IFLA International Newspaper Conference


11-13 April 2012
BnF, Paris

With the support of:
The European(a) Newspapers Project

A Gateway to European Newspapers Online


Thorsten Siegmann, Staatsbibliothek zu Berlin, Germany
Content

Project Profile
• Aims
• Consortium
• Framework

Areas of activity
Europeana Newspapers – why newspapers?

"Die Zeitungen sind die Sekundenzeiger der Geschichte."
(Newspapers are the sweep hands of history)  
Arthur Schopenhauer

Why newspapers?
• Relevant to all citizens
• Highly relevant to European policies incl. Europeana
• Newspapers in libraries – between
  • Heaven = solid and complete originals, excellent microfilm copies
  • and Hell = frail and crumbly originals, missing editions, incomplete supplements, poor microfilm copies, legal uncertainties with contemporary material
Europeana Newspapers: Aims and Objectives

Europeana Newspapers

- aims at the aggregation and refinement of newspapers for *The European Library* and *Europeana*.

- will use refinement methods for OCR, OLR (article segmentation), and named entity (NER) and class recognition

- the libraries participating in the project will provide around 18 million digitised newspaper pages to Europeana

- Further libraries will be encouraged to contribute newspapers to Europeana and TEL by the project
Project Profile: Consortium & stakeholders

• 17 partners from 12 countries within the consortium
  • National libraries
  • University libraries
  • SME

• External partners and stakeholders:
  • Involvement of libraries outside the project consortium

• Framework:
  • Funded as a Best Practise Network in the ICTPSP programme of the European Commission
  • Project Duration: February 2012 – January 2015
1. Staatsbibliothek zu Berlin (project co-ordinator)
2. National Library of the Netherlands
4. Österreichische Nationalbibliothek
5. National Library of Finland
6. Staats- und Universitätsbibliothek Hamburg
7. Bibliothèque nationale de France
8. National Library of Poland
9. University of Salford
10. CCS Content Conversion Specialists GmbH
11. Stichting LIBER
12. National Library of Latvia
13. National Library of Turkey
14. University Library of Belgrade
15. University of Innsbruck
16. Landesbibliothek Dr. Friedrich Tessmann
17. The British Library
Project Profile: Objectives

1) Selection, Refinement & Aggregation of content
   • Make Europeana the largest provider of pan-European newspaper collections
   • Provision of more than **18 million newspaper pages** to Europeana, many of those with full-texts
   • Support move from images to texts in Europeana

2) Analysis of existing newspaper collections
   • Survey of newspaper holdings in Europe

3) Quality Assurance & Best practise recommendations
   • Contribute to optimised workflows and data aggregation infrastructures
   • Provide best practice recommendations for digitization, refinement, workflows, metadata etc. and evaluation tools

4) Presentation and full-text search
   • Improve access to newspaper collections within Europeana
1) Selection, Refinement & Aggregation of content

- Aggregation of 18 million pages of digitised newspapers to *Europeana* and to *The European Library*
  - 8 million pages “as is” (content providers)
  - 10 million refined pages: OCR (UIBK, Austria)
  - 2 million refined pages: OCR/OLR (article segmentation) (CCS, Germany)
- Analysis of available digital newspaper collections and selection of subsets suitable for refinement

www.europeana.eu/

www.theeuropeanlibrary.org/
1) Refinement – OCR and OLR

- 10 million refined pages: OCR (UIBK, Austria)
- 2 million refined pages: OCR/OLR (article segmentation) (CCS, Germany)

- UIBK enriches the OCR with structural information from their Document Understanding Platform
- CCS produces OCR and verification of column recognition, zoning, article segmentation, and page class recognition
- CCS provides libraries with a client technology for manual correction of recognition and segmentation results

CCS: Column recognition, article segmentation

UIBK: Detection of headings, footnotes, etc.
Table of contents extraction
1) Refinement - Named Entity Recognition

- KB provides named entities recognition (NER) for material from up to three languages (Dutch, English, and German)
2) Analysis of existing digitised newspaper collections

• Project partners and others will be contacted until summer 2012 to analyse the extent of digitised newspapers collections at their institutions
  • Results will be embedded in “Zeitschriftendatenbank” of Staatsbibliothek zu Berlin (Union Catalogue of Serials)
  • Potential new partners for the extension of the network will be suggested by survey
• May also be useful to judge technical status of digitised data and as part of gathering descriptive metadata

• If you hold digital newspaper collection and like to participate in the survey → please contact: survey@europeana-newspapers.eu/
3) Analysis of work & Best Practise Recommendations

• Analysis of metadata formats in use by libraries in digitisation projects
• Align metadata models with the METS/ALTO standard and release best practise recommendation on how to apply these formats in newspaper digitisation and refinement
• Usability of the recommendation will be tested through an evaluation cycle
• Provide recommendations on best practices for refinement of digitized newspaper collections for Europeana
4) Presentation & Access to full-texts

- Within the lifetime of the project, a content browser will be built within TEL portal so that users can...
- Search full text, e.g.
  - by search term,
  - by named entities
  - by collections of newspapers
  - by date ....
- See newspaper images
- Be linked to relevant library sources
- This browser will be built in TEL during project; and exported to Europeana after the project
5) Dissemination

- Objectives:
  - Establishment of publicity
  - Increasing usage of Europeana
  - Awareness raising among target groups

- Tasks:
  1. Media Communication
  2. Workshops and conferences
     - Three main dissemination workshops
     - National information days
     - Network extension
  3. Exploitation
... and more will come soon

• Detailed information will be available soon:
  
  → http://www.europeana-newspapers.eu/
  (Launch: End of April)

• Participating in the survey on European Newspapers:
  
  → survey@europeana-newspapers.eu
Thank you for your attention!

Thorsten Siegmann, Staatsbibliothek zu Berlin

siegmann@eu1914-1918.eu

www.europeana-newspapers.eu/
(Launch: End of April)