



européana
newspapers



Europeana Newspapers Interim Report

Prague, 18th October 2013

Günter Mühlberger, Innsbruck University

- ENP Newspaper Viewing Application
- Best Practise Approach
 - Metadata format
 - Structural metadata
 - Workflow
 - Evaluation
- Outlook

ENP Newspaper Viewer

The screenshot shows the ENP Newspaper Viewer interface. At the top, there is a navigation bar with the logo 'The European Library' and the tagline 'Connecting knowledge'. Below the logo is a search bar with the text 'Search...' and a 'GO' button. The main content area is titled 'Search Newspapers' and features a search bar with the text 'Search within newspapers...' and a 'GO' button. Below the search bar are filters for 'Search Options' (Text, Title), 'Library' (All), and 'Language' (All). A timeline for 'Year of Publication' is shown, ranging from 1800 to 1950. The 'Discover Newspapers' section displays a carousel of newspaper thumbnails and a 'Top 5' list of newspapers: Der Humorist, Innsbrucker Nachrichten, Das Vaterland, and Militär-Zeitung.

Summary

- e2014 more than 10 mill. of digitized pages shall be available via this newspaper browser
 - 2 Mill. pages will come with structural markup on article level
- Site shall be integrated into the current TEL application
- Also newspapers with restricted access (copyright) shall be linked in the application (e.g. Turkish newspapers)
- Sustainability
 - TEL (membership model) will sustain it
 - Further integration of newspapers has to be clarified

Pros & Cons

- Pros

- Ambition was “one place for all digitized European newspapers”
- Newspapers are distributed over many countries
- Browsing e.g. via calendar would allow to compare newspapers from one day or several countries
- Centralized searching

- Cons

- Setting up a central server for all newspapers requires more IT capability than originally planned
- Distributed solutions require a lot of work (adapting the viewer to several local solutions)
- Dependency on local solutions
- Copyright restrictions lead to the situation that the user will get “more” at local sites than in the central place

ENP as best practise project

- ENMAP (Europeana Newspaper METS ALTO Profile)
 - Has been developed in order to be able to manage delivery/distribution of 10 mill. newspaper pages from more than 10 countries, etc.
 - Issue level
 - MODS for descriptive metadata
- ENMAP structural map
 - Higher ambition
 - Idea is to provide a data dictionary to support common understanding of newspapers

Example

ONB/ANNO Austria Newspaper Online - Mozilla Firefox

anno.onb.ac.at/cgi-content/anno?aid=bn&datum=18700604&zoom=16

ANNO Zeitungen
→ Jahresübersicht
→ 1870
→ 4. Juni 1870

Innsbrucker Nachrichten
→ Jahresübersicht
→ 1870

4. Juni 1870

Innsbrucker Nachrichten, 4. Juni 1870

Österreichische Nationalbibliothek
ANNO
Historische österreichische Zeitungen und Zeitschriften

Start | Firefox | Notizen Pra... | 2 Adobe A... | 2 Windows... | Microsoft Po... | consortium... | 09:26

Current proposal

- Newspaper content items (~ articles)
 - Are defined as containing the “content” of a newspaper
 - Since newspapers consist of a lot of different content “article” is a too narrow notion
 - Are clearly separated from each other both by content and by layout features (e.g. separators)
 - May be part of a section
- Structural elements
 - Content items consist of structural elements
 - Structural elements are e.g. headlines, sub-headlines, by-lines, paragraphs, author names, etc.

Classification of NCI

- Five main classes
 - Information
 - Opinion
 - Advertisement
 - Entertainment
 - Metadata
- These classes contain “typical content items”
 - Information: news, breaking news, background news, etc.
 - Opinion: Columns (by well-known persons), commentaries, letters-to-the-editor, book reviews, etc.
 - Advertisement (including also classified advertisement)
 - Entertainment: Serial novels, poems, jokes,...
 - Metadata: Title section, imprint

Rationale for data dictionary

- Rationale for data dictionary
 - Desiderata of metadata schemas
 - Communication among libraries and with technical providers
 - Crowd sourcing
 - Automated processing (Natural Language Processing, Data Analytics, Information Extraction, etc.)
- Technical realisation
 - Paper describing main Newspaper Content Items and Structural Elements
 - METS Structural Map

Best practise project: Workflow

- Several simple tools to support workflow for newspapers
 - File Analyzer Tool (FAT)
 - Binarization and Conversion Tool (BCT)
 - DateApplierTool (DAT)
 - Structify (Structural Metadata Tool)
- Named Entities Recognition (NER)
 - German, Dutch, French language
- OCR workbench in Innsbruck
 - 32 parallel FineReader SDK processes
 - Between 10.000 and 100.000 newspaper pages per day
 - Sustained after the end of the project

Finally: Evaluation of OCR results

- What can be done with erroneous OCR results?
 - Evaluation of 50 pages per library
 - Manual ground truth
 - Sophisticated evaluation methods by University of Salford
- Other approaches
 - Evaluating Models of Latent Document Semantics in the Presence of OCR Errors (WALKER, 2010)
 - Topic modeling suffers from OCR errors but still useful results are possible even with high Word Error Rates
 - Mapping Texts: Combining text-mining and geo-visualization to unlock the research potential of historical newspapers (TORGET, et.al, 2010)
 - Simple evaluation method: “good” words (=those found in a dictionary) vs. “bad” words
 - Need for transparency

MAPPING TEXTS Stanford University
University of North Texas

Assessing Digitization Quality
Scans of Texas Newspapers, 1829-2008

This visualization plots the quantity and quality of 232,567 pages of historical Texas newspapers, as they spread out over time and space. The graphs plot the overall quantity of information available by year and the quality of the corpus (by comparing the number of words we can recognize to the total number scanned). The map shows the geography of the collection, grouping all newspapers by their publication city, and can show both the quantity and quality of the newspapers from various locations. Clicking on a particular city will provide a detailed view of the individual newspapers, where you can examine both the quantity and quality of information. A timeline of historical events related to Texas is also available for context.

Time Quantity of Recognized and Unrecognized Text, 1829-2008

Zoom: 1d 3d 1m 3m 1y Max

Total Words Scanned 563.85 k • Correct Words Scanned 454.13 k | Februar 01, 2008

Space Collection Quantity and Quality by Location

Abilene, Texas, 1829 - 2008
Good words: 72,186,505 Total words: 93,508,784
77%

- The HSU Brand 100%
- The McMurry Bulletin 80%
- The Optimist 80%
- The Reata 80%
- The War Whoop 80%

Zoom level: To data, All years, Manual

Legend: Ratio of Good to Bad Words

Showing cities having publications with 50.0% - 100% correct words.

Circle Scaling: Log, Linear

Circle size is relative to total

- Project runs until January 2015
 - Still 3 mill. pages to process
 - Additional OCR processing in the last six months of the project
 - Release of the ENMAP format and structural metadata dictionary
 - Attempts with automatically detecting structural metadata
 - Release of the TEL Newspaper Site
- A lot to do 😊



europaana
newspapers



Thank you for your attention!

Günter Mühlberger |
<guenter.muehlberger@uibk.ac.at>