Europeana Newspapers Interim Report

Prague, 18th October 2013

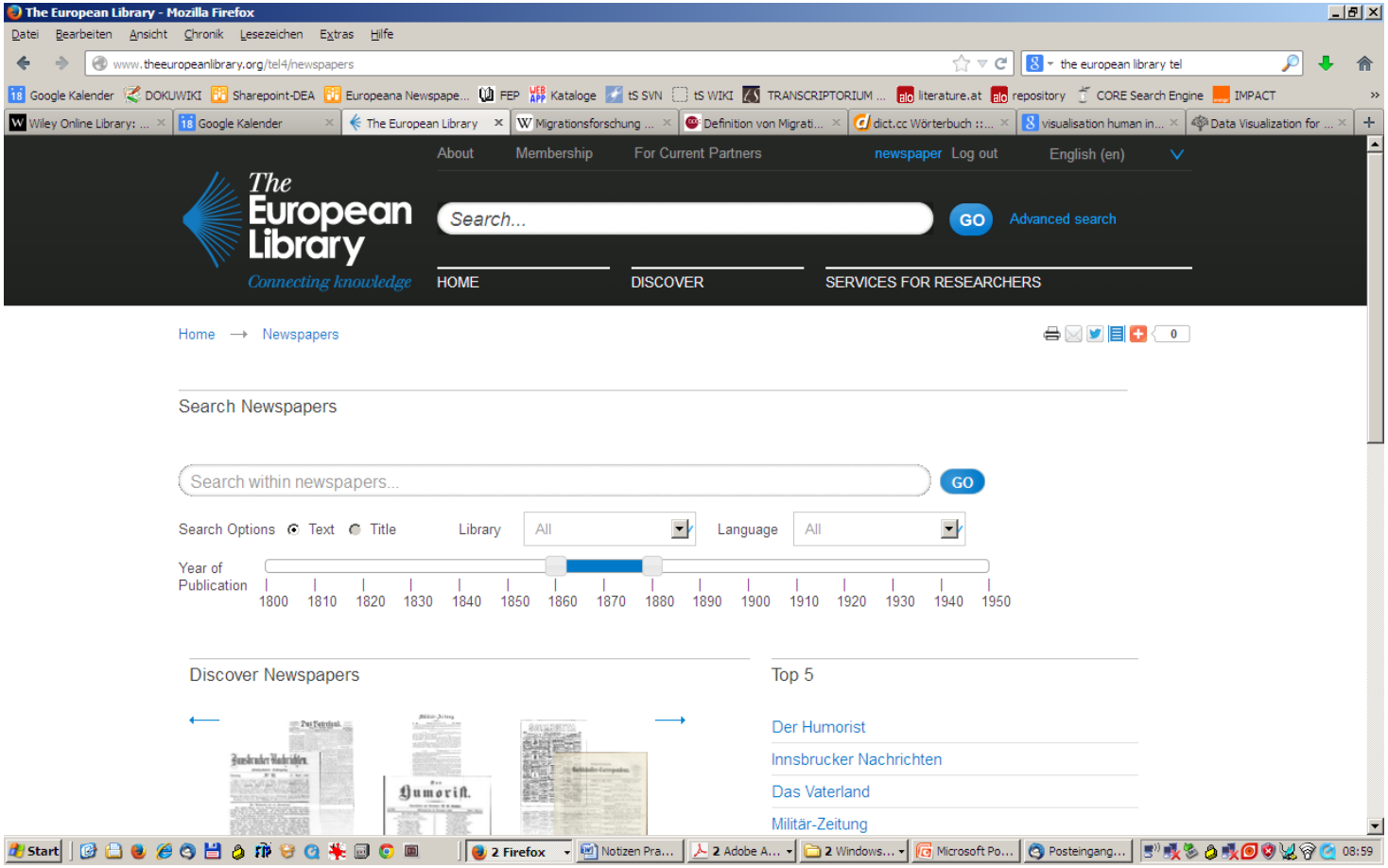
Günter Mühlberger, Innsbruck University

Agenda

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

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ENP Newspaper Viewer



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

ENP as best practise project

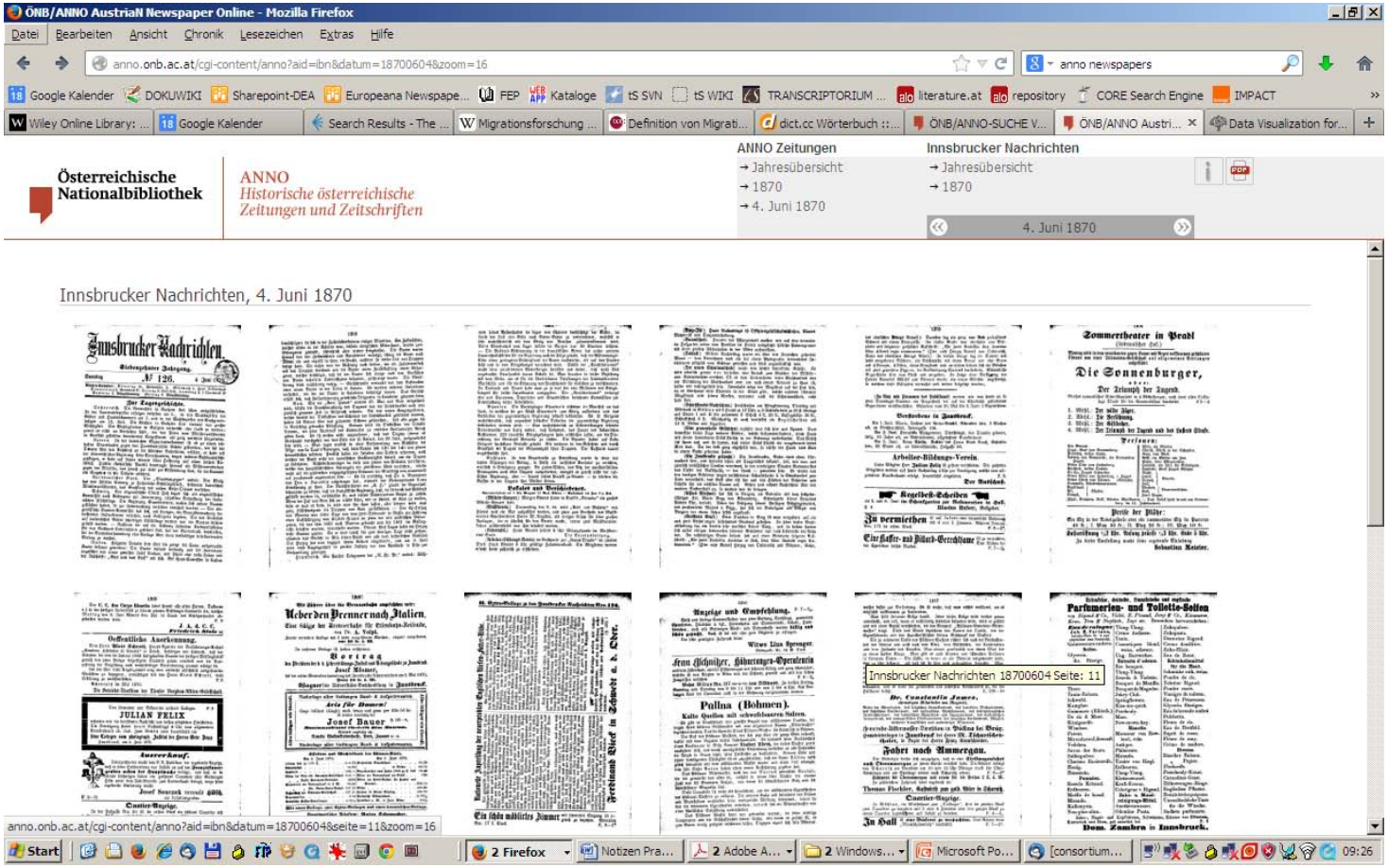
- ENMAP (Europeana Newspaper METS ALTO Profile)
  - Has been developed in order to be able to manage delivery/distribution of 10 mill. newspaper pages from more than 10 countries, etc.
  - Issue level
  - MODS for descriptive metadata
- ENMAP structural map
  - Higher ambition
  - Idea is to provide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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Example



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

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.

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

Finally: Evaluation of OCR results

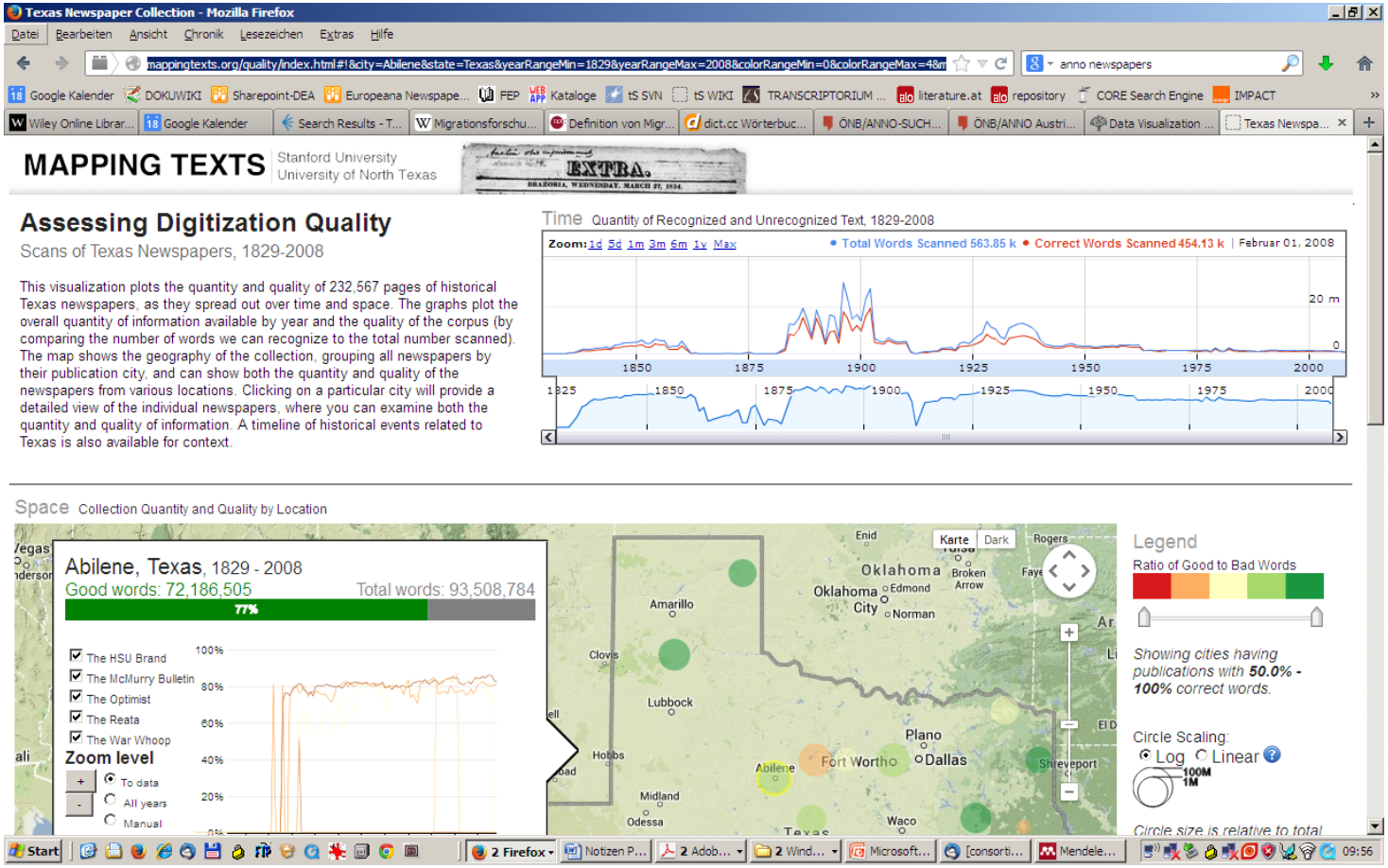
- What can be done with erroneous OCR results?
  - Evaluation of 50 pages per library
  - Manual ground truth
  - Sophisticated evaluation methods by University of Salford
- Other approaches
  - Evaluating Models of Latent Document Semantics in the Presence of OCR Errors (WALKER, 2010)
    - Topic modeling 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

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

mappingtexts.org



This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

Outlook

- 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 ☺

europeana newspapers

This project is partially funded under the ICT Policy Support Programme (ICT PSP) as part of the Competitiveness and Innovation Framework Programme by the European Community
http://ec.europa.eu/ict_psp

ICT PSP ICT POLICY SUPPORT PROGRAMME

europeana
newspapers

ICT PSP
ICT POLICY SUPPORT PROGRAMME

Thank you for your attention!

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