Newspapers to the People
IFLA Newspapers/GENLOC Pre-Conference Satellite Meeting
Singapore, 14 – 15 August 2013

Europeana Newspapers - A Gateway to European Newspapers Online

Clemens Neudecker and Lotte Wilms, KB National Library of the Netherlands

1. Introduction

For many years now libraries and publishers are digitising extensive newspaper collections to meet the demand of their users and already millions of articles have been scanned. Despite this, access to these collections is still erratic and often limited to local access points. The project Europeana Newspapers is set out to greatly increase the accessibility to digitised newspaper collections – both those already in existence and those created in the future. The project brings together key stakeholders to make the process of digitisation more efficient in areas such as image refinement and the development of newspaper related metadata by providing best practice recommendations.

Within the project, ten million newspaper pages will be refined with Optical Character Recognition and Optical Layout Recognition by, respectively, the University of Innsbruck and Content Conversion Specialists GmbH. This whole workflow is supervised by the National Library of the Netherlands, one of the founding members of the impact Centre of Competence in text digitisation.

With this paper we would like to present our work with a wide variety of digitised European newspapers, ranging from 1618 to 2002 in over twenty languages, and with fourteen different metadata and master file formats. We want to share our experiences with the challenges faced when refining 10 million newspaper pages, from the selection of the material down to the final online presentation.

The search for an optimal workflow shall be addressed, but naturally also the choices made to process material from twelve different content providing institutions, such as the national libraries of The Netherlands, France, Austria, Finland who have been running big newspaper digitisation projects, but also smaller institutions that have just begun digitising newspapers on a larger scale such as the Landesbibliothek Friedrich Teßmann or the Hamburg State and University Library.

2. Background of Europeana Newspapers

A group of 18 European partner institutions have joined forces in the "Europeana Newspapers" project and will, over a period of three years, provide more than 18 million newspaper pages to the Europeana service.

Europeana is the online collection of European cultural heritage, from book to photo and from painting to artifact. All items have a thumbnail with a link back to the holding institution. There are currently over 26 million objects available from more than 22.000 institutions and 34 countries. The portal saw 3.5 million visits in 2012, from January to September alone. The actual number is probably even higher, as all content is also made available via an API[1].

The Europeana Newspapers Project (funded under the European Commission's Competitiveness and Innovation Framework Programme 2007 – 2013) started in February 2012 and aims at the aggregation and refinement of newspapers through The European Library. In addition, the project addresses challenges particularly linked with digitised newspapers:

- use of refinement methods for OCR, OLR/article segmentation, and named entity recognition (NER) to enhance search and presentation functionalities for Europeana customers,
- quality evaluation for automatic refinement technologies,
- transformation of local metadata to the Europeana Data Model (EDM),
- metadata standardization in close collaboration with stakeholders from the public and private sector.

3. Refinement of digitised newspapers

Structural refinement of digitised documents is a complex and often challenging task that can be broken down into various sub-tasks. Typically the whole process starts out with image capture. This can be done with either a regular scanner and the original paper document, or without, where the library digitises the mostly already available microfilm or microfiche. The pros and cons of scanning from the original or microfilm/microfiche have been widely discussed throughout the Optical Character Recognition (OCR) community[2]. Especially for newspapers, which are mostly very large pages with small characters, scanning from the original image instead of the mostly available microfilms might provide a better end result,not only in OCR quality, but also in the viewing experience of the user. Microfilm and -fiche are black and white, while scanning from the original is mostly in colour and thus resembles more closely the original document. Especially with old newspapers, the user sees all discolourations, the colour of the paper and where possible the photos and advertisements in their original print.

After the image capture, the next step is image enhancement. This again includes various processes such as cropping black borders form the pages, splitting double pages into single ones, straightening of

[1] For more information about Europeana, see http://europeana.eu/portal/aboutus.html
[2] See for example http://www.digitisation.eu/nc/training/knowledgebank/view/article/digitising-surrogates-scanning-from-microfilm-impact-case-study/

text lines and finally binarisation (the transformation into a black-and-white image). Binarisation is required by the OCR software to recognise text. To put it simply, black is something the software should focus on, while white is not. If an image is in colour, the difference between text in various colours and the background in various colours might be too much for the software to handle, resulting in misrecognised text.

Segmentation (also sometimes referred to as "zoning") follows, which aims to hierarchically break down a document into distinct sections consisting of text elements and non-textual elements such as illustrations or tables, then paragraphs, followed by lines, words and finally glyphs. This again is essential for the OCR software, with images being something that should not be scanned for text for example.

Only then text recognition is applied, sometimes iteratively. After the text has been detected and exported to a file , yet another (optional) processing step is verification of the recognised words against a dictionary and the further enrichment of the recognised text with semantic information, such as for example tagging of the person or place names being mentioned, called Named Entities.

Within Europeana Newspapers, the University of Innsbruck is the main provider for Optical Character Recognition. Around 8 million newspaper pages are foreseen for OCR through 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 extensive experience with OCR from various European project such as MetaE, IMPACT and others and are currently providing the technical and administrative backbone of the European eBooks on demand service EOD, which also includes OCR as part of its offerings.

The University of Innsbruck is making use of 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, the University of Innsbruck has been changing their OCR service to use the FineReader Engine SDK instead of the Recognition Server, because it gives more flexibility in the configuration while maintaining aptitude for large-scale processing. The version of the FineReader SDK which is being used is 11.0, the most recent release to date.

In addition to the roughly 8 million pages of "regular" OCR provided to the project by the University of Innsbruck, another 2 million pages will be OCR'ed with additional structural refinement, such as separation of articles and page classification, by Content Conversion Specialists (CCS), Hamburg. CCS makes use of their docWorks software technology that figuratively senses, recognises, reads, understands and controls documents. Along with state-of-the-art OCR, this software allows also structure recognition, a very important feature for digitised newspapers. Adding this information to the scans once again helps with the final presentation of the newspaper. Instead of accessing the whole newspaper page at once, the user sees only the selected article highlighted. It also becomes possible to request the text of a single article, instead of the whole page. This provides benefits for the presentation of newspapers, as they often consist of much text per page. Narrowing this down to a single article increases the readability of such a newspaper.

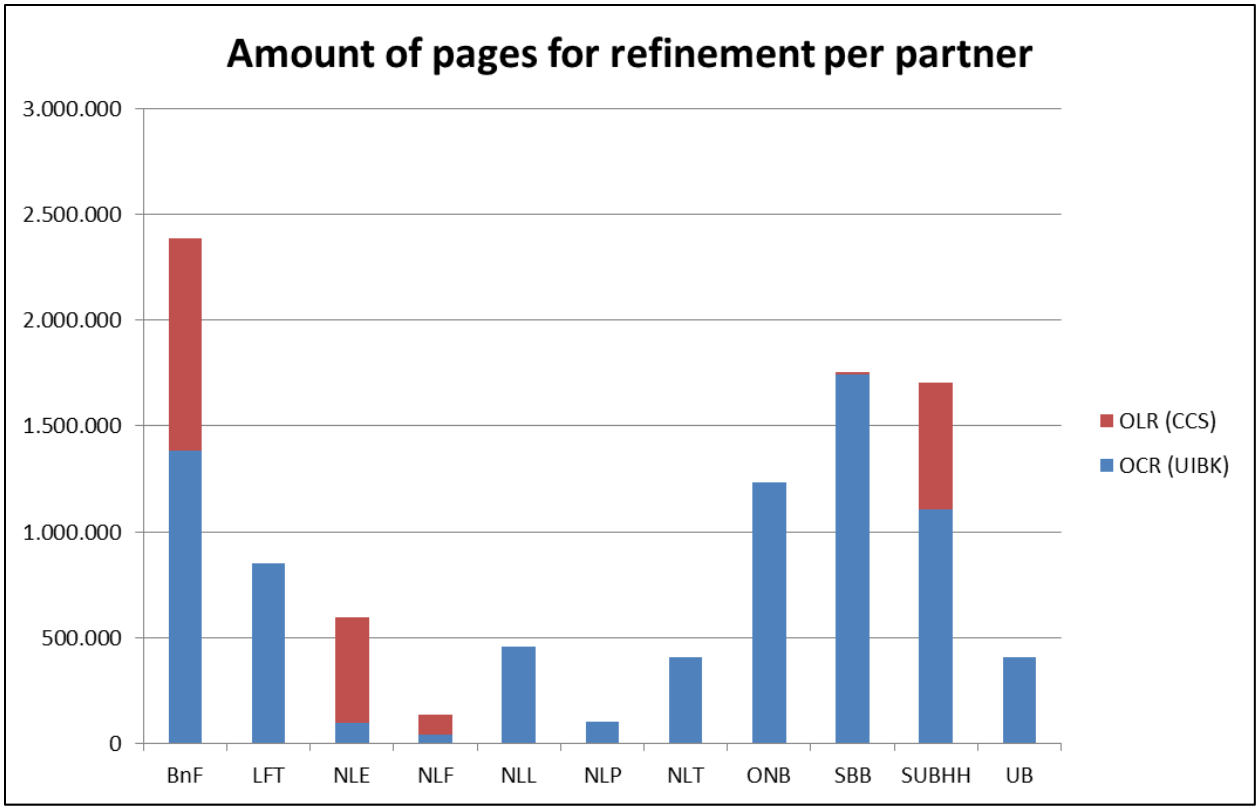
For a subset of the OCR'ed content from partners in either Dutch, English or German language, the National Library of the Netherlands will provide recognition of named entities such as person or place names. The NER system will build on preliminary work that was carried out in the IMPACT project. The Stanford NER tagger will be used and further adapted for the project. These adaptations are mainly focused on the use on historical language, such as matching a historical variant to a modern lemma and increased robustness for OCR errors. As can be seen in the paragraphs below, the data set of Europeana Newspapers covers a wide range of years, but most of the content available was published before 1940.

4. The data set

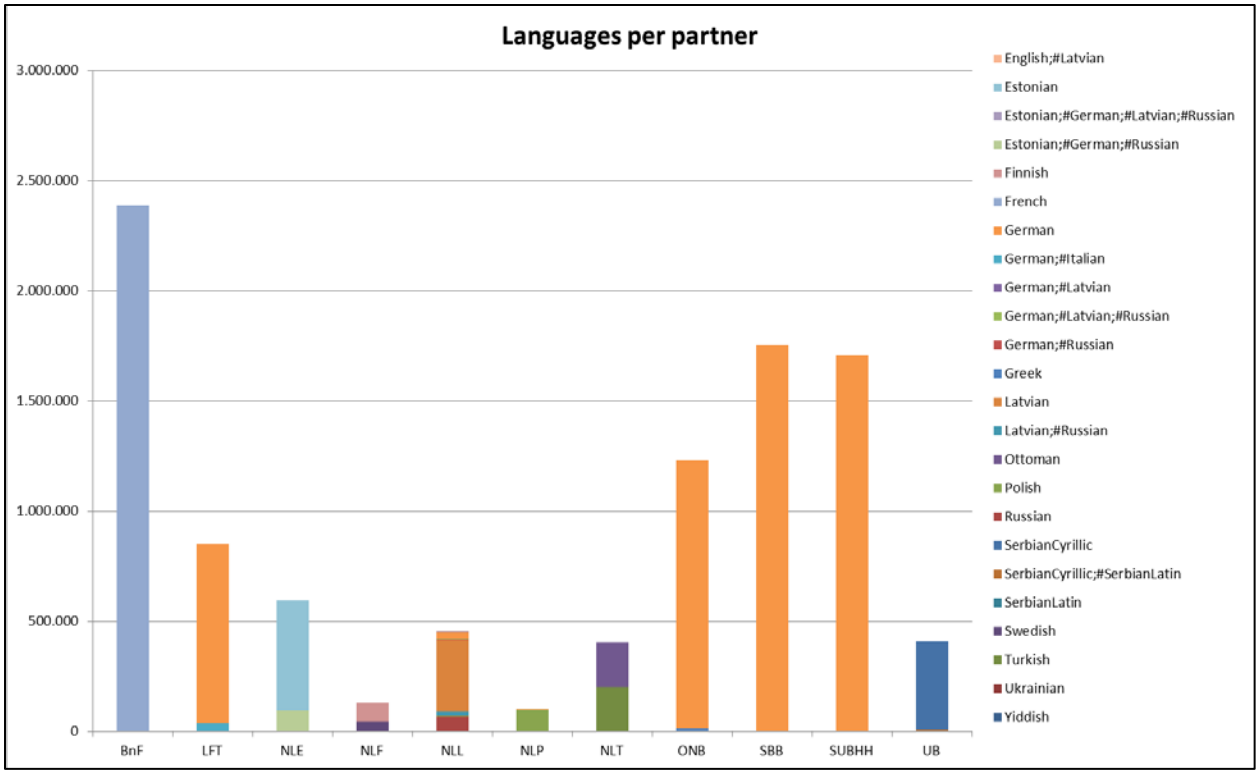
The Europeana Newspaper data set currently consists of over 18 million items, divided over 12 partners[3]. The partner libraries contributing data to the project are the National Library of France (BnF), the National Library of the Netherlands (KB), Dr. Friedrich Teßmann Library (LFT), the National Library of Estonia (NLE), the National Library of Finland (NLF), the National Library of Latvia (NLL), the National Library of Poland (NLP), the National Library of Turkey (NLT), the National Library of Austria (ONB), State Library of Berlin (SBB), State and University Library of Hamburg (SUBHH) and the University of Belgrade Library (UB). Not all 18 million items will be ingested in the refinement workflow, as can be seen in the table below. Several partners will still add more material to the data set at a later date. Also, the project has welcomed associated partners in 2013, who have the option to contribute data via the project to Europeana. These libraries are not shown in the overviews[4]. The selection for NER has not yet been finalised at the time of writing, as these pages will be selected based on OCR success rates.

Library	No Refinement	OCR (UIBK)	OLR (CCS)	Total
BnF		1.385.727	1.002.761	2.388.488
KB	1.921.946			1.921.946
LFT		857.485		857.485
NLE		94.701	500.256	594.957
NLF		40.665	91.428	132.093
NLL		454.639		454.639
NLP		99.795		99.795
NLT		406.481		406.481
ONB	5.691.024	1.232.434		6.923.458
SBB		1.745.400	10.000	1.755.400
SUBHH	508.800	1.105.200	602.200	2.216.200
UB		408.181		408.181
Total	**8.121.770**	**7.830.708**	**2.206.645**	**18.159.123**

[3] For the full list of project partners see http://www.europeana-newspapers.eu/consortium/project-partners/.
[4] For the list of associated partners, see http://www.europeana-newspapers.eu/consortium/associated-partners/.

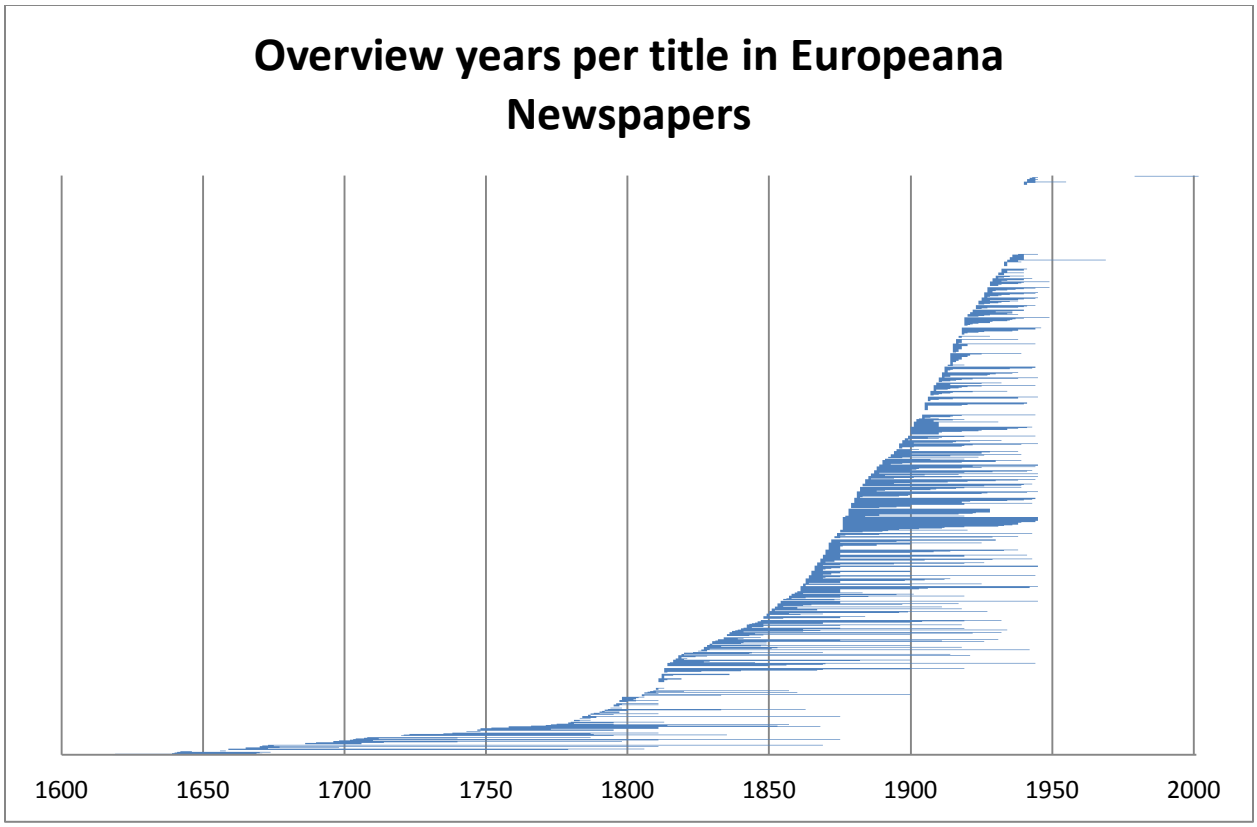


This graph shows the amount of pages supplied for refinement per content holding partner, with a subdivision between the work done by Content Conversion Specialists and that of the University of Innsbruck. The total amount of pages for refinement is currently 7.9 million for UIBK and just over 2.2 million for CCS. Each partner has indicated the font type per newspaper title that will be ingested into the refinement workflow. This is necessary for the setup of the refinement software and can be linked to the language up until a certain point.



The newspaper set in Europeana Newspapers contains 20 languages in total. Several newspapers combine two or more languages in one title. The language with the least amount of pages is Ukrainian and the language that is most common in the set is German. The various languages in the newspaper set often come with their own characteristics, such as font type. The most popular language in the set, German, is often printed in a gothic typeface, whereas the second most common language, French, is typically printed in a Latin type. Serbian is printed in both Cyrillic and Latin, so these titles have been indicated separately. A special case is the National Library of Turkey who hold digitised newspapers in Ottoman alphabet. Currently there is no OCR on the market that would support recognition of Ottoman characters. Some testing with Arabic OCR and training Tesseract did not deliver very encouraging results, so the project decided to split the Turkish set in half with 50% of the pages being in Latin alphabet and scheduled for refinement during 2013. The other half of the set will be Ottoman based on the developments that can still be expected with regard to Ottoman support in for example Abbyy FineReader in the course of 2014.

The covered years per title, library and language differ greatly over the entire selection in the Europeana Newspapers Project. The following graph gives an overview of the publication years of the selected newspapers sorted by date of publication.



It is interesting to note the differences in the cut-off date of the newspapers with regard to the copyright. In the partner view graph it is noticeable that some institutions have opted to adhere to the 'safe'-date of 1870-1900, whereas other institutions have made arrangements with copyright holders and have the possibility of providing more recent newspapers up to 1940 for aggregation. There are a few institutions that can even include modern newspapers from 1940 onwards.

5. Workflow

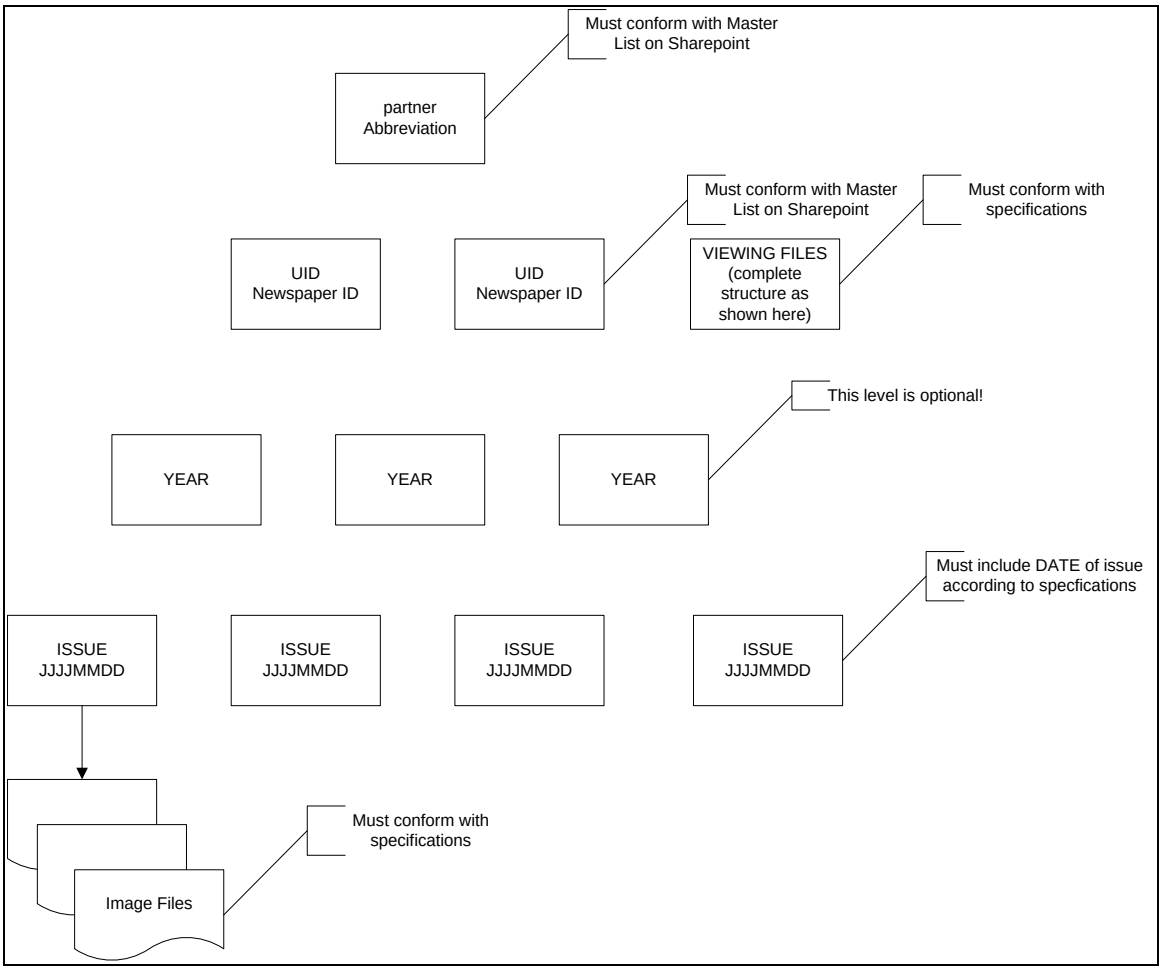
Deciding on a workflow for the refinement of 10 million newspaper pages, by three service providing institutions is a challenging task. However, the partners in the Europeana Newspapers Project have set up the optimal workflow for this undertaking and will combine efforts in creating a valuable new resource for the Europeana users.

The master list is where the process of refinement begins. The libraries indicate their newspapers there, with all relevant information concerning the newspaper titles they wish to contribute. Consequently, they prepare their data according to the specifications with the help of the tools made available by the University of Innsbruck, before sending it via hard disk to either UIBK or CCS. The material gets processed accordingly (OCR/OLR and NER), while the libraries can keep track of the status via various tracking tools on the project extranet. Finally, the libraries receive their enriched newspapers back as zipped METS/ALTO packages (corresponding to Submission Information Packages in OAI terms) according to the ENMAP profile defined in a work package specifically set up for this task. UIBK and CCS also forward all material to Europeana, for the final publication via the newspaper content browser

produced by The European Library (TEL), as to combine the 10 million refined newspapers pages into a unique resource for the European public.

Delivery from content holders to refinement partners needs to happen according to a strict system, in order to always guarantee a controlled processing of the approximately 10 million files that will be put through the refinement workflows altogether. The system is constituted by exact requirements for directory structure and file naming in the delivery package to guarantee the controlled processing of the approximately 10 million files delivered from content providers. Data that does not precisely follow this schema will be rejected and needs to be re-delivered according to the exact specifications. To check the data does indeed adhere to the diagram below, UIBK has provided all libraries with a tool that goes through the folders and displays any errors it sees. The library can then correct these errors until none more show up. This ensures no data is sent to the service providers with the wrong folder structure. This is essential to be able to monitor the data as it goes through the various steps of the refinement workflow. Without it, pages or even whole titles might get lost or disconnected from their corresponding metadata.

Below diagram illustrates the structure of the delivery package and the according file naming requirements.



Delivery of files from libraries to refinement partners is dealt with using external hard disks. The drives must have a connector for either USB 3.0 or eSATA protocol in order to ensure quick transfer times and should be formatted with the NTFS file system. The hard disks contain complete newspaper titles, with the master (binarised) images and viewing copies packed together in a ZIP archive. Due to the large file size of most of the digitised newspapers, we have chosen to binarise all images before the actual OCR process. This dramatically decreases the images in size, thereby making it much easier to send them to the service providing partners. Since this is a step which is necessary for the OCR workflow anyway, the libraries can do this in-house with the provided tools before sending the files. However, we do make sure that the user does not notice this conversion, as this is not beneficial for their viewing experience.

To ensure we have the best results for the user, there are two options with regard to viewing copies of the images. One option is applicable to libraries that have an image server in their institution. The newspaper browser can then just request the image needed, which is mostly in colour or greyscale. Another option is the creation of a colour/greyscale viewing copy, which is smaller than the original scan but maintains a high enough quality for the user to easily navigate an entire newspaper page. This choice can be made within the provided pre-processing tools. If a library needs to create the viewing copies, this is done simultaneously with the binarisation and stored in the appropriate folder structure automatically.

6. Newspaper browser

All newspaper images processed within the Europeana Newspaper project will be delivered to both the library providing the image, but also to the project partner The European Library (TEL). Being part of the Europeana Foundation, they will ensure all content will be aggregated for Europeana. At the same time they will provide the public with a single portal to search, browse and read the collection of the project and associated partners providing data to the project.

All decisions in the refinement workflow have been made with this portal in mind to guarantee the users gets the optimal experience when visiting the rich collection of European newspapers. At the time of writing, the wireframes of the portal have been made available for feedback[5]. Later in 2013, a beta version should be up and running with a final version at the end of the project in December 2014.

7. Conclusion

The Europeana Newspapers data set consists of a wide range of European newspapers from the 17th to the 20th century in twenty different languages and combinations thereof with a current total of 16 million pages, with more than half of those intended for refinement.

The selection of the set was done by the participating libraries themselves and takes into consideration the condition of the material, the demands of their users and the access conditions with regards to copyright. By selecting the titles with the utmost care, the overall set provides an insight into the European newspaper collection as a whole. The Europeana Newspapers public will be able to enjoy the elaborate selection of the project partners and search or browse through three centuries of European news articles and events from eleven countries, but at the same time get an insight into the past of the whole of Europe. This is all made possible by setting up the optimal workflow for the refinement of 10 million newspaper pages from 12 different institutions. The choices made for the refinement process have been a balance between the most practical options, while at the same time ensuring the best viewing experience for the final users.

Europeana Newspapers will provide a rich user experience with access to more than 18 million newspaper pages from partners around Europe. With this, we hope to not only reach the users of such material, but also other libraries and content holding institutions to encourage them to digitise, OCR and make available as much newspapers as possible. This way, the future users of Europeana can not only benefit from our project and associated partners, but from every institution with newspapers in their collection, making the provided corpus even richer and the possible research questions to be unleashed on the set even more elaborate.

[5] See http://www.europeana-newspapers.eu/building-a-content-browser-for-digital-newspapers/