DELIVERABLE

Project Acronym:	Europeana Newspapers
Grant Agreement number:	297380
Project Title:	A Gateway to European Newspapers Online

D4.2 Newspaper Aggregation and Indexing Plan

Revision: 1.1

Authors: **Alastair Dunning (The European Library/Europeana Foundation)**

Markus Muhr (The European Library/Europeana Foundation)

Chiara Latronico (The European Library/Europeana Foundation)

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

Revision	Date	Author	Organisation	Description
0.1	27.02.2012	Markus Muhr	The European Library/Europeana Foundation	Initial version
0.2	27.02.2012	Alastair Dunning	The European Library/Europeana Foundation	Revisions
1.0				Published
1.1	25.06.2013	Alastair Dunning	The European Library/Europeana Foundation	Adjusted Names of Associate Partners

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This deliverable contains original unpublished work except where clearly indicated otherwise.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.

Table of Contents

1. Executive Summary

The Newspaper Aggregation and Indexing Plan for Europeana Newspapers provides management for the ingestion process. The content consisting of metadata, full text from OCR results and/or viewing images will be aggregated and ingested into The European Library, who are also responsible for designing the content browser for the project. The metadata will then be forwarded to Europeana and also the Zeitschriftendatenbank at the SBB.

This aggregation and indexing plan provides a realistic schedule for ingestion of content from the partners involved in OCR, those involved in named entity recognition and the Associate Libraries. This plan covers the process from getting the data into the central storage on The European Library hosting site, feeding the Newspaper content browser and finally the delivery of metadata to Europeana in the Europeana Data Model (EDM).

This plan cuts the ingestion process into small manageable pieces with quarterly deliverables. This allows for flexible changes in the schedule if problems occur as well as giving transparency to the aggregation process. This process covers everything that will be published and accessible in the content browser. Ingested data will include metadata for all 18 millions pages and full-text from OCR for around 10 millions pages.

The European Library will also host an image server for viewing pages if there is not such one in place at the partner and if the partner is willing to provide the images to The European Library. The partners choices for how they wish their content to be displayed are shown in appendix A of this document.

The schedule will be reviewed and updated if necessary every 6 months.

At the end of the project in January 2015, all the content from the libraries listed in the content table will be available through The European Library and delivered in EDM on a title level to Europeana. Furthermore, full-text content for 10 million pages will be indexed and accessible through the content browser.

This document is structured into 5 major sections. Firstly, it will start with a summary on the objectives and scope of the plan. Secondly, it will outline how the information was gathered by mainly explaining how the survey report guided the development of the aggregation and indexing plan as well as the collaboration with the partners. Thirdly, it will provide a risk analysis. Then, it will outline how the process will be implemented as well as monitored. Finally, a general overview for the individual quarters will be presented.

This Newspaper Aggregation and Indexing Plan is vital to the project and should be continuously revisited to allow performance evaluation and risk mitigation.

2. Introduction

This document provides an overview of the aggregation and indexing plan of the project. It provides a realistic schedule for ingesting 18 million pages from newspapers and also for additional aggregation of newspaper content from associated partners. Since 10 millions of these will be delivered by University of Innsbruck and CCS following their OCR scheduling, this aggregation plan follows their agreed deliverables in the OCR process. These 10m pages constitute a large chunk of data, so that the metadata including OCR full-text and potentially viewing images will be delivered via hard disks. For the remaining 8 million pages, most of the images will be accessed via an image server on the partner’s site, the metadata will be transferred via relevant Internet protocols. This plan will be split into quarterly scheduling blocks and should be reviewed on a regular basis at the end of every second quarter (6 months). The aggregation plan was established based on a survey to provide us with more information about what and how the data will be delivered. By December 2014, project month 36, 18 million pages should be accessible through the content browser either by their metadata and/or full-text.

3. Developing the aggregation and indexing plan

3.1 Europeana Newspaper Content And Associate Partners

The initial step for the aggregation plan was to gather all the information from the different partners including title names, number of issues, number of pages, etc. To get this initial information a survey was created and placed on sharepoint. The results were composed in an excel sheet on a title level. This document also served for the OCR partners to select and schedule the OCR process for different libraries. One major part includes the format of metadata as well as the kind of delivery to The European Library.

The Euroepana Newspaeprs survey identified associate partners and gave an overview of their digitised content. The 11 Associate Partners have now been chosen the metadata from their digitised collection has also been included in this document

3.2 OCR Process Plan

Since the aggregation of the newspaper content in The European Library and furthermore indexing the full-text depends on the OCR partners University of Innsbruck (UIBK) and CCS for over half of the newspapers, this aggregation plan follows closely the overall processing plan for OCR. The European Library will get the data from them via hard disks and will further send the data to ULCC in London, the hosting provider of The European Library, where the data will be manually uploaded onto our servers. The European Library will also make a copy of the hard disks and store those copies on site in Den Haag for backup. The format provided by UIBK as well as CCS will be an agreed METS/ALTO one. This allows the same treatment for all data provided by the two OCR partners.

3.3 Content Browser Option Paper

The European Library provided the partners with an options paper on how their content should be presented in the content browser as well as how the viewing images or thumbnails for the search result page should be served.

For all the data aggregated in the Europeana Newspaper project, metadata will be harvested and furthermore indexed and is therefore accessible within the content browser. Full-text is available for more than half of the 18 million pages and will be indexed as well. Some providers outlined that they have existing full-text available that might go into the content browser as well.

However, there were considerable differences on how to serve the viewing images. Basically, there are three options:

- The partner provides an image server that can be accessed by The European Library. No viewing images will be delivered to The European Library by the partner, instead The European Library will request preview images from those image servers and will embed the results into the results shown in the content browser.
- The partner has no image server in place, but is willing to provide the viewing images and approves the storage and usage inside an image server maintained by The European Library. In this case the delivered data includes besides metadata and/or full-text viewing images. Preview images shown in the content browser will be requested from The European Library's own image server.

- The final option is to neither provide the viewing images to The European Library nor giving access to an image server on site. In this case, the content browser will still show metadata and/or full-text, but only a link back to the specific page on the partner's website is provided instead of viewing images.

The paper with the different options are attached to this document.

3.4 Merging Plans to Aggregation and Indexing Plan

These three resources are the basis for this aggregation and indexing plan. The information was condensed and made centrally available through a Sharepoint system provided by the University of Innsbruck.

The overall aim of the aggregation plan was to balance the workload over the whole project. For this reason, each quarterly chunk includes deliveries from CCS and UIBK which consists of full-text and metadata in a METS/ALTO format. Depending on the option that has been chosen by the partner viewing images can be part of these deliveries. Furthermore, a quarterly plan includes newspaper content where only metadata will be delivered. Titles for which only metadata will be delivered includes partners in the proposal like ONB as well as associated partners.

Over the course of the project numbers of metadata and full-text will steadily grow. Since serving images on site by The European Library involves the setup of a large storage system, the aggregation plan and the OCR plan start with datasets where the images are served by the partner themselves or are not accessible at all. This gives The European Library sufficient time to prepare the storage system. Furthermore, the first prototype of the content browser will only be ready later in the year, so the indexing timeline is a shifted in the aggregation plan and will begin at the end of the second year of the project.

4. Risk Analysis

Risk in delays is omnipresent, so we need to have an agile approach to address changes in delivery times from partners. For this reason, we will adapt this document over time. Delivery cycles are quarterly to allow us some freedom on changing scheduling. This document should be revisited after each quarterly delivery. Furthermore, the updates should continue to go along with changes in the OCR scheduling, since they directly affect the aggregation status. We don’t expect changes on the content browser options for partners, but we may get in new partners delivering additional newspaper content, so that we may extend the aggregation plan to more content from new partners. There is some risk involved that The European Library will not have the infrastructure for serving images as well as the index to provide full-text retrieval ready, but this risk has been mitigated by postponing datasets that need the infrastructure ready to a later stage.

Risks that can become apparent:

- Aggregation of content from OCR partners is delayed due to
  - processing problems (e.g. OCR partner has unforeseen technical problems and OCR process stands still for days)
  - transfer problems (e.g. hard disks get lost on the way)
  - corrupted hard disks (e.g. data on the hard disks cannot be read by The European Library)
- Aggregation of metadata from partners is delayed due to
  - infrastructure problems at partner site
  - metadata that is not ready at provided dates
  - The European Library´s lack of support for certain metadata formats or transfer protocol
- Aggregation and indexing of data at The European Library is delayed due to
  - unready full-text indexing capabilities
  - infrastructure that is not ready and in place
  - the fact that expected infrastructure is not sufficient for requirements
- Delivery to Europeana is delayed due to
  - unavailable ingestion resources
  - insufficient EDM support
  - blocked ingestion process (e.g. Too many collections in pipeline)

Besides mitigation risk by having a dynamic ingestion plan, bigger problems will be communicated to the work package leader and ultimatively escalated to the project coordinator in severe cases.

5. Aggregation and Indexing Process

Aggregation will start in April 2013 (Q2 2013) with the first deliveries from University of Innsbruck and CCS. We expect then to get quarterly deliveries from the OCR partners. In addition harvesting of metadata from the National Library of Austria (ONB) will start.

Indexing of aggregated collections will start for metadata information immediately and will be accessible on the website of The European Library (www.theeuropeanlibrary.org). The full-text index as well as the content browser will be available as a prototype at the end of 2013 (?? Hmm September would be better). From that time on, full-text will be indexed at the same time as the aggregation and the indexing of metadata. Furthermore, the content will be delivered to Europeana in EDM provided by The European Library from April on as well. In addition, the image server hosted by The European Library will be set up in the first half of 2013 as well and will be available together with the first version of the content browser in the second half of 2013.

The first delivery in April will contain 1 million records from Innsbruck with 20 titles. Q2 in 2013 will ingest 2 millions from Innsbruck, 50 000 from CCS and 2 millions in addition as plain metadata. This will roughly continue for each of the following quarters. An overview is provided in the following table.

Note, that roughly 3 million pages are missing in this first version of the aggregation plan as University of Innsbruck has not received plans on delivery of images from the partner libraries and they cannot plan their OCR process. Since we are depending on University of Innsbruck, we will add the aggregation of these 3 million pages later in a later version. Furthermore, the dates for the delivery of the results for named entity recognition has not been specified yet either. Furthermore, first named entity recognition results will be delivered to The European Library at the end of 2013. The first results will be for content from the Austrian National Library and Teßmann Library.

Quarter	Action	# pages	# full-text	# images
Q1 2013	Planning/Sample	-	-	-
Q2 2013	3 Providers	7.643.817	1.952.793	857.485
Q3 2013	4 Providers	10.963.784	390.366	99.701
Q4 2013	10 Providers	8.795.800	1.115.000	1.523.800
Q1 2014	9 Providers	6.697.992	2.355.594	233.648
Q2 2014	7 Providers	455.000	455.000	255.000
Q3 2014	5 Providers	551.428	551.428	255.000
Q4 2014	4 Providers	552.723	552.723	309.962

There are four key activities that run sequentially. First, aggregation of data by The European Library that consists partly of hard disks (UIBK, CSS) and partly of metadata harvested via the Internet from partners. The data on the hard disks need then to be sent further to London to be put onto the servers of The European Library (hosted by University of London Computer Centre). The second phase includes ingesting the metadata into the services at The European Library and will then be accessible from the portal (www.theeuropeanlibrary.org) . Third phase: indexing the full-text in a separate search index optimized for this purpose and uploading if necessary the viewing images to the image server. After this step, the content browser will be filled with the data and the newspapers can be retrieved from there. Fourth, on a title level the objects are delivered to Europeana in EDM and will then be retrievable from www.europeana.eu at the latest one month after delivery. This last step is repeated to provide the objects on a title level to the Zeitschriftendatenbank at the SBB.

The aggregation plan follows the common ingestion plan of The European Library, so that the ingestion process will benefit from the experience of the team there. The last quarter in the aggregation plan is smaller in size to give The European Library some buffer to finalize the aggregation process and to make sure all relevant content is accessible through the content browser and from Europeana.

6 Schedule of Aggregation and Indexing Plan

The aggregation of newspaper full-text, metadata and/or images will start from April on, so the first planned quarter is the 2nd one of 2013 (April till June). The other quarterly aggregation have been planned for July 2013, October 2013, January 2014, April 2014, July 2014 and October 2014. The last quarter of the project can be used as buffer.

At the end of the project 18 000 000 pages are accessible through the content browser that originate from 2000 newspaper titles. 2000 records representing these titles will be delivered to Europeana and will be accessible through the Europeana portal (www.europeana.eu). Indirectly all the content is also available via Europeana. All the pages are accessible on an individual level either through full-text or full-text and metadata search in the content browser developed by The European Library.

However, the first version of the content browser will not be available in the first 2 quarters for this aggregation plan, so that indexing of this content will start after the first quarter. Retrieval of metadata will be made available right after starting aggregation in the portal of The European Library (www.theeuropeanlibrary.org). Titles that have been ingested will also be delivered and made available when the aggregation starts. Full-text and viewing images will be made available with the first version of the content browser in the fourth quarter of 2013.

The tables in the following chapters show the aggregation plan per quarter. The table columns are the acronym for the provider, the country of the provider, the distribution way (via CCS, UIBK or directly from partner for non-refined material), the number of pages in the definition of done, the accumulated number of pages, and the number of pages, full-text and images to be aggregated and indexed in the specific quarter.

Acronyms used in the following tables for OCR partners:

CCS – Content Conversion Specialists

UIBK – University of Innsbruck

Acronyms used in the following tables for partner libraries:

BnF - Bibliotheque Nationale de France / National Library France

KB - Koninklijke Bibliotheek / National Library of the Netherlands

LFT Landesbibliothek Dr. Friedrich Teßmann / Teßmann Library

NLE - Eesti Rahvusraamatukogu / Estonian National Library

NLF – Kansalliskirjasto / National Library of Finland

NLL - Latvijas Nacionala Biblioteka / National Library of Latvia

NLP - Biblioteka Narodowa / National Library of Poland

NLT - Milli Kutuphane Baskanligi / National Library of Turkey

ONB - Österreichische Nationalbibliothek / Austrian National Library

SBB - Staatsbibliothek zu Berlin / Berlin State Library

SUBHH - Staats- und Universitätsbibliothek Hamburg / State and University Library

UB - Univerzitet u Beogradu / University Library of Belgrade

ULBT - Universitäts- und Landesbibliothek Tirol / University Innsbruck Library

Acronyms used in the following tables for associated partners:

NLW - National Library of Wales

NLB - St. Cyril and Methodius National Library / The National Library of Bulgaria

NLC - National Library of Czech Republic

NLH - National and University Library in Zagreb

KBR - Koninklijke Bibliotheek van België / Bibliothèque royale de Belgique

NUK - Narodna in univerzitetna knjižnica / National and University Library of Slovenia

NLPR - National Library of Portugal

NRO - National Library of Romania

LI - Landsbókasafn Íslands - Háskólabókasafn / National and Univeristy Library of Iceland

BNE – National Library of Spain

BNL - Bibliothèque nationale de Luxembourg / National Library of Luxembourg

6.1 April – June 2013 (2nd quarter 2013)

Provider	Country	Via	DoW	# accum	# pages	# full-text	# images
BnF	FRA	CCS	3.000.000	5.000	5.000	5.000	0
LFT	GER	UIBK	840.082	857.485	857.485	857.485	857.485
ONB	AUT	UIBK	7.057.772	1.090.308	1.090.308	1.090.308	0
ONB*	AUT	ONB	7.057.772	6.781.332	5.691.024	0	0

*Over 2nd and 3rd quarter

6.2 July – September 2013 (3rd quarter 2013)

Provider	Country	VIA	DoW	# accum	# pages	# full-text	# images
BnF	FRA	CCS	3.000.000	55.000	50.000	50.000	0
NLE	EST	CCS	594.657	5.000	5.000	5.000	5.000
BnF	FRA	UIBK	3.000.000	255.000	200.000	200.000	0
NLE	EST	UIBK	594.657	99.701	94.701	94.701	94.701
NLF	FIN	UIBK	114.000	40.665	40.665	40.665	0
NLT	TUR	NLT	418.345	8.990	8.990	0	0
LI	ICE	LI	0	4.112.602	4.112.602	0	0
BNE	ESP	BNE	0	5.831.826	5.831.826	0	0
BNL	LUX	BNL	0	620.000	620.000	0	0

*Over 2nd and 3rd quarter

6.3 October – December 2013 (4th quarter 2013)

Provider	Country	Via	DoW	# accum	# pages	# full-text	#images
BnF	FRA	CCS	3.000.000	355.000	100.000	100.000	0
NLE	ÉST	CCS	594.657	114.701	15.000	15.000	15.000
SUBHH	GER	UIBK	1.545.000	400.000	400.000	400.000	400.000
SBB	GER	UIBK	1.710.000	400.000	400.000	400.000	400.000
UB	SRB	UIBK	387,700	200.000	200.000	200.000	200.000
SUBHH	GER	PR	1.545.000	908.800	508.800	0	508.800
NLW	GBR	NLW	0	1.100.000	1.100.000	0	0
NLB	BUL	NLB	0	12.000	12.000	0	0
NLC	CZE	NLC	0	5.760.000	5.760.000	0	0
NLH	CRO	NLH	0	300.000	300.000	0	0

6.4 January – March 2014 (1st quarter 2014)

Provider	Country	Via	DoW	# accum	# pages	# full-text	#images
BnF	FRA	CCS	3.000.000	555.000	200.000	200.000	0
NLE	EST	CCS	594.657	164.701	50.000	50.000	50.000
SUBHH	GER	CCS	1.545.000	1.008.800	100.000	100.000	100.000
NLP	POL	UIBK	99.106	83.648	83.648	83.648	83.648
KB	NET	KB	2.000.000	1.921.946	1.921.946	1.921.946	0
KBR	BEL	KBR	0	3.500.000	3.500.000	0	0
NUK	SLO	NUK	0			0	0
NLPR	PRT	NLPR	0	400.000	400.000	0	0
NRO	ROM	NRO	0	442.398	442.398	0	0

6.5 April – June 2014 (2nd quarter 2014)

Provider	Country	Via	DoW	# accum	# pages	# full-text	#images
BnF	FRA	CCS	3.000.000	755.000	200.000	200.000	0
NLE	EST	CCS	594.657	314.701	150.000	150.000	150.000
SUBHH	GER	CCS	1.545.000	1.108.800	100.000	100.000	100.000
SBB	GER	CCS	1.710.000	405.000	5.000	5.000	5.000

6.6 July – September 2014 (3rd quarter 2014)

Provider	Country	Via	DoW	# accum	# pages	# full-text	#images
BnF	FRA	CCS	3.000.000	1.005.000	250.000	250.000	0
NLE	EST	CCS	594.657	464.701	150.000	150.000	150.000
SUBHH	GER	CCS	1.545.000	1.208.800	100.000	100.000	100.000
SBB	GER	CCS	1.710.000	410.000	5.000	5.000	5.000
NLF	FIN	CCS	114.000	46.428	46.428	46.428	0

6.7 October – December 2014 (4th quarter 2014)

Provider	Country	Via	DoW	# accum	# pages	# full-text	#images
BnF	FRA	CCS	3.000.000	1.202.761	197.761	197.761	0
NLE	EST	CCS	594.657	594.657	129.962	129.962	129.962
SUBHH	GER	CCS	1.545.000	1.388.800	180.000	180.000	180.000
NLF	FIN	CCS	114.000	91.428	45.000	45.000	0

7. Summary and Future Work

The aggregation and indexing plan is the central point for ingestion into The European Library and Europeana on time. The scheduling involves all partners: The European Library, Europeana, University of Innsbruck, CCS, KB and all data providers.

The aggregation and indexing plan was created by The European Library in cooperation with all partners in the project foremost the OCR delivery organisation University of Innsbruck and CCS as well as KB who will deliver enriched data namely named entity recognition and the partners. This plan covers the whole aggregation and indexing process done by the European Library.

After creating this aggregation and indexing plan in an initial version, The European Library will start to carry out the aggregation depending on the first delivery from University of Innsbruck. Any potential problems need to be identified quickly and addressed immediately. This aggregation plan should be revisited after each quarterly aggregation.

At the end of project in 2014 all the content listed in this plan should be accessible through web pages at Europeana or The European Library.

A. Content browser option by library

Library	Country	Metadata	Full-text	Images	Serving images	Content Browser Choice
UB	SRB	408.181	408.181	408.181	Images served by TEL	Option 1
SBB	GER	248.200	248.200	248.200	Images served by TEL	Option 1
ONB	AUT	6.781.332	1.090.308	0	Images served by ONB	Option 1
NLT	TRK	8.990	0	0	No images in content browser, referral to webpage at NLT	Option 4
NLP	POL	83.648	83.648	83.648	Snippet images served by TEL, referral for full images	Option 2
NLL	LTV	460.781	460.781	460.781	Images served by TEL	Option 1
NLF	FIN	132.093	132.093	0	Images served by NLF	Option 1
NLE	EST	594.663	594.663	594.663	Images served by TEL	Option 1
LFT	GER	857.485	857.485	857.485	Snippet images served by TEL, referral for full images	Option 2
KB	NET	1.921.946	1.921.946	0	Images served by KB	Option 1
SUBHH	GER	2.216.200	1.707.400	2.216.200	Images served by TEL	Option 1
BnF	FRA	2.388.488	2.388.488	0	Images served by TEL	Option 1