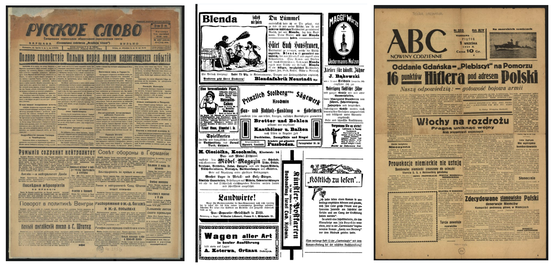
Making Europe's Historical Newspapers Searchable

Clemens Neudecker, Staatsbibliothek zu Berlin - Preußischer Kulturbesitz, and Apostolos Antonacopoulos, PRImA Research Lab, University of Salford, Greater Manchester

Facts & Figures: Over the course of 2012—2015, the Europeana Newspapers project brought together 18 partners from all across Europe to tackle newspaper digitisation on a massive scale.

The main project goals were to:

- Aggregate metadata about Europe's digitised historical newspapers and ingest it into the European digital cultural heritage platform Europeana
- Convert 10 million pages to fully searchable text by means of Optical Character Recognition
- Create a special content viewer to improve online newspaper browsing
- Establish best-practices for the scalable digitisation of historical newspapers



Example newspaper pages from the collection

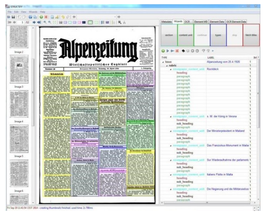
Portal: At the end of the project, the online newspaper browser at The European Library provides access to millions of pages of historical newspapers.

It is possible to search the full-text derived from OCR, browse the more than 1,000 newspapers by title, find specific issues by date of publication in a calendar or explore newspapers per country on a map of Europe.

You can access the newspaper portal at this URL: www.theeuropeanlibrary.org/tel4/newspapers

Software: The project partners have developed a number of free and open source software:

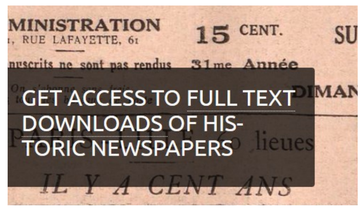
- Tools for working with the PAGE format have been made available by PRImA Research Lab: https://github.com/PRImA-Research-Lab
- Tools for Named Entity Recognition and Disambiguation and a corpus have been released: https://github.com/EuropeanaNewspapers
- The University of Innsbruck developed a number of tools for various data preparation and pre-processing tasks: https://github.com/dea-uibk
- Structify, an advanced tool for viewing and editing structural metadata in METS/ALTO format



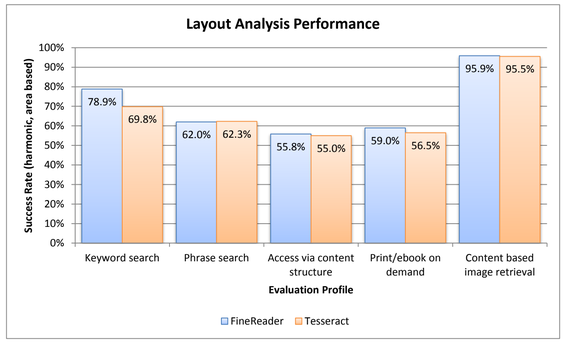
Structify: a tool for structural annotation of digitised newspapers

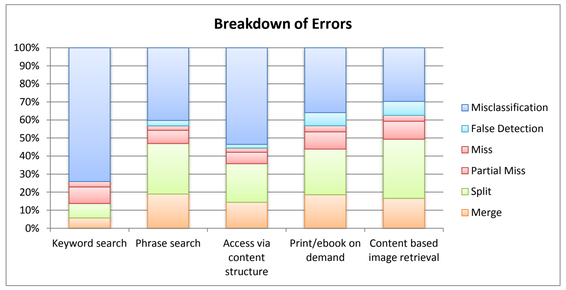
Datasets: The Europeana Newspapers project partners are committed to providing unrestricted access to the data produced by the project. Three datasets can be freely accessed:

1. A Ground Truth dataset can be obtained via http://primaresearch.org/datasets/ENP
2. A dataset for training and evaluation of Named Entity Recognition of historical newspapers in Dutch, French and German has been released https://github.com/EuropeanaNewspapers/ner-corpora
3. The full-text of newspapers in the public domain is made available for download at http://research.europeana.eu/itemtype/newspapers



More than 5 million pages can already be downloaded as ZIP archives.

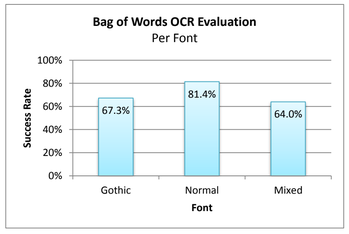




Evaluation: To illustrate the usefulness of the Ground Truth dataset, state-of-the-art page analysis systems have been evaluated against it to ascertain an objective measure of their performance on historical newspapers. ABBYY FineReader Engine 11 and Tesseract 3.03 have been evaluated to establish a baseline with regard to both layout analysis and text recognition performance.

The charts above show the layout analysis performance per real-world use scenario (as identified by libraries and users). The breakdown of errors shows the different (weighted) error types that are measured by the performance analysis system.

The chart on the right indicates the performance of FineReader in recognising text in three different font situations as used in the production workflow.



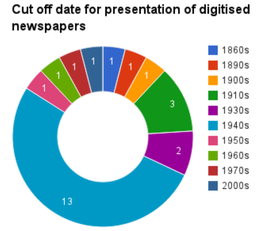
Scale: The newspaper collection covers a timeframe from 1618 - 1990, 4 alphabets and 40 languages.

Approximately 12 million pages were processed:

- 10 million pages OCR using ABBYY FineReader Engine 10 & 11
- 2 million pages OCR including article segmentation using CCS docWorks technology
- The total amount of words recognised is estimated to be more than 70,000,000,000 (70 billion)
- As the scanned images were coming from 12 libraries, digitised in distinct projects and configurations over many years, there was major variation in the data in terms of quality, file formats, metadata standards and availability

Challenges & Lessons learned: The key findings of the project can be summarised as follows:

- OCR at such scale requires a highly specified workflow & pre-processing incl. validation checks
- To successfully process all images (>400 TB) in the project lifetime, all images were binarized to reduce file size - resulting in 90 % reduction of data volume vs. 1 % lower word accuracy
- Recognition of pages with mixed fonts (e.g. Antiqua/Gothic) or with multiple languages significantly increased processing time in FineReader
- Accurate layout analysis is challenging - quality could benefit greatly from advances in this area
- Recognition of large tables (e.g. stock markets) sometimes resulted in crashes of the OCR engine
- OCR for Ottoman font (resembling Arabic) was not available and results obtained from Arabic OCR were poor (at about 20% recognition rate)
- Due to copyright issues, most libraries only digitise newspapers printed before 1940 and make them available for online publication



Staatsbibliothek zu Berlin
Preußischer Kulturbesitz

Staatsbibliothek zu Berlin - Preußischer Kulturbesitz
Berlin, Germany
www.staatsbibliothek-berlin.de

University of Salford MANCHESTER PRImA Research Lab

Pattern Recognition and Image Analysis Research Lab, School of Computing, Science and Engineering, University of Salford, Greater Manchester, United Kingdom, **www.primaresearch.org**

www.europeana-newspapers.eu

europeana newspapers

This work has been supported in part through the EU CIP ICT-PSP Programme grant Europeana Newspapers (Ref: 297380).