

DELIVERABLE

Project Acronym: Europeana Newspapers

Grant Agreement number: 297380

Project Title: A Gateway to European Newspapers Online

D2.1 Data set for refinement of digitised newspapers

Revision: 1.0

Authors: Clemens Neudecker, KB
Lotte Wilms, KB

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	08-11-2012	Clemens Neudecker	KB	Created
0.2	11-01-2013	Lotte Wilms, Clemens Neudecker	KB	Updated
0.3	25-01-2013	Tiina Hölttä	NLF	Updated
0.4	21-02-2013	Ulrike Kölsch	SBB	Internal review
1.0	21-02-2013	Clemens Neudecker	KB	Finalised after internal review

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This deliverable contains original unpublished work except where clearly indicated otherwise.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	4
2. Content selection	5
2.1 Criteria	5
2.2 Process	5
2.3 Initial data set	6
2.4 Evaluation data set.....	6
3. Europeana Newspapers data set.....	7
3.1 Overview	7
3.2 Visualisations	7
3.2.1 Volume.....	8
3.2.2 Fonts.....	9
3.2.3 Language.....	10
3.2.4 Timeframe.....	12
3.3 Access conditions.....	14
4. Library data sets	15
4.1 BnF – Bibliothèque nationale de France	15
4.2 KB – Koninklijke Bibliotheek.....	15
4.3 LFT – Landesbibliothek Dr. Friedrich Tessmann	16
4.4 NLE – Eesti Rahvusraamatukogu	17
4.5 NLF – Kansalliskirjasto.....	17
4.6 NLL - Latvijas Nacionāla Biblioteka.....	18
4.7 NLP - Biblioteka Narodowa	18
4.8 NLT - Milli Kutuphane Baskanligi	19
4.9 ONB - Österreichische Nationalbibliothek	19
4.10 SBB - Staatsbibliothek zu Berlin.....	20
4.11 SUBHH - Staats- und Universitätsbibliothek Hamburg	21
4.12 UB - Univerzitet u Beogradu.....	21
5. Conclusion	23

1. Executive Summary

The main objective of this Work Package is to enhance and refine digitised newspaper pages with Optical Character and Layout Recognition as well as Named Entities Recognition as part of the general aggregation process. Altogether more than 10 million digitised newspaper pages will be processed with the various refinement processes in the project.

This document describes the data set of digitised newspapers that has been selected for refinement in the scope of Work Package 2 and ingested into Europeana through Work Package 4.

The document begins with an outline of the process for selecting the source material and the specific criteria that have been applied therein, such as suitability and availability of the source material, relevance to end-users etc. It then continues with a description of the mechanisms that have been put into place to guarantee a smooth processing of the large amounts of material that have been selected. Metadata on all newspaper pages that will be refined in the project are collected centrally and the progress of refinement is then tracked via a central list.

This is followed by a detailed record of the overall data set with regard to the amount of titles and pages selected as well as the languages and timeframe covered by the selection. The characteristics of the overall data set have been visualised to provide a better overview of the nature of the content.

Finally, the individual contributions of each content providing partner are described in a bit more detail.

2. Content selection

2.1 Criteria

The following criteria have been guiding the selection of suitable digitised newspaper titles:

- Availability

Naturally, the possibility to make the selected newspaper available on a very open, non-restrictive basis is one of the key aspects in the selection process. Libraries are required to provide full metadata with a [CC-0](#) license in order to get their data into Europeana, as specified in the [Europeana Data Exchange Agreement](#) (see also 3.3 – Access conditions).

- Relevance to end-users

There is a strong interest at the libraries to select material that is also of particular interest (e.g. often requested) to the end user.

- Digitisation quality

Given the complexity of the refinement process and the technical constraints inherent in the software tools it becomes obvious that it will only make sense to select a digitised newspaper for refinement when it is also likely that the process will return reasonable results. Therefore only high resolution uncompressed master images will be processed with the refinement technologies.

- Document characteristics

Document characteristics such as the language of the document, the complexity of the layout, the font that has been used for printing and the overall condition of the document also play an important part in the selection process.

- Technical considerations

Other technical considerations, e.g. with regard to the file formats and metadata standards used by the content providers, etc.

2.2 Process

The selection of the data set proceeded in several steps. In the first months of the project, the process was discussed among the technical partners, with input from a content provider. The decision was made to gather all information on one location, namely a Sharepoint list, to be filled out by the libraries themselves. This 'master list' also evolved over time, as detailed in deliverable D2.2.

After a proposal for the refinement workflow was agreed upon by the technical partners, the content providers were asked to provide feedback during the WP2 workshop, held on 14 May 2012 in Hamburg. Several adjustments were made to the process and then finalised with the creation of the *Master list* on the Europeana Newspapers Sharepoint in June. The libraries were asked to input their data before the end of July for the first iteration of the selection. The list was then again discussed in the technical meeting in Innsbruck on 29 September 2012, where the libraries were

asked to refine some titles metadata and enter the latest additions. In November 2012, the final list was used as input for the refinement workflow and the scheduling of the process.

A small number of partners have not been able to fill out the list completely in 2012, but will do so as soon as possible. This does not have a negative effect on the timing of the refinement process.

2.3 Initial data set

In order to get a first impression of the kind of content available from the libraries already very early on in the project, and also to prepare sample input images for some early testing and experimentation with the tools, an initial data set was compiled during the first three months of the project.

This initial data set contains about 100 exemplary scanned images from the digital newspaper collections of all content holders in the project.

2.4 Evaluation data set

For the purposes of evaluation in the scope of WP3, a special database has been set up by the University of Salford. This database collects a subset of the overall data set (the “evaluation data set”) which will be used to assess the performance of the various refinement methods through scenario-driven in-depth evaluation using ground-truth.

University of Salford MANCHESTER

PRImA Pattern Recognition & Image Analysis

Introduction Dataset Adv Search Your Cart Login

Europeana Newspapers Evaluation Dataset

This online repository is the main point of reference for all activities related to evaluation within the scope of the Europeana Newspapers project. Its main goal is to provide a representative collection of all the types of newspapers which are and/or might be subject of ongoing or future digitisation activities. As such, it is hosting scanned images, metadata and ground truth (a representation of the ideal result of a processing step like OCR or layout analysis) on the level of individual newspaper pages.

Statistics

Collection

- 961 total images
- 8 PAGE files
- 32.3 GB of space used
- 34.44 MB per image (avg)

Last updated: 28/1/2013 8:59

Usage

- 18,606 thumbs browsed
- 500 images accessed

Since: 13/9/2012

Total number of pageviews since 2 July 2012: 2,889

W3C XHTML 1.0 Best viewed in 1024x768 - Maintained by: [Christos Papadopoulos \(e-mail\)](#), [Apostolos Antonacopoulos \(e-mail\)](#) - © 2008-12 W3C CSS

3. Europeana Newspapers data set

3.1 Overview

The Europeana Newspaper data set currently consists of over 16 million items, divided over 12 partners. Not all of these items will be ingested in the refinement workflow, as can be seen in the table below. Several partners will add more material to the data set at a later date.

	BnF	KB	LFT	NLE	NLF	NLL	NLP
No Refinement		1.921.946					
OCR (UIBK)	1.385.727		857.485	94.701	40.665	460.781	83.648
OLR (CCS)	1.002.761			499.962	91.428		
Total	2.388.488	1.921.946	857.485	594.663	132.093	460.781	83.648

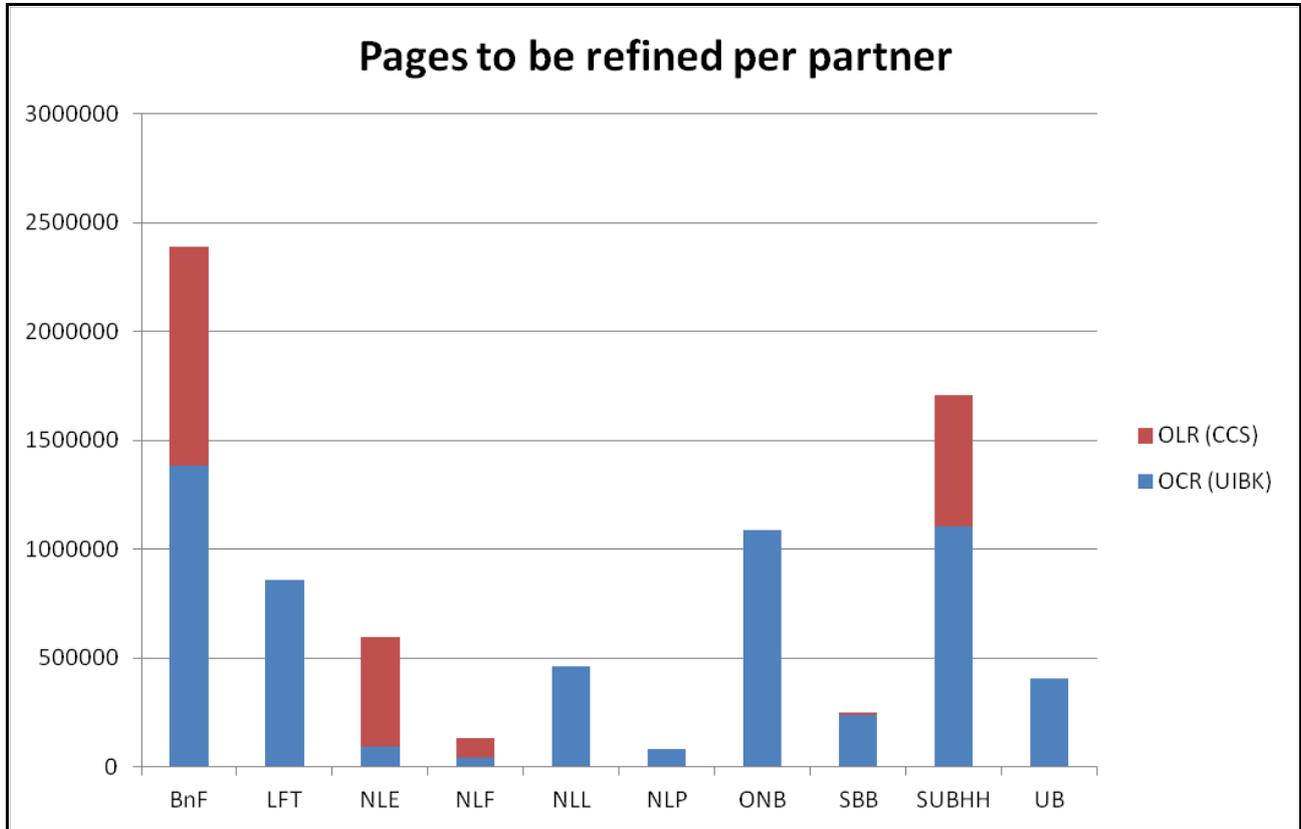
	NLT	ONB	SBB	SUBHH	UB	Total
No Refinement	8.990	5.691.024		508.800		8.130.760
OCR (UIBK)		1.090.308	238.200	1.105.200	408.181	5.766.836
OLR (CCS)			10.000	602.200		2.206.351
Total	8.990	6.781.332	248.200	2.216.200	408.181	16.103.947

3.2 Visualisations

To show a clear view on the composition of the collection, several visualisations are made of the Europeana Newspaper Projects (ENP) data set. The visualisations are included below and subdivided into the categories 'Volume', 'Fonts', 'Language' and 'Timeframe'. A short commentary is given with additional information per graph.

The visualisations only show the material that will be processed in the refinement workflow (i.e. with OCR or OLR). The amount of material that is part of the project, but will not be refined with either OCR or OLR can be seen in the table above.

3.2.1 Volume

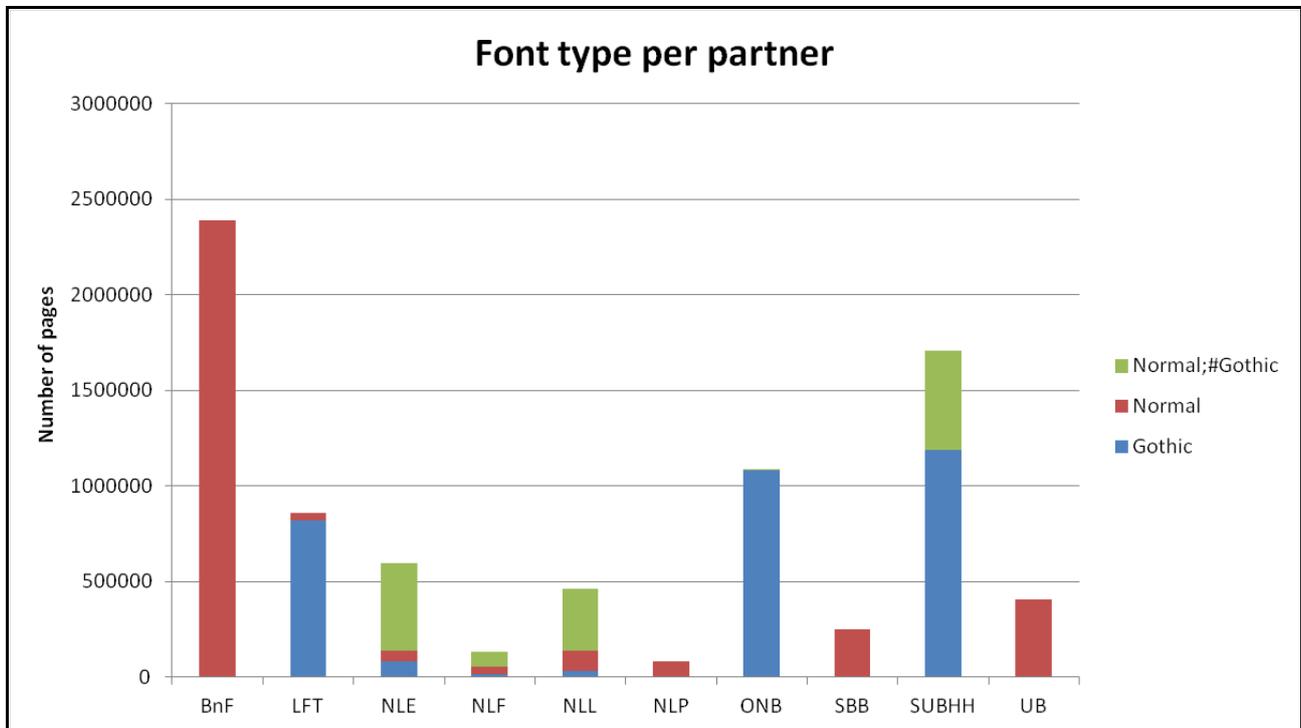


This graph shows the amount of pages supplied for refinement per content holding partner, with a subdivision between the work done by Content Conversion Specialists (CCS) and that of the University of Innsbruck (UIBK). The total amount of pages for refinement is currently 5.8 million for UIBK and just over 2.2 million for CCS.

Please note that the amount of pages for the National Library of Turkey is missing in this overview due to the issues with their font type (which is not currently supported by the off-the-shelf OCR solutions available in the project).

A few other libraries have not yet inputted their total collection, as there are some difficulties with their data selection.

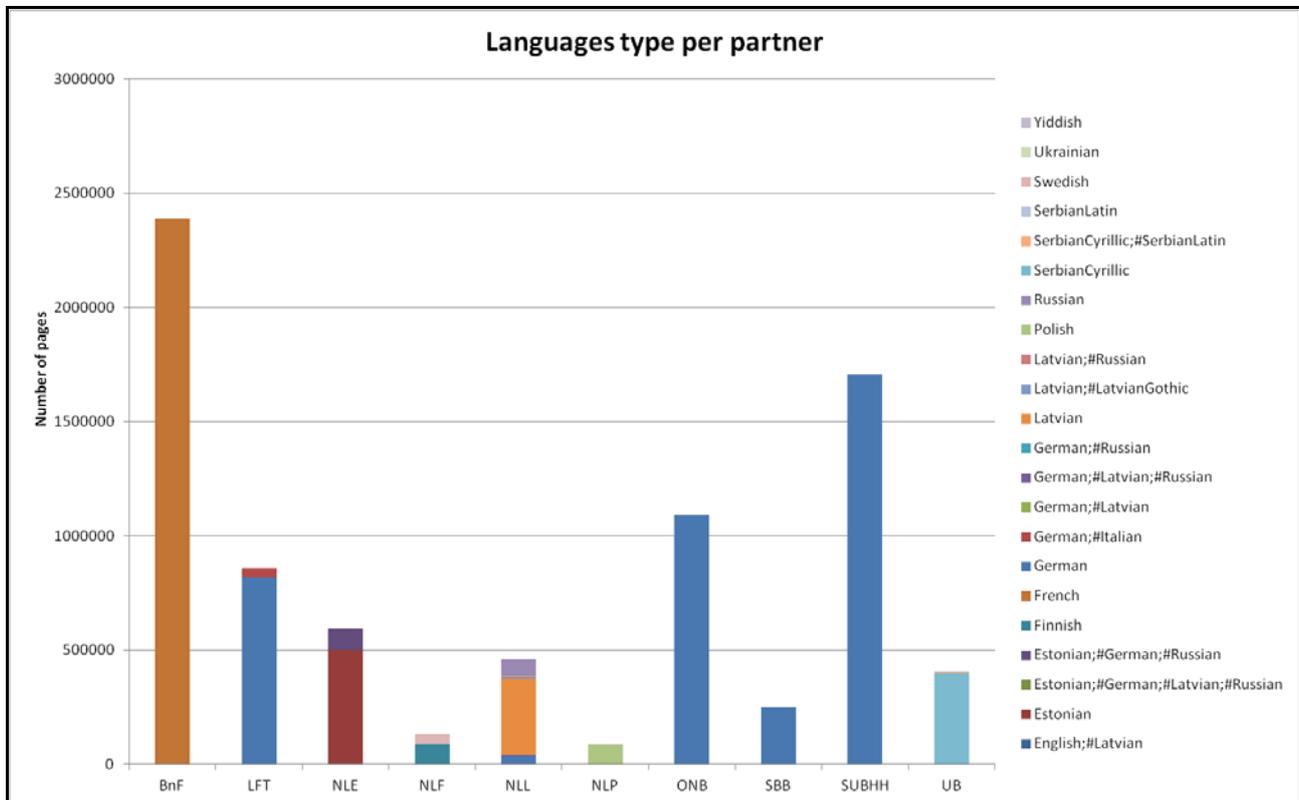
3.2.2 Fonts



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. A more elaborate overview of the font types per language can be seen in the paragraph below.

Please note that the National Library of Turkey is again not included in this graph. Their initial selection consists entirely of the Ottoman Turkish alphabet, which is not compatible with the current OCR techniques. A work around has been proposed, which is to select only newspaper pages in Latin alphabet from the National Library of Turkey for the refinement process.

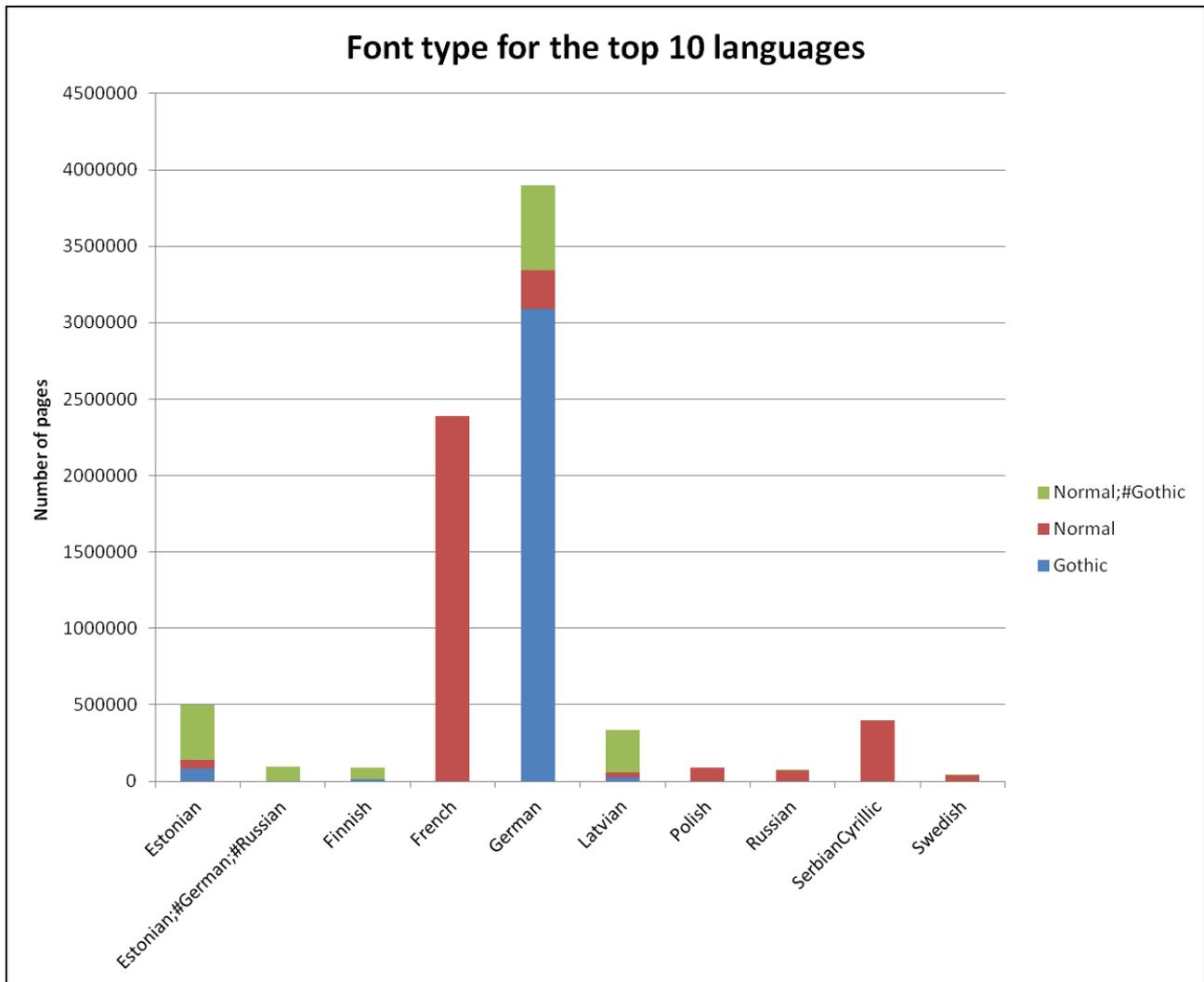
3.2.3 Language



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.

When looking at the two workflows for CCS and UIBK, UIBK has the most variation with 14 languages and combinations thereof and CCS the least with only 6 languages. However, CCS will process more pages in French than UIBK. On the other hand, the amount of German pages that will be processed at UIBK is more than the total amount to be processed at CCS (almost 3.3 million pages).

There is a definitive lack of English language content in the data set – the idea is to mitigate this with the addition of new content through the associated partners that will join the project in the course of 2013.



The various languages in the newspaper set often come with their own characteristics, such as font type. As can be seen in the graph above, 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 small subset of the Latvian content is in “Latgalian”, i.e. in archaic spelling. Tests have been made with a special OCR module developed by Abbyy and the National Library of Latvia and have shown good success rates for this historic material.

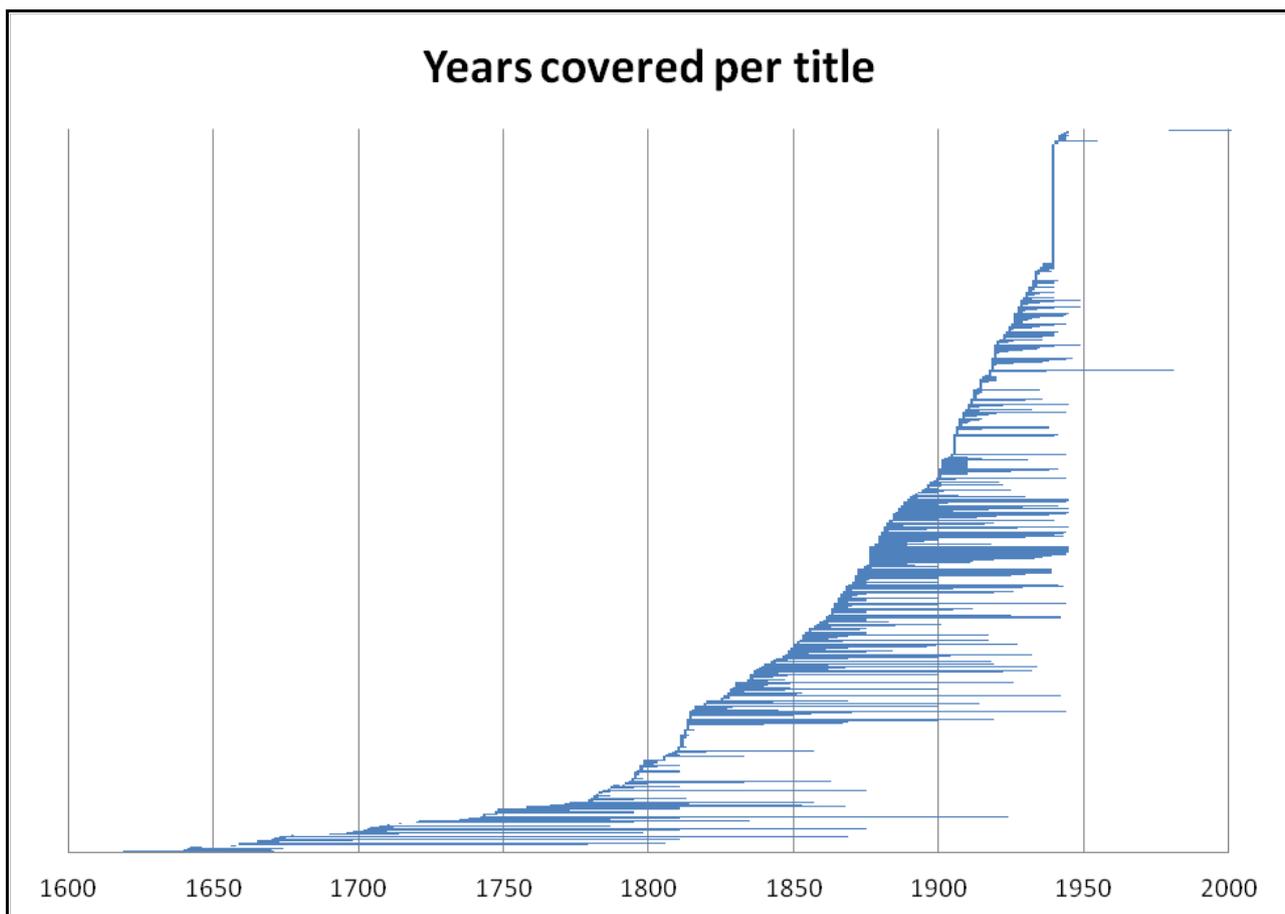
This graph only shows the top ten languages when looking at the amount of pages, due to the high variety of languages as seen in the first graph.

3.2.4 Timeframe

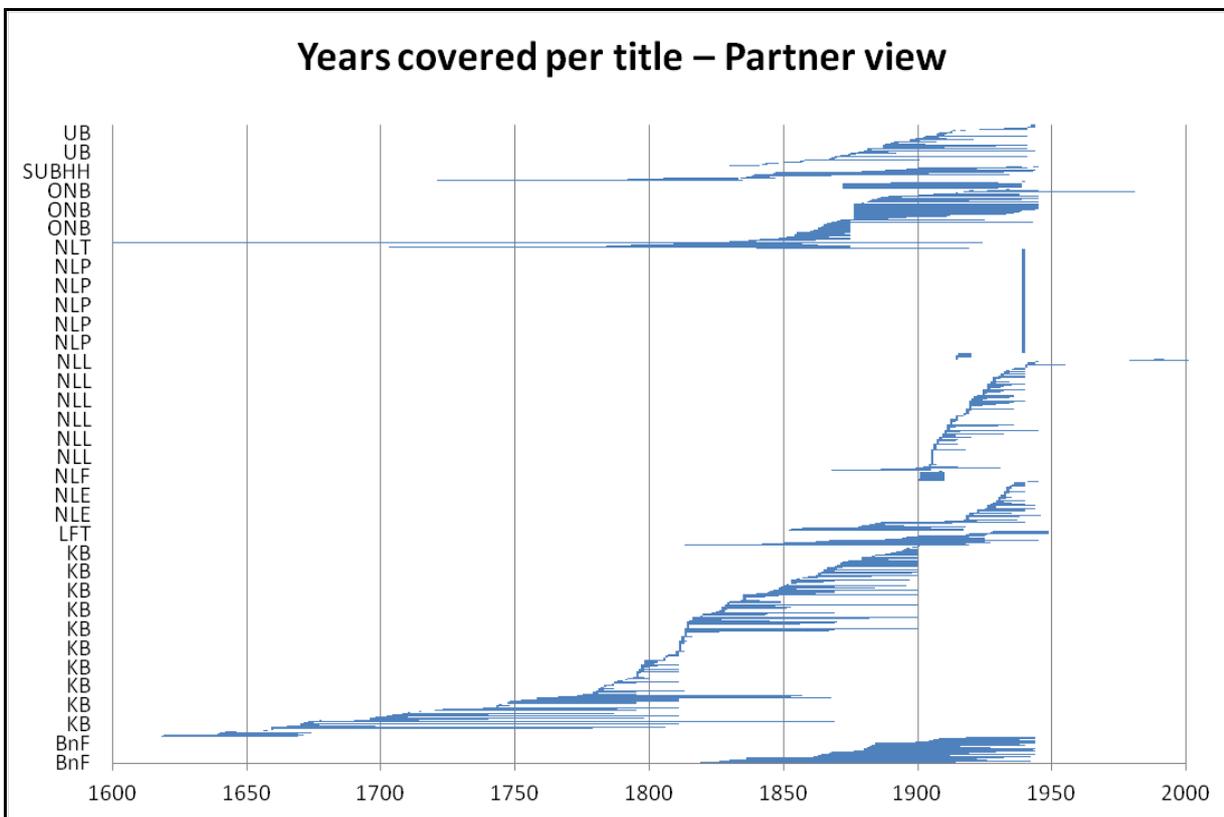
The covered years per title, library and language differ greatly over the entire selection in the Europeana Newspapers Project. The following graphs give an overview of the publication years of the selected newspapers in different views, starting with the entire collection sorted by date of publications, followed with the entire collection sorted by institution and finally the entire collection sorted by language.

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.

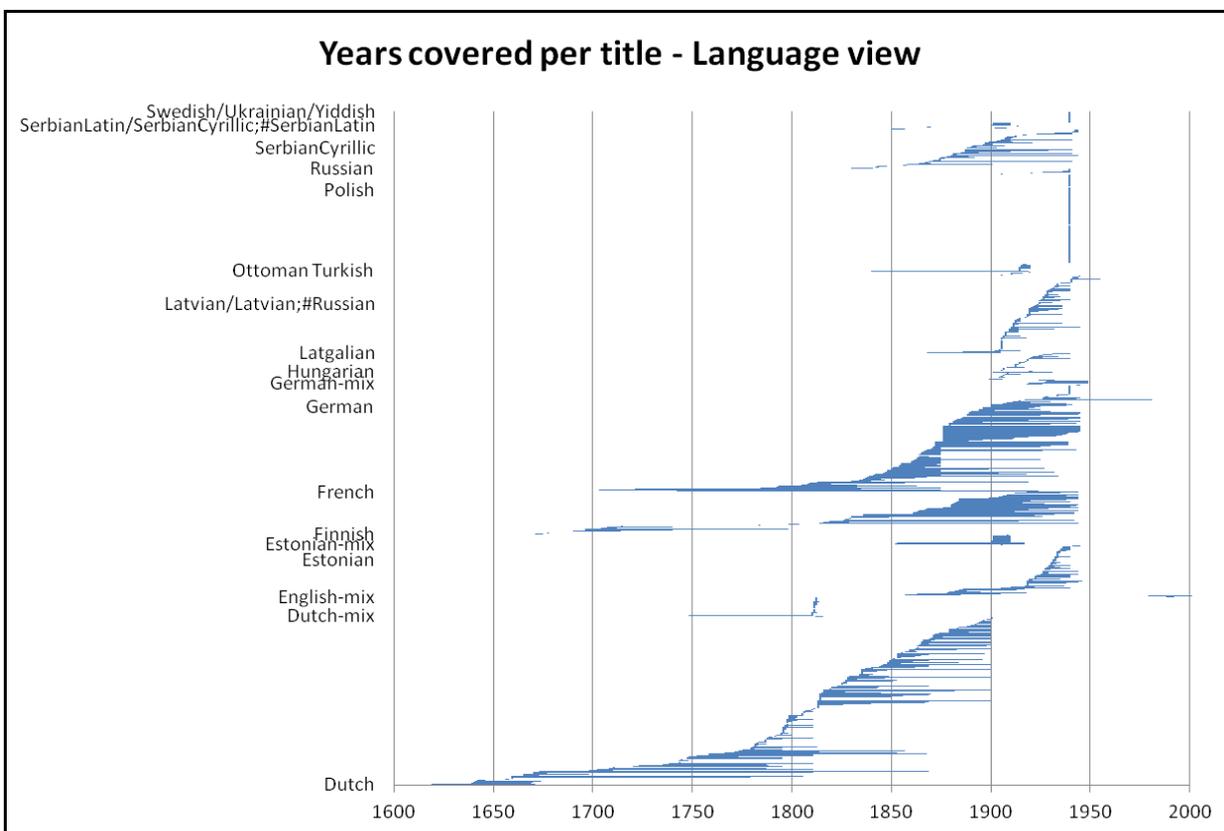
Please note that in order to provide a better overview of the total selection these visualisations do include also those titles that will not be processed with the refinement workflow, but are part of the overall content to be aggregated.



Years covered per title – Partner view



Years covered per title - Language view



3.3 Access conditions

For the online presentation system (“Newspaper Browser”) to be built in Work Package 4, the libraries were presented with a choice of six options for the access conditions that are going to be implemented for their selection:

- (1) Full page view: The user gets to see the full page, including OCR
- (2) Snippet view: The user only gets to see a snippet of the full page
- (3) Plain text view: The user gets to see the plain text, but not the digital facsimile
- (4) Metadata only: The user gets to see the descriptive metadata about the object only
- (5) Full page view – using remote image server: Same as (1), with the difference that images are provided through a remote image server by the library
- (6) Snippet view – using remote image server: Same as (2), with the difference that images are provided through a remote image server by the library

Options 1-4 are also outlined in more detail in a document that was produced by TEL and circulated to all library partners in November 2012.

4. Library data sets

All libraries participating in Europeana Newspapers have selected the titles for inclusion in the Europeana Newspaper Projects data set themselves, with regard to their own collection criteria and wishes. As their collections naturally differ greatly per partner, each library has provided several paragraphs on their institution, newspaper collection and the main drivers for selection of the set for Europeana Newspapers.

4.1 BnF – *Bibliothèque nationale de France*

The Bibliothèque nationale de France (BnF) is one of the largest public and research library in the world today. The BnF offers access to its digital library [Gallica](#) obtained through the library's commitment to the digitisation of selected items of its collections. Gallica now contains around 2 million digitised documents: manuscripts, sound materials and music score, books, images and over 800,000 newspapers issues, in French and other languages.



The BnF contributes **39** newspaper titles to the project, comprising **2.388.488** pages in total.

4.2 KB – *Koninklijke Bibliotheek*

The KB National Library of the Netherlands is a research library with a broad collection in the fields of Dutch history, culture and society, and as national library collects and maintains all publications that appear in and about the Netherlands.

The KB initiated the Databank of Digital Daily (DDD) newspapers project at the end of 2006. The project has realised the large-scale digitisation of Dutch national, regional, local and colonial newspapers and makes these freely accessible on the Internet via www.kranten.kb.nl. The DDD portal currently contains eight million pages, from the first newspaper dated 1618 up to the newspapers of the twentieth century. It is one of the largest digitisation projects of historical material in the Netherlands. All newspaper pages have been processed with OCR and layout recognition software and are thus searchable on article level. The selection for Europeana Newspapers entails all material from the Databank of Digital Daily newspapers that is in the Public Domain.



The KB contributes **201** newspaper titles to the project, comprising **1.921.946** pages in total.

4.3 LFT – Landesbibliothek Dr. Friedrich Tessimann

The Teßmann Library in Bozen, Italy, is the provincial library for the German and the Ladin speaking population in South Tyrol.

The “[Digital Newspapers Archive](#)” of the Teßmann Library went online in 2010. It presently offers access to 43 digitised newspapers and periodicals published in the area of historic Tyrol (boundaries before 1918) in the 19th and beginning 20th century. The digital copies were made partly from microfilm, partly from the original print versions; the collection offered in the portal is a junction of the newspaper stocks of several institutions from North, South and East Tyrol as well as the Trentino, as a project of the “European region Tyrol”, and in order to be able to offer a virtually complete collection of historical regional newspapers. The scope of the digitisation and of the portal’s creation was to preserve the originals and to render them accessible on the internet.

The main criteria for the selection of titles for the Europeana Newspapers Project were the following:

- Publication period and copyright: it was important that the issues of a newspaper were no longer protected by copyright.
- Local historical relevance: periodicals published in Tyrol that were of particular importance for that age as well as nowadays for historians and annalists.



The LFT contributes **15** newspaper titles to the project, comprising **857.485** pages in total.

4.4 NLE – Eesti Rahvusraamatukogu

The National Library of Estonia functions as a national library, parliamentary library, research library, library research and development centre and a cultural centre.

The creation of digital collections has taken place at the National Library since 2003, when a project-based digitisation of Estonian newspapers issued before 1940 began. The digitised Estonian newspapers database *DEA* provides access to Estonian periodicals from 1821 - 1944 and the newspapers of the Estonian diaspora since 1944.

At the moment 1.2 million images are accessible via DEA at <http://dea.nlib.ee>. At the end of the Europeana Newspapers Project half of them will be available in Europeana as full text. The main criteria for selecting the titles for Europeana Newspapers were the importance of the publication in its time and the quality of master files, enabling the best possible result for OCR.



The NLE contributes **42** newspaper titles to the project, comprising **594.663** pages in total.

4.5 NLF – Kansalliskirjasto

The National Library of Finland is the oldest and largest scholarly library in Finland as well as one of the largest independent institutes at the University of Helsinki. It is responsible for the collection, description, preservation and accessibility of Finland's printed national heritage and the unique collections under its care.

The National Library of Finland has an almost complete collection of newspapers published in Finland (regardless of language) from 1771 to the present. The newspapers are received as deposit copies from the printers, and they are catalogued in the FENNICA database. The Library has domestic newspapers from the period 1771-1910 and their subject-related index 1771-1890 have been digitised and can be read online in the Finnish Historical Newspaper Library (www.digi.nationallibrary.fi). Digitised newspapers from 1911 onwards are accessible only at Finnish Legal Deposit Libraries all over the country, using dedicated workstations.

The addressed Digital Collections of the National Library of Finland contains about 6.8 million digitized pages: newspapers, journals and ephemera. Of these page volumes, over 2 million pages consist of newspapers. The volume of newspaper issues is around 480,000.

NLF's selection criteria for the OCR process in Europeana Newspapers are to fill gaps in the Finnish Historical Newspaper Library. The title selections for OLR are done purely by ranking the volumes of all published titles and pages digitised in Finland in period 1900–1910 after studying literature and NLF's databases. The final choice consists of three newspapers (Hufvudstadsbladet, Uusi Aura and Wiipuri), which belong to the largest newspapers of their era ranked by volume.



The NLF contributes **11** newspaper titles to the project, comprising **132.093** pages in total.

4.6 NLL - *Latvijas Nacionāla Bibliotēka*

The NLL (Latvijas Nacionāla Bibliotēka) is the national library of Latvia.

Over time NLL has created many individual digital collections. In 2000, the first digital collection of Latvian newspapers was published online. The mass-digitisation of newspapers started in 2009, which included about 1000 titles or almost three million pages of historic periodicals. The project included scanning, article-level segmentation and OCRing of newspapers dated from 1760-ies to 2010-s. The resulting digital collection is available online: www.periodika.lv.



The NLL contributes **114** newspaper titles to the project, comprising **460.781** pages in total.

4.7 NLP - *Biblioteka Narodowa*

The NLP (Biblioteka Narodowa) is the national library of Poland and was established under the Decree of the President of the Republic of Poland of February 24, 1928. The National Digital Library [cBN Polona](http://cBN.Polona) provides access to 712 issues of journals from the holdings of the National Library. These are periodicals and newspapers published in the 19th and 20th centuries, as well as the underground press from the communist period.

The National Library uses originals as well as microfilm copies to digitise newspapers and journals. A plan is in place to digitise all newspapers and periodicals from the holdings of National Library. Digital copies are made primarily for objects exposed to destruction, those belonging to the public domain and those most frequently ordered by library users. Recently digitised and published were newspapers and journals from the period between 1914 and 1939 for the Europeana Newspapers Project.

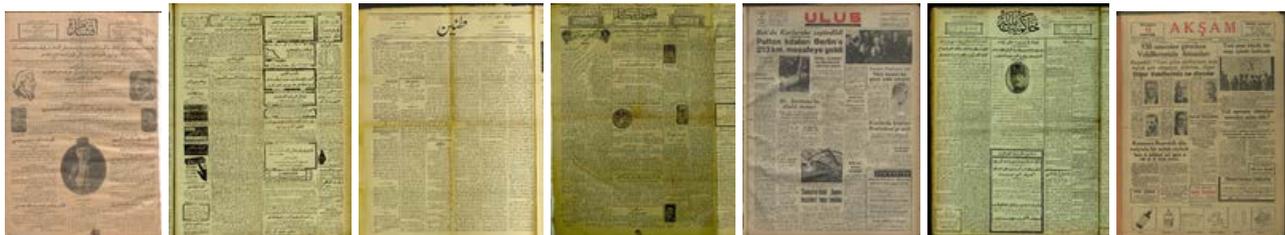


The NLP contributes **116** newspaper titles to the project, comprising **83.648** pages in total.

4.8 NLT - Milli Kutuphane Baskanligi

The National Library of Turkey was established with the intention that it would become a center for a national network of knowledge and information and it would thus contribute to cultural development, economic growth, and creation of a knowledge society in our country.

The National Library at this time maintains a collection of 3,052,268 items, comprising books, periodicals, and non-book materials. The collection of periodicals consists of 1,462,243 issues/volumes of daily papers, magazines, bulletins, almanacs, and similar materials.



The NLT contributes **1** newspaper title to the project, comprising **8.990** pages in total.

Note: Due to the unsuitability of current OCR technologies for Ottoman alphabet, the NLT is currently evaluating a new selection of content, with the aim of selecting several newspapers in Latin alphabet that can then be successfully processed with the refinement technologies in Europeana Newspapers.

4.9 ONB - Österreichische Nationalbibliothek

The ONB (Österreichische Nationalbibliothek) is the national library of Austria. The guiding selection criteria of the ONB for digitisation are:

- newspapers which have bad paper or cover conditions
- newspapers which are not saved via microfilm yet or
- newspapers which are already microfilmed but the film has a bad condition

Moreover, the ONB digitises newspapers which are:

- frequently requested or used by our customers
- used in different projects running in the ONB
- asked by certain customer requests or
- requested by each department of the ONB

Currently the ONB has about 6.9 Million digitised newspaper pages in the [ANNO](#) (AustriaN Newspapers Online) portal. The digitisation of newspapers at the ONB is not running on a project basis. It has a more initiative character which means that the entire digitisation process has an “open end” disposition and there is no end in sight yet.



The ONB contributes **237** newspaper titles to the project, comprising **6.781.332** pages in total.

4.10 SBB - Staatsbibliothek zu Berlin

The SBB (Staatsbibliothek zu Berlin) is the largest research library in Germany and maintains the “[Zeitschriftendatenbank](#)”, the national online portal for newspapers and periodicals. The spectrum collection of newspaper collection of the Berlin State Library includes daily newspapers, weekly newspapers and newspaper-like publications and shows the largest and most comprehensive collection of German newspaper libraries. In addition to the original newspapers (approximately 180,000 volumes newspapers) and microforms (more than 150,000), a comprehensive, constantly growing stock as online databases and CD-ROM editions are offered. The newspapers from Berlin and Brandenburg are almost complete in paper or microform copies and the historical inventory includes the major German newspapers, in particular newspapers in the country Prussia.

The Berlin State Library has one of the most extensive collections of newspapers from Berlin. Therefore some titles of the selection for the project are mid-to late-19th Century important titles like the "Berliner Boersen-Zeitung" (1864 - 1940) and the "Berliner Tageblatt" (1872 - 1939).



The SBB contributes **8** newspaper titles to the project, comprising **248.200** pages in total.

Note: Due to pending digitisation projects, the number of pages from the SBB in the current data set is not yet fully determined. An additional 1.5 million newspaper pages will still be selected further on in the project.

4.11 SUBHH - Staats- und Universitätsbibliothek Hamburg

Hamburg State and University Library Carl von Ossietzky is the largest general academic library in the city state of Hamburg and, at the same time, the central library of the University and other Hamburg colleges.

The newspaper collection consists of about 18.000 volumes and 23.000 microfilm roles. The print volumes contain a complete legal deposit collection of newspapers printed in Hamburg from 1943 up to now and fragments of older Hamburg newspaper series. Almost the whole newspaper collection had got lost in July 1943. Additionally, there are diverse series of regional, national and international newspapers.

The seven newspapers chosen for Europeana Newspapers are the first set of digitised newspapers of the library. The library holds complete microfilm runs of these titles which are frequently requested. They were selected because they represent different types of the daily press: two famous national papers (*Hamburgischer Correspondent*, 1721-1934; *Hamburger Nachrichten*, 1792-1939), a regional paper from the mid-sized former Danish city of Altona (*Altonaer Nachrichten*, 1850-1941), a liberal 'Generalanzeiger' (*Hamburger Anzeiger*, 1888-1945) with a strictly political 'sister'-paper from the same publishing house (*Neue Hamburger Zeitung*, 1896-1922), a local paper (*Norddeutsche Nachrichten*, 1879-1943) and finally the foremost economic paper from Hamburg (*Börsenhalle*, 1805-1904, incorporated in the *Correspondent* 1905-1934).



The SUBHH contributes 16 newspaper titles to the project, comprising 2.216.200 pages in total.

4.12 UB - Univerzitet u Beogradu

The UB (Univerzitet u Beogradu) is the central library within the University of Belgrade. The Library collection consists of 1.5 million objects and tens of thousands of electronic books and papers are also accessible in the library.

For the Europeana Newspaper project University of Belgrade, University library is providing a selected range of digitised newspapers from 19th and 20th century in Serbia. All items are in Cyrillic script and Serbian language and have a great importance for researchers in humanities and social sciences looking into historical period of 19th and 20th century in Serbia. The materials consist of more than 400.000 pages. Parts of the materials are available online with the rest following shortly.



The UB contributes **43** newspaper titles to the project, comprising **408.181** pages in total.

5. 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 set is still growing due to some difficulties with the selection and content provision process at some of the partners, but this will not affect the scheduling of the refinement activities negatively.

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.