DELIVERABLE

Project Acronym:	Europeana Libraries
Grant Agreement number:	270933
Project Title:	Europeana Libraries: Aggregating digital content from Europe's libraries

D4.4 – Full-text content delivered to Europeana

Authors:

Andreas Juffinger (TEL)

Error! Reference source not found. (TEL)

Gilberto Pedrosa (IST)

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

Revision History

Rev.	Date	Author	Org.	Description
1.0	11/20/2012	Andreas Juffinger	TEL	Initial draft of feature list
1.1	12/12/2012	Andreas Juffinger	TEL	Incorporated the December delivery.
1.2	12/17/2012	Andreas Juffinger	TEL	Improvements.
1.3	08/01/2012	Andreas Juffinger	TEL	Review Comments

Statement of originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Index

1. Introduction

This delivery represents a short report about the full-text aggregation and delivery process carried out during the Europeana Libraries project. The aim of Task 4.4 was to make the existing full-text corpus available to Europeana and to extend and integrate the full-text harvesting mechanisms into the ingestion pipeline.

1.1 Status End of Year One:

Given that the final version of EDM (Europeana Data Model) -XML was not finalised in 2011 it was agreed that the best way to provide full-text material to Europeana was to use the existing ESE (Europeana Semantic Elements) model with an additional element from the *The European Library's* namespace, which holds the URLs to the full-text.

The status of the full-text delivery to Europeana as of End of December 2011, in terms of digital objects and full-text pages is presented in Table 1.

Country of origin	Digital objects provided	Pages provided	Total Pages
Estonia	40,637	713,933	713,933
Latvia	21,190	195,075	195,075
Norway	7,242	1,600,000	1,600,000
Poland	2,964	436,198	436,198
Sweden	58,277	253,653	253,653
Total:	130,310	3,198,859	3,198,859

Table 1 Full-text objects made available for Europeana

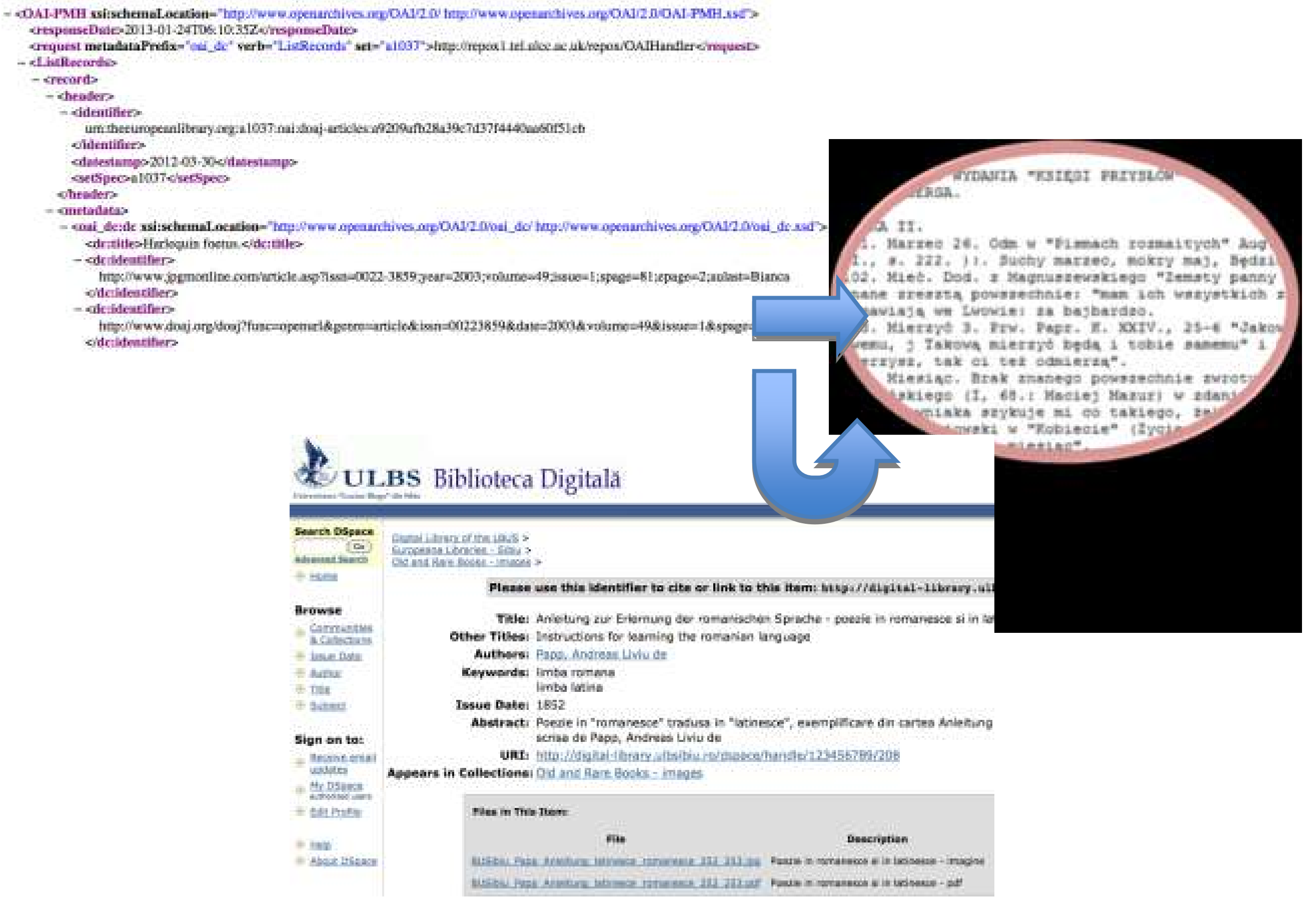
1.2 ESE/EDM Delivery

Europeana Libraries is listed as one of the case studies/best practices for the Europeana Data Model in the Europeana Professional Environment [6]. However, during the lifetime of the project Europeana did not yet support EDM in its production system; this led to the decision that all records from the Europeana Libraries should be provided in ESE 3.4.

This also influenced the full-text delivery, which follows the ESE 3.4 delivery format. As described in D5.2, EDM is now integrated into the ingestion workflow of *The European Library* and all data available in ESE is also available in EDM, subject to additional validation as soon as EDM is in production.

1.3 Efficient Full-Text Aggregation

During experiments, which have been carried out by the technical project partners, we have identified that the harvesting of full-text on page level (one link per page) is sub optimal and an unnecessary overhead. Furthermore, the analysis of the exchange of full-text data between partners and *The European Library* has shown that the harvesting process is computationally very expensive for provider and aggregator. As outlined in Chapter 2, we have implemented a full-text harvesting module that is triggered whenever an appropriate full-text URL is found in the OAI-PMH record. The document is downloaded and analyzed. In the case where the resulting document is an HTML page, it can be necessary to extract another link from that page which points to the actual full-text.



```
- <OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
    <responseDate>2013-01-24T06:10:35Z</responseDate>
    <request metadataPrefix="oai_dc" verb="ListRecords" set="a1037">http://repox1.tel.ulcc.ac.uk/repox/OAIHandler</request>
  - <ListRecords>
    - <record>
      - <header>
        - <identifier>
            urn:theeuropeanlibrary.org:a1037:oai:doaj-articles:a9209afb28a39c7d37f4440aa60f51cb
          </identifier>
          <datestamp>2012-03-30</datestamp>
          <setSpec>a1037</setSpec>
        </header>
      - <metadata>
        - <oai_dc:dc xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/oai_dc/ http://www.openarchives.org/OAI/2.0/oai_dc.xsd">
            <dc:title>Harlequin foetus.</dc:title>
          - <dc:identifier>
              http://www.jpgmonline.com/article.asp?issn=0022-3859;year=2003;volume=49;issue=1;spage=81;epage=2;aulast=Bianca
            </dc:identifier>
          - <dc:identifier>
              http://www.doaj.org/doaj?func=openurl&genre=article&issn=00223859&date=2003&volume=49&issue=1&spage=
            </dc:identifier>
```

Digital Library of the ULBS > European Libraries - Sibiu > Old and Rare Books - images >

Please use this identifier to cite or link to this item: http://digital-library.ul

Title: Anleitung zur Erlernung der romanischen Sprache - poezie in romanesce si in la

Other Titles: Instructions for learning the romanian language

Authors: Papp, Andreas Liviu de

Keywords: limba romana
limba latina

Issue Date: 1852

Abstract: Poezie in "romanesce" tradusa in "latinesce", exemplificare din cartea Anleitung scrisa de Papp, Andreas Liviu de

URI: http://digital-library.ulbsibiu.ro/dspace/handle/123456789/208

Appears in Collections: Old and Rare Books - images

Files in This Item:

File	Description
[illegible]	Poezie in romanesce si in latinesce - imagine
[illegible]	Poezie in romanesce si in latinesce - pdf

Figure 1 Flow of Full-Text Harvesting

An aggregator is bound to this interaction between provider and aggregator. The costs are still high, but necessary. In a trusted network between aggregators and Europeana one can significantly reduce the costs by working with batches rather than individual items.

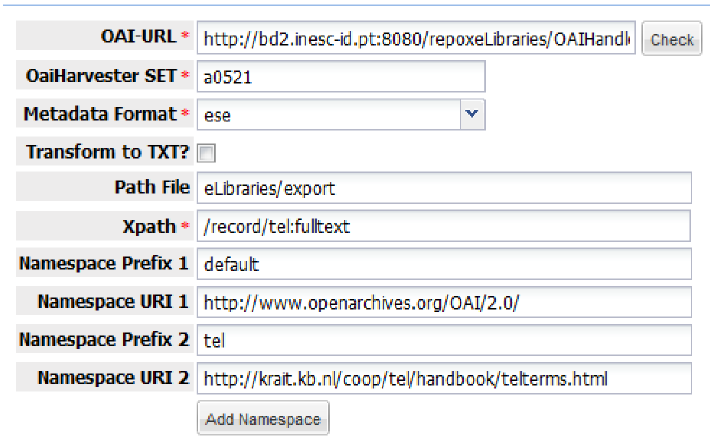
There are two dimensions which mainly contribute to the costs: Firstly, the number of interactions between data provider and data harvester. Each interaction naturally costs the overhead of establishing a network connection and some content negotiations. For example, the 49,801 digital items are split into 534,000 individual pages, and would cost an overhead (for downloading each item individually) of 500ms * 534,000 = 267,000 sec = 74h. For this reason, the aim must be to minimize the number of interactions. Secondly, storage and provision of individual items is expensive. For example, the 49,801 digital items from Austria have a storage footprint of about 4.8 GB; packed into one compressed archive the whole data can be exchanged in a single file with about 1.4 GB data.

The lessons learned from the provision after the first year are therefore:

- Create meaningful batches, which represent significantly more than one digital item. This forms one collection in the Europeana setting, given that all other aggregation interactions also happen on a collection level.
- Provide the metadata and full-text in one archive to simplify the handling. This is similar to the best practices in related projects (such as Europeana Newspapers), in which one exchanges the digital surrogates together with metadata in archives.

2. Full-text Harvesting

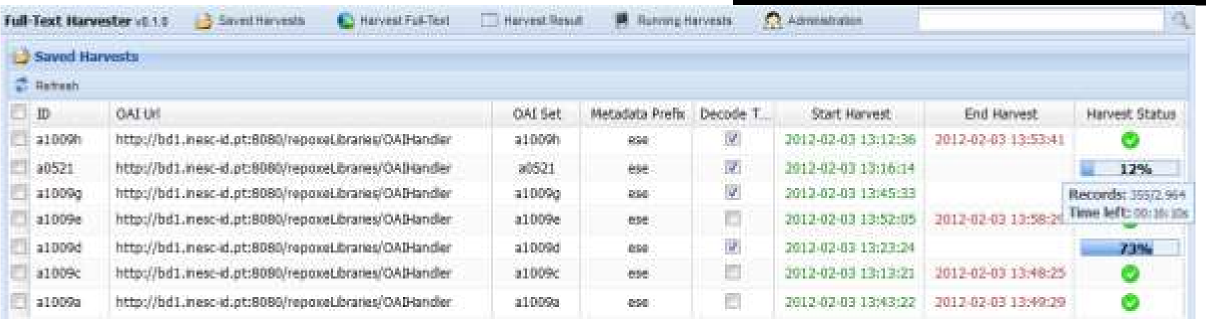
The Full-Text Harvester is the service responsible for analyzing the ingested records and harvesting the content linked from them (mostly PDFs in this project). It is composed of a web API, which exposes all its features through a REST based protocol, and a graphical user interface that allows a user to visually manage and monitor the harvests.



OAI-URL *	http://bd2.inesc-id.pt:8080/repoxeLibraries/OAIHandl	Check
OaiHarvester SET *	a0521	
Metadata Format *	ese	
Transform to TXT?	☐	
Path File	eLibraries/export	
Xpath *	/record/tel:fulltext	
Namespace Prefix 1	default	
Namespace URI 1	http://www.openarchives.org/OAI/2.0/	
Namespace Prefix 2	tel	
Namespace URI 2	http://krait.kb.nl/coop/tel/handbook/telterms.html	
	Add Namespace	

WYDANIA "KSIĘGI PRZYSŁOW
...ERGA.

...A II.
...1. Marzec 26. Odm. w "Pismach rozmaitych" Aug
...., s. 222.) 1. Suchy marzec, mokry maj, Będzi
...02. Miet. Dod. z Magnuszewskiego "Zematy panny
...znane zresztą powszechnie: "mam ich wszystkich z
...wymawiają we Lwowie: za bajbardzo.
...03. Mierzyć 3. Prw. Papr. E. XXIV., 25-6 "Jako
...wemu,) Takową mierzyć będą i tobie samemu" i
...rzysz, tak ci też odmierzą".
Miesiąc. Brak znanego powszechnie zwrot...
...skiego (I, 68.: Maciej Mazur) w zdani...
...niaka szykuje mi co takiego, że...
...wski w "Kobiecie" (Życie...
...miesiąc".



☐	ID	OAI Url	OAI Set	Metadata Prefix	Decode T...	Start Harvest	End Harvest	Harvest Status
☐	a1009h	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a1009h	ese	☑	2012-02-03 13:12:36	2012-02-03 13:53:41	✓
☐	a0521	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a0521	ese	☑	2012-02-03 13:16:14		12%
☐	a1009g	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a1009g	ese	☑	2012-02-03 13:45:33		Records: 355/2.964 Time left: 00:16:10s
☐	a1009e	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a1009e	ese	☐	2012-02-03 13:52:05	2012-02-03 13:58:2	
☐	a1009d	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a1009d	ese	☑	2012-02-03 13:23:24		73%
☐	a1009c	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a1009c	ese	☐	2012-02-03 13:13:21	2012-02-03 13:48:25	✓
☐	a1009a	http://bd1.inesc-id.pt:8080/repoxeLibraries/OAIHandler	a1009a	ese	☐	2012-02-03 13:43:22	2012-02-03 13:49:29	✓

The Full-Text Harvester stores the downloaded documents in a local file repository from where one can easily create an archive file for further processing and in our case for delivery to Europeana.

2.1 Robots Exclusion Protocol

The Robot Exclusion Standard, also known as the Robots Exclusion Protocol or robots.txt protocol, is a convention to prevent cooperating web crawlers and other web robots from accessing all or part of a website which is otherwise publicly viewable[1].

Many content and full-text providers, which publicly give access to the text, try to “exclude” the automatic download of full-text. For the time being, we only configure full-text harvesting if a library includes links to full-text and we have an explicit agreement for the usage of the content for indexing purposes. For OpenAccess material we follow the Budapest Open Access Initiative, which clearly identifies, in its original definition of Open Access from 2001, that OA is not only about making research outputs freely available for download and reading. The aspect of reuse, which includes being able to index and pass OA content to software, is firmly embedded in the definition.

[1] http://en.wikipedia.org/wiki/Robots_exclusion_standard

3. Full-Text Delivery

As outlined above we have changed the exchange pattern after year one from OAI-PMH + individual download to batch download per collection. In addition to the data provision in year one, we have provided the same data now also in batch mode. The full-text delivery of year two is exclusively available in batch mode to Europeana.

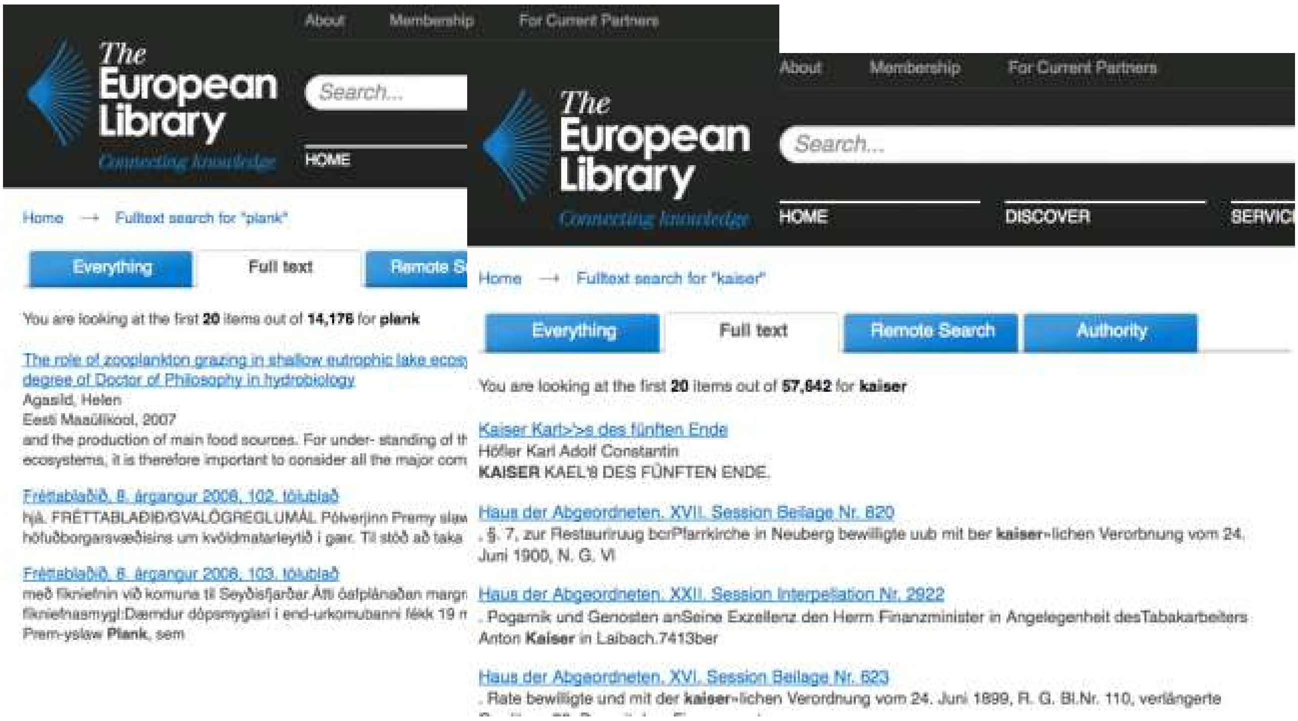
3.1 Full-text available from TELPlus Project

As of end of December 2012, we have been able to provide the whole corpus of the TELPlus project in 15 batches (one per country, except Spain which where two batches).

1st Year	130,310	3,198,859	3,198,859
Austria	49,801	534,000	534,000
Czech Republic	7,549	2,579,511	2,579,511
France	155,563	8,242,908	8,242,908
Hungary	4,039	237,914	237,914
Iceland	198,046	5,727,149	5,727,149
Lithuania	14,298	125,477	125,477
Slovakia	131,012	185,000	185,000
Slovenia	58,974	328,502	328,502
Spain	358,357	3,033,525	3,033,525
Sweden (cor. page number)	0	1,155,442	1,155,442
Total:	1,107,949	25,348,287	25,348,287

Table 2 Full-text objects made available for Europeana

In this way we have provided more than 25 million pages to Europeana and fully searchable in the new *The European Library* portal.



Home → Fulltext search for "plank"

You are looking at the first **20** items out of **14,176** for **plank**

The role of zooplankton grazing in shallow eutrophic lake ecos
degree of Doctor of Philosophy in hydrobiology
Agasild, Helen
Eesti Maaülikool, 2007
and the production of main food sources. For under- standing of th
ecosystems, it is therefore important to consider all the major com

Fréttablaðið, 8. árgangur 2008, 102. tölublað
hjá. FRÉTTABLAÐIÐ/GVALÖGREGLUMÁL Pólverjinn Premy slaw
höfuðborgarsvæðisins um kvöldmatarleytið í gær. Til stóð að taka

Fréttablaðið, 8. árgangur 2008, 103. tölublað
með fíkniefnin við komuna til Seyðisfjarðar Átti ósplánaðan margr
fíkniefnasmygl:Dæmdur dópsmyglari í end-urkomubanni fékk 19 m
Prem-yslaw **Plank**, sem

Home → Fulltext search for "kaiser"

You are looking at the first **20** items out of **57,642** for **kaiser**

Kaiser Karl>'>s des fünften Ende
Höfler Karl Adolf Constantin
KAISER KAEL'8 DES FÜNFTEN ENDE.

Haus der Abgeordneten. XVII. Session Beilage Nr. 820
. §. 7, zur Restauriruug bcrPfarrkirche in Neuberg bewilligte uub mit ber **kaiser**-lichen Verorbnung vom 24. Juni 1900, N. G. VI

Haus der Abgeordneten. XXII. Session Interpellation Nr. 2922
. Pogamik und Genosten anSeine Exzellenz den Herrn Finanzminister in Angelegenheit desTabakarbeiters Anton **Kaiser** in Laibach.7413ber

Haus der Abgeordneten. XVI. Session Beilage Nr. 623
. Rate bewilligte und mit der **kaiser**-lichen Verordnung vom 24. Juni 1899, R. G. Bl.Nr. 110, verlängerte

3.2 Additional Europeana Libraries Full-text

	Available Records	Available PDFs	Estimated Pages
OpenAccess: DOAJ	918,000	238,577	1,500,000
OpenAccess: DART	67,868	32,368	3,000,000
Tartu University Library	17,828	7,021	500,000
Bavarian State Library	640,293	0[2]	0
Hungarian Parliament	78,418	0[3]	0
Sibiu University Library	18	0[4]	0
Total Available:	1,722,425	277,966	5,000,000

Table 3 Europeana Libraries full-text objects available for Europeana

[2] License issues. Google and DFG funded digitization.

[3] About 1.8 mill pages in PDF/JPEG but no OCR access.

[4] About 1000 pages in PDF but no OCR access.

4. References

[1] D4.1 – Requirements Infrastructure and Harvester (Extended Revised Version)

[2] D4.3 – Report on how the full-text content will be made available to Europeana

[3] Europeana Professional – Technical Requirements - http://pro.europeana.eu/technical-requirements

[4] Europeana Professional – Europeana Data Model (EDM) Documentation - http://pro.europeana.eu/web/guest/edm-documentation

[5] Europeana Professional – Legal Requirements for Providing Data - http://pro.europeana.eu/web/guest/licensing

[6] Europeana Professional – Case Studies for EDM

http://pro.europeana.eu/europeana-libraries-edm