Making Historic Newspapers Available Online: Why, Where and How

IFLA Newspaper Pre-Conference
14 August 2014, Geneva

Hans-Jörg Lieder, Staatsbibliothek zu Berlin – Preußischer Kulturbesitz | Berlin State Library
Why Newspapers?

Cons:

• Originals are cumbersome objects
  • Prone to damage and destruction due to paper quality
  • Missing issues and pages
  • Difficult to deal with from a catalogueing point of view
  • Poor bindings
  • Funny fonts and fading ink
• Microforms may also be cumbersome objects
  • Skewed images, text loss
  • More missing issues and pages, plus duplicate pages
That’s Why!

Pros:

• “Newspapers are the second hand of history”
  • Provide insights into history’s microstructure
  • Unlimited thematic scope
  • Interesting for all fields of scholarship, but also for the layman
  • Massive digital newspaper text corpora allow for new ways of research
  • A European perspective: significant contribution to the shaping of identities of peoples and individuals
The Europeana Newspaper Project – Who?

Blue – Content Providers

Yellow – Service Providers

Green – Associated Partners
The Europeana Newspaper Project – What?

<table>
<thead>
<tr>
<th>Country</th>
<th>Partner</th>
<th>Titel</th>
<th>Startdate (overall)</th>
<th>Enddate (overall)</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>BnF</td>
<td>80</td>
<td>1814</td>
<td>1944</td>
</tr>
<tr>
<td>Germany</td>
<td>SBB</td>
<td>6</td>
<td>1872</td>
<td>1940</td>
</tr>
<tr>
<td></td>
<td>SUB-HH</td>
<td>16</td>
<td>1721</td>
<td>1945</td>
</tr>
<tr>
<td>Netherlands</td>
<td>KB</td>
<td>203</td>
<td>1618</td>
<td>1900</td>
</tr>
<tr>
<td>Italy</td>
<td>LFT</td>
<td>15</td>
<td>1813</td>
<td>1949</td>
</tr>
<tr>
<td>Estonia</td>
<td>NLE</td>
<td>43</td>
<td>1852</td>
<td>1944</td>
</tr>
<tr>
<td>Finland</td>
<td>NLF</td>
<td>11</td>
<td>1900</td>
<td>1910</td>
</tr>
<tr>
<td>Latvia</td>
<td>NLL</td>
<td>117</td>
<td>1868</td>
<td>1955</td>
</tr>
<tr>
<td>Poland</td>
<td>NLP</td>
<td>118</td>
<td>1914</td>
<td>1939</td>
</tr>
<tr>
<td>Turkey</td>
<td>NLT</td>
<td>22</td>
<td>1818</td>
<td>1928</td>
</tr>
<tr>
<td>Austria</td>
<td>ONB</td>
<td>275</td>
<td>1686</td>
<td>1945</td>
</tr>
<tr>
<td>Serbia</td>
<td>UB</td>
<td>45</td>
<td>1830</td>
<td>1944</td>
</tr>
</tbody>
</table>

20 languages

ca. 950 titles

ca. 10m pages refined

- 8m OCR
- 2m OLR
- 2m NER
The Europeana Newspaper Project – What else?

- Tools for informed selection of newspapers for digitisation
- Specifications and tools for the creation and validation of OCR-ready images
- Large-scale, highly automated workflows for refinement (OCR, OLR, NER)
- Metadata best practice recommendations
- Transmission of data to European Portals and the Union Catalogue of Serials
- Presentation of results
What does it look like … in TEL?
What does it look like … in the Union Catalogue of Serials?
What about Services?

• Richest service portfolio available at local web pages (if you’re lucky)
  • Calendar navigation, search in texts
  • filters to narrow down queries or result sets
  • mark-ups, annotations, links to other information resources, etc.

• Services at TEL
  • Calendar navigation, search in texts
  • Filters for searches: title, date, owning library
  • Filters for results: title, date, owning library, country, language
What about further Services? – An Example

Empfindsamkeit
(ca. 1720-1800)

= Sentimentalism
What about further Services?

- Natural language processing
- Text mining
- Visualisations
- Cross-media linking
- Semantic field analysis
- Links to other resources, librarian and non-librarian
- ...
- LIBERATE YOUR DATA AND LEARN FROM YOUR USERS!
Digital Text Corpora: The Inconvenient Truth

Bag of Words OCR Evaluation
Per Language

Language Setting

Success Rate

Dutch | 82.4%
Czech | 85.3%
English | 50.9%
Estonian | 75.6%
Finnish | 67.5%
French | 83.4%
German | 84.1%
German/DiGen | 93.1%
Hungarian | 87.0%
Latvian | 87.0%
Polish | 57.6%
Russian | 68.3%
Serbian/Cyrillic | 76.1%
Swedish | 82.6%
Ukrainian | 54.1%
Yiddish | 32.7%
What About Digital Text Corpora?

• Provide possibilities for corrections where data is presented
• Options for improvement
  • Automated corrections (index and page level)
  • Software aided corrections
  • Crowdsourcing
• Challenges: data synchronisation, update intervals, versioning…
Thank you for your attention!

IFLA Newspaper Pre-Conference
14 August 2014, Geneva

Hans-Jörg Lieder, Staatsbibliothek zu Berlin – Preußischer Kulturbesitz | Berlin State Library