DELIVERABLE

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

D2.3 Upgraded content

Revision: 1.0

Authors: **Clemens Neudecker, KB**

Lotte Wilms, KB

Günter Hackl, UIBK

Siobhan Bolli, CCS

Contributions: **WP2 participants**

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	17-06-2014	Clemens Neudecker, Lotte Wilms	KB	Created
0.2	30-06-2014	Clemens Neudecker, Lotte Wilms	KB	Updated
0.3	23-07-2014	Clemens Neudecker	KB	Updated
0.4	24-07-2014	Hans-Jörg Lieder	SBB	Internal review
1.0	30-07-2014	Clemens Neudecker	SBB	Final version

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This deliverable contains original unpublished work except where clearly indicated otherwise.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 objective of Work Package 2 Refinement is the upgrading of 10 million scans of digital newspaper pages from participating libraries with Optical Character Recognition (OCR), Optical Layout Recognition (OLR) and Named Entities Recognition (NER). The main output of this deliverable is the actual refined content, i.e. 10 million files containing OCR results in ALTO format, combined with a METS container according to the ENMAP (Europeana Newspapers METS/ALTO Profile) specified in Work Package 5 Metadata Best Practice Recommendations. 8 million pages are processed with OCR only by the University of Innsbruck (UIBK), and 2 million pages are processed with OLR by Content Conversion Specialists (CCS), Hamburg. In addition, suitable content in Dutch, German and French language is further refined with NER by the KB National Library of the Netherlands.

This document provides a detailed overview of the distribution of upgraded content between the 12 participating content holding institutions, and also gives an account of the three main workflows for refinement, namely OCR, OLR and NER.

The scheduling, provisioning and upgrading of 10 million scanned newspaper pages from 12 partners and with three different technology providers required a highly streamlined workflow. This was put into place in the first year of the project according to the *Specification of requirements of OCR and structural refinement-services for digitised newspapers in Europeana* (D2.2).

Due to the large amount of content and the complexity of the material, naturally some deviations from the initial planning were to be expected. In some cases, libraries had difficulties providing large exports from their data repositories, since, in order to ensure the best possible quality of refinement results, master files with high optical resolution and file size had to be used. In addition, the scale of content made it necessary to enforce a strict regulation of the data packages that were provided by content holders to the technical partners, so that the processing could happen in a fully automated way, which also proved to be a challenge for some libraries.

These issues were mitigated by the provision of clear instructions and software tools that support the libraries in the process of data preparation, and as a side-effect, drastically reduced the file size of the images that have to be provided to the technical partners, thus speeding up the actual refinement.

In two cases, due to project-external reasons, the digitisation (i.e. scanning) of the newspapers was not fully completed before the refinement of this content was scheduled in the project. To guarantee that the total amount of 10 million pages will nevertheless be reached in time, and the facilities available for processing of content at the technical partners were not idling, several minor changes had to be made to the division of content across partners and the scheduling of deliveries.

Thanks to these measures, by the end of July 2014 (M30), a total of **9.017.641 / 7.961.247** pages (for a detailed explanation of the figures, see the table under Chap. 2 Content) have been successfully upgraded. However, due to the availability of the technical resources, it was decided to continue the refinement process until at least October 2014 (M33), so that even more content can be added. This has been agreed with the Work Package leader responsible for aggregation, so that it can also be ensured that all of this content will still be fed into the presentation platform provided by The European Library, and Europeana.

2. Content

Partner	Pages planned	Pages delivered M30	Pages done - OCR M30	Pages done - OLR M30 [7]	Total refined content M30
BnF	2.380.000	2.174.452	1.287.606	786.888	2.074.494
SBB[1]	1.700.000	916.187	873.700	7.432	881.132
ONB	1.600.000	1.364.106	1.364.106	-	1.364.106
LFT[2]	840.000	1.021.362	1.021.362	-	1.021.362
NLE[3]	560.000	591.702	94.698	448.599	543.297
NLF[3]	100.000	131.713	40.659	82.464	123.123
NLL	450.000	454.639	454.639	-	454.639
NLP	100.000	108.432	108.432	-	108.432
NLT[4]	400.000	406.460	222.473	-	222.473
SUBHH	1.430.000	1.153.632	605.197	170.669	775.866
UB	387.000	392.323	392.323	-	392.323
∑	**9.947.000**	**8.715.008**	**6.465.195**	**1.496.052**	**7.961.247**
SBB+[5]	-	286.662	286.662	-	286.662
ONB+[6]	-	769.732	769.732	-	769.732
∑+	9.947.000	9.771.402	7.521.589	1.496.052	**9.017.641**

(Figures of refinement progress as of 31-07-2014)

[1] SBB will only deliver around 1.100.000 pages in due time, but deliveries of content for refinement will continue for a certain time as long as ingests to The European Library and Europeana can still be guaranteed.

[2] To compensate pages currently missing from SBB, LFT agreed to deliver 150.000 pages more for OCR than originally planned.

[3] Due to a mistake in the original Description of Work, NLF had no pages foreseen in the OLR workflow. To mitigate this, NLE kindly agreed to reduce their content in the OLR workflow by 60.000 pages which have in turn been granted to NLF.

[4] Unfortunately there is no OCR technology available at the moment that is suitable for Ottoman script, which is the main script in historical newspapers from the National Library of Turkey. Accordingly, the decision was taken to OCR 202.662 pages in Latin alphabet and only perform

layout recognition for the remaining 203.798 pages in Ottoman, thereby facilitating the subsequent transcription of those Ottoman pages.

[5] An additional 286.662 journal pages from SBB were processed with OCR as part of collaboration with the Europeana Collections 1914-1918 project[1].

[6] ONB delivered an additional 769.732 pages that have been OCR'd and indexed for the project, but for which only metadata will be shown in the TEL browser. However, thanks to the indexing, it will be possible for users to find these newspapers with a full-text search via both, TEL and Europeana.

[7] Due to the specific workflow for OLR - which includes also a quality control step including manual corrections by libraries - there is currently less data listed as "completed" in the CCS workflow as has already been refined. The total number of pages already in the pipeline is 2.027.890, with 531.838 pages awaiting verification by libraries before finally being forwarded to TEL and Europeana in the course of August - October.

[1] http://www.europeana-collections-1914-1918.eu/

3. Process

Organising the refinement of 10 million pages requires a very clear and highly standardised workflow to ensure that all parties are aware of their tasks at all times and can always track their data at any point in the defined workflow. The setup of this process is described in more detail in deliverable D2.2 *Specification of requirements of OCR and structural refinement-services for digitized newspapers in Europeana*, but a short recap will be given here.

All libraries input the information about the data they selected for refinement into a central Master List on the project extranet. This list contains all necessary information for the refinement, such as language, image size, font type and total number of pages. Next to this, the list also contains the metadata about the object itself that will be shown in the image browser, such as title, date range and the identifier of the library.

Once the libraries completed the Master List, they prepared their files according to the specifications and tested their compliance with a specifically built tool, the File Analyzer Tool[2]. This tool checks the validity of the images, metadata and the standardised directory structure. If the files were checked by the FAT tool and no errors occurred, the libraries can be safe to continue with the creation of viewing copies needed for the presentation and the conversion of the images to bitonal (i.e. black-and-white), using the File Binarization and Conversion Tool[3]. This was done to reduce the file size of the images, thereby making it possible to ship the large amounts of data via regular cost-efficient hard disks. However, this is only an intermediate step in the technical refinement process. Should a library have scans in greyscale or colour available, these will still be used instead for rendering the presentation for the end-user. The final pre-processing step at the libraries was the copying of the data to the hard disks and shipping it to the technical partners for the actual refinement.

3.1 OCR@UIBK

Within the Europeana Newspapers Project, UIBK is the main provider for OCR. Around eight million newspaper pages will be enriched with OCR by the University of Innsbruck as part of the project.

The work is carried out within the “Abteilung für Digitalisierung und elektronische Archivierung” (Department for Digitisation and Digital Preservation”) who have gained extensive experience with OCR by participating in various European project such as MetaE, IMPACT and others and are also currently providing the technical and administrative backbone of the European eBooks on demand service EOD, which also includes OCR amongst its services.

UIBK uses the state-of-the-art commercial application for OCR, ABBYY’s FineReader, which was also developed further as part of the IMPACT project. In the course of 2012, UIBK modified their OCR service platform to use the FineReader Engine SDK instead of the Recognition Server

[2] More information about the FAT (File Analyzer Tool) tool can be found in deliverable D2.2.

[3] More information about the BCT (Binarization and Conversion Tool) tool can be found in deliverable D2.2.

because it gives them more flexibility in the configuration while maintaining aptitude for large-scale processing. The version of the FineReader SDK that is being used is 11.0, the most recent and robust release and available at the time of refinement.

3.2 OLR@CCS

In addition to the roughly eight million pages of “regular” OCR provided to the project by the University of Innsbruck, another two million pages are OCR’d with additional structural refinement, referred to as OLR (Optical Layout Recognition). In addition to merely recognising the text, OLR includes advanced features such as the separation of articles and the classification of pages (e.g. advertisements, titles pages, etc.).

The OLR workflow is run by CCS, a company specialising in newspaper digitisation and with ample experience collected in large-scale newspaper digitisation projects. CCS uses their in-house docWorks software technology for the project. This is an intelligent application for automatic conversion, structuring, and indexing of printed or electronic documents such as books, journals, newspapers and magazines. With docWorks, it is possible to locate and categorise key data from a variety of documents.

After raw data verification and ingest, the conversion process starts off with page analysis to determine page frames, followed by several zoning and structure recognition steps, where each element is assigned a specific zone category, and the individual elements (e.g. headlines, text blocks, illustrations) are grouped together into articles.

Each automatic detection step can also be followed by manual verification, depending on the quality level required. For mass digitisation projects, manual verification is typically reduced to a minimum, but the content holders participating in the OLR workflow were provided with three alternative solutions for manual quality assurance by CCS, thus allowing them to at least sample the quality, perform some manual corrections and understand the importance of good raw material for optimal automatic results.

3.3 NER@KB

For a subset of the OCR’d content from partners in Dutch, German and French language, the National Library of the Netherlands provides tools and technologies for the extraction of named entities such as person and place names or the names of organisations. This greatly improves the usability of the full-text for further text-mining and scientific purposes, since users can search for specific individuals or places. This also allows the cross-linking of the refined newspaper content with other online information resources such as authority files and the linked open data cloud.

The NER system that is implemented by KB for the project builds on prior work and experience derived from the IMPACT project. The Stanford University NER tagger, a mature and widely used machine learning tool for NER, is used and was also further extended for the project. The software itself, including all necessary adjustments as well as the data that is used for training, will subsequently be published under an open license[4] and has already attracted users external to the project, e.g. from the Stanford University NLP group or the Translantis[5] project in the Netherlands.

Also in the NER workflow a slight change of plans occurred. In the course of the first year, following the analysis of the content from libraries, it became obvious that only little content is going to be provided in English language (since the British Library is a networking partner only and does not provide digitised newspaper pages to the projects data set). The development of NER tools for English would therefore not have reached the desired impact with regard to the content available in the project. Accordingly, given that a significant share (approx. 25%) of the full text produced by the project is in French language, an investigating into the feasibility of extending the NER workflow for the project with support for French was conducted. Following this, in December 2012, the National Library of France (BnF) submitted a proposal to the Project Management Board for the development of NER resources for French via collaboration with the ACASA LIP6 Group of the Université Pierre et Marie Curie in Paris. An agreement was reached that ACASA LIP6 will develop technical resources for French NER in close collaboration with the BnF and the Europeana Newspapers project. These technical resources will follow the design principles of the NER approach chosen by KB, and will be made available under equal terms as for the other languages already supported.

In 2013 the KB submitted a proposal to the Project Management Board about further extending the NER workflow with disambiguation and linking services for named entities. Disambiguation of entities and referencing of authority files are more advanced ways to refine NER results, which in turn require a more sophisticated implementation, but at the same time allow for a much more useful presentation of the results. The proposal was approved by the Project Management Board and KB will deliver additional modules for these purposes in the second half of 2014.

The upgrading of full-text with NER is an incremental process. Results from a first iteration feed into a tuning step that will enhance the results of the second iteration and so forth. The first iteration for Dutch content from KB has already been completed, as also the first iteration for German content from Dr Friedrich Tessmann Library and the Austrian National Library. The European Library is working on a prototype implementation for the incorporation of these results into the main presentation interface, a first version of which is scheduled for release in July 2014.

[4] https://github.com/KBNLresearch/europeananp-ner

[5] http://translantis.wp.hum.uu.nl/

4. Issues

Although the refinement workflow of the project was set up with the utmost care, some issues did occur when the actual refinement was started. One of the most important issues encountered was the fact that several libraries were over-confident regarding the availability of their material. Some libraries still needed to digitise parts of their collection that were supposed to be ready for enrichment in the project, while another library was just in the middle of setting up a new digital discovery system that hindered them in delivering their data according to the original planning. In addition, there were issues due to force majeure, such as the evacuation of one libraries department for several months due to severe problems with the air-conditioning system.

Altogether, this lead to several delays in the content deliveries and thus required a significant amount of change management in the refinement progress.

Furthermore, while the libraries had provided estimates of how much material was available for enrichment before the project started, sometimes this estimation turned out not to be correct or precise enough, or the availability of the material they had selected was problematic due to rights issues. Consequently, some of the content had to be de-selected because the issues were too great and solving them was not possible within the project lifetime. This meant that new content had to be located in the consortium and associated partners' network to include in the refinement process.

Thankfully, due to a lot of good will at all partners and pro-active management of these issues, the refinement capacity of the technical partners could nevertheless be utilised to the fullest extent. In some cases the contingent of newspaper pages provided to the project had to be shifted between partners, and in other cases the scheduling of deliveries was adjusted to those other partners that were ready and could deliver their content ahead of schedule. At all times there was parallel processing of content from multiple content holders by the technical partners which required very precise tracking and communication.

Several tools were put into place in order to guarantee a timely delivery of content to The European Library and Europeana. A central calendar and tracking table (typically updated on a daily basis) on the project extranet form the basis for the organization of this work, and frequent communications have been made to guarantee that all affected partners are transparently made aware of any changes and the impact they might have on their tasks and deadlines.

Lessons learned and measures taken in the management of these issues will also be captured in a further deliverable D2.4 *Recommendations on best practices for refinement of digitized newspapers in Europeana* that will be delivered at the end of the project, in January 2015 (M36).

5. Conclusion

Refining 10 million newspaper pages from 12 different content delivering partners with 3 different technical providers proved to be a challenging objective in the Europeana Newspapers project. After carefully setting up a highly standardised and automated workflow and a schedule for each library, some issues where nevertheless encountered in the delivery of the content and the availability of some of the newspapers.

All of these concerns were solved in time and alternative solutions were identified, discussed and agreed by the consortium partners whenever this was necessary. This means that several libraries had the possibility to process more material than was originally planned, since other libraries were not able to deliver in time the exact contents that were spelled out in the original planning.

Unfortunately, due to the issues mentioned above, the refinement process was not yet finally concluded within the allotted time. However, it was still possible to produce quality results for the planned amount of about 10 million pages. Thanks to a flexible setup, the technical partners were able to accommodate to occurring changes, and have agreed to provide extra capacity for refinement beyond the scheduled end date in July 2014 (M30). It is envisaged that there will be more deliveries from libraries in the timeframe August – October 2014, which has been agreed with The European Library as feasible dates to still guarantee a presentation of these contents via the Europeana Newspapers portal browser and Europeana.

Many positive reactions by researchers[6] and the general public were received by project partners regarding the provision of newspapers as full-text resources, and - to some extent - with additional refinement, indicating the usefulness of this work, and the potential wider impact and outreach it can have for the project and TEL/Europeana.

[6] http://www.europeana-newspapers.eu/category/interviews-with-researchers/