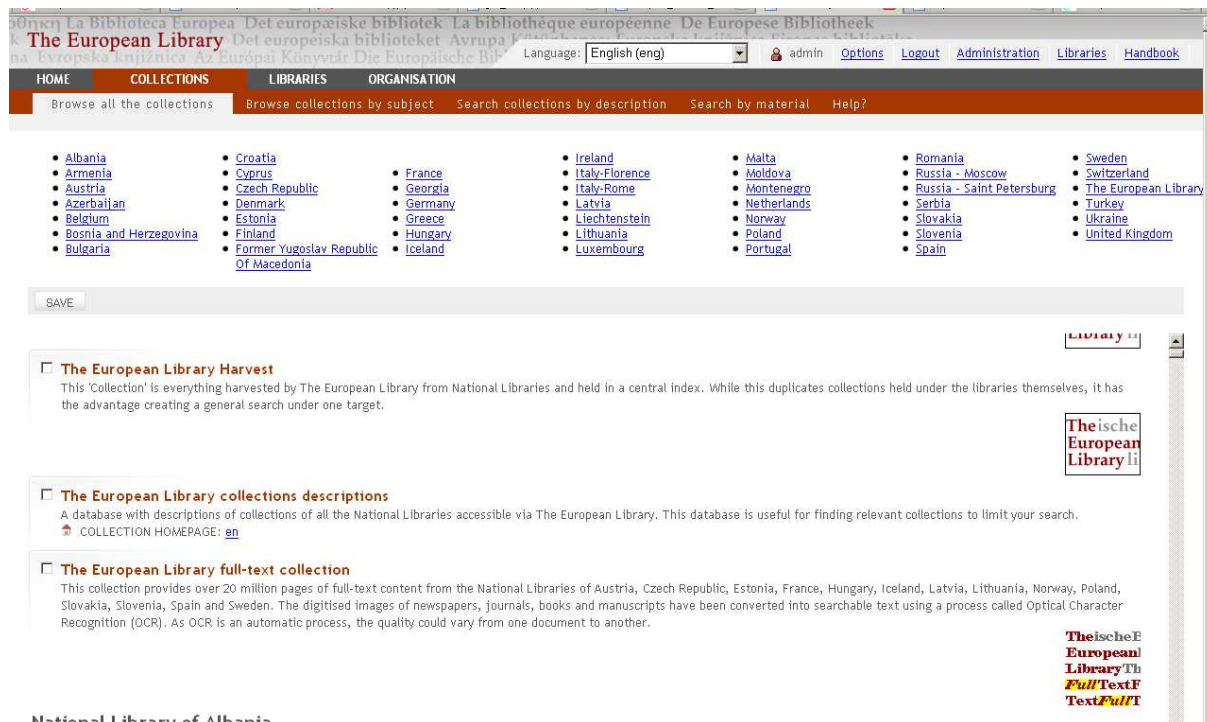


Figure 3: Selecting the full-text search collection from the collections' page

A collection description and new logo were created for the full-text collection.

The "See online" is used to redirect to the native environment for a detailed view of the document. The “see online” button will be deployed before the end of the year.

3. Additional information

As of the end of 2009, query of the full-text is demonstrated in the test version of the TEL portal and not in production. This is because the central index is still hosted in IST which is a research center and cannot guarantee full robust support of the service in production. While The European Library ordered new servers to accommodate fully the index in production, it took longer than expected to deliver and install the hardware and therefore, as of the end of 2009, the index was not migrated to ULCC.

As a follow-up, the following actions are planned at the start of 2010:

- At the beginning of January TEL will transfer the full-text harvester and indexer software as well as migrate the SOLR full-text index in the dedicated servers of our data hosting center in [ULCC](#) in London and provide query in the production portal.

Provision of access to newly OCR-ed material through
The European Library
TELplus D 1.6, Final, December 2009
Georgia Angelaki, Nuno Freire, Michael Kranewitter,
Sally Chambers



- We will fully test the new harvesting and indexing tools that IST will provide to TEL.
- It is expected that the full-text index will be searchable in The European Library production portal from Release 2.4 currently scheduled for Thursday 28th January.
- A re-harvest and re-indexing of all full-text is planned from all the partners to make sure that the process works on our end in a robust way. Priority will be given to the remaining content that it was not possible to harvest before the end of December 2009.
- We will advertise the new service offered by TEL around mid- February and after the whole process has been completed and search performs in a reliable way.