Retrodigitization and Electronic Representation of the “Vorwärts”

A Workshop-Report

Olaf Guercke
Table of Content

1. Why do we digitize the „Vorwärts“
2. What has been done so far
   • The scan-process: digitizing in-house
   • The internal structure of the newspaper corpus
   • The previous workflow
3. What should be done from now on
   • Objectives of the web presentation
   • Research project: What has been implemented so far
   • Research project: Possibilities and challenges for the future
Why do we digitize the „Vorwärts“

- By far the most important source for social democracy in German Kaiserreich and Weimarer Republik
- Content important for both historical and cultural research
- Currently only accessible via microfilm edition
  - Huge workload for researchers
  - In many cases practically not accessible
- Fits in the narrowest mission of the Library of the Friedrich-Ebert-Stiftung
Why do we digitize the „Vorwärts“

Aim:
a) Complete digitization of the „Vorwärts“, 1876-1878 ; 1891-1933
b) Providing it to the public via researchable web-presentation

Extent:
200,000 DIN A3 newspaper-pages
19,000 issues
44 volumes
The „Vorwärts“ and 9 different supplements
Font: German gothic
Warnung!
Wer mit dem Faschismus spielt, der spielt mit Deutschlands Untergang!

Bomden im Arbeiterviertel
Schweigende Polizei — Die Cöll als Täter erkannt

Extra-Ausgabe.
Vorwärts
Zentralorgan der Sozialdemokratischen Partei Deutschlands.

Warnung!
Wer mit dem Faschismus spielt, der spielt mit Deutschlands Untergang!

Bomden im Arbeiterviertel
Schweigende Polizei — Die Cöll als Täter erkannt

Protest des österreichischen Proletariats!
Arbeiter, Parteigründern!

Wir wollen keinen Krieg! Nieder mit dem Kriege!
Sieg der internationale Börsenverbrüderung!

Olaf Guercke - Retrodigitization and Electronic Presentation of the "Vorwärts"
What has been done so far

Scan process: Challenges of in-house digitization

• Advantages
  – Complete control over production process with sensitive, partly damaged material

• Challenge 1: Suitable devices
  – Extensive market research

• Challenge 2: Manpower
  – 1 Librarian (25 hours) and 2 scan operators (23 hours in total)
  – 55,000 images scanned and postproduced since June 2015
  – Support your scan-operator!!
What has been done so far

Scan process: File formats

• Tiff uncompressed - ca. 55 MB per Image
• Jpeg 100% - ca. 18 MB per Image
• Jpeg 80% - ca. 3 MB per Image
• Pdf (Jpeg 80%) of whole issues

• All formats are generated automatically by the scanner-software
What has been done so far

Internal structure of the newspaper-corpus

- Periodicity: 4 issues per week - daily appearance - 2 issues per day
- Numbered sequentially during each volume
- Additional numbering for morning/evening-issues
- 9 different supplements with independent numbering
- Various special cases
What has been done so far

Mirroring the structure in filenames and folders

• Filename (example): vw-1923-12-25-325_b165-004.tiff
• Translation: „Page 4 of the Vorwärts issue nr. 1923.325, evening-edition nr. 1923.165, released in December 25th“
• Folder-structure: file-format/newspaper-title/volume/release-date/issue

• Filenames and folders are automatically generated by the scan-software
What has been done so far

Workflow:

- Paper originals bound in quarterly folios are examined and documented with excel-spreadsheet (volume, date, number, pages per issue, conservation-status)
- Folios are opened carefully by a bookbinder
- Originals are scanned and stored in archive-boxes
- Quality control and upload of the scans
- Missing or heavily damaged pages are replaced with help of ULB-Bonn
- All these steps are performed parallel
Die Wählerin

Frauenstimme

Es geht vorwärts!

Nummer 42

VOLK UND ZEIT

10. Jahrgang

Berlin, den 16. Oktober 1906

Jugend-Borwärts

Nummer 4

Jugend und Frieden.

Beitrag zum internationalen Sozialismus.

29.03.2016

Olaf Guercke - Retrodigitization and
Electronic Presentation of the "Vorwärts"
What should be done from now on

Objectives of the web-presentation

• 2/3 of actual client inquires: Researchers and students

• Client request:

  For the preparation and design of a multiple-semester seminar at the historian faculty of his university, a docent needs access to reports from the “Vorwärts” about left-wing labor-unrest in Halle in the years of 1919, 1920 and 1921.

• Solution:
  - Keyword-search on basis of OCR, ideally to be narrowed down to certain timeframes
  - Browsing approach via volume, release date and issue
What should be done from now on

Additional important objectives:

- Traceability via URNs at a page-level
- Managable mechanics to block content with copyright-related issues at a page level
What should be done from now on

Research project: Library of the FES – Imageware Components GmbH (IWC)

Aim:
• Presenting the „Vorwärts“ in the web with the presentation-software MyBib eL®
• Developing the software to meet the requirements of such a presentation with the input of expertise from both sides
• Web presentation will be hosted by hbz
What should be done from now on

Research project: What has been implemented so far

• Viewer / display:
  – Pages can be scaled continuously
  – Keynote search / Trunkation, / Search for phrases and multiple words
  – Hit-lists / highlighting of hits

• OCR
  – Fraktur-OCR: ABBYY vs. Tesseract
  – Pretest: ABBYY provides much better results with much faster performance

29.03.2016

Olaf Guercke - Retrodigitization and
Electronic Presentation of the "Vorwärts"
What should be done from now on

Research project: Possibilities and challenges for the future

• Configuration of the „advanced search“
  – Search in certain time-periods
  – Search with all Boolean operators
  – Fuzzy search

• Search entry „browsing“
  – Structurizing the browsing opportunity to three levels: volume, date of appearance (calendar-function), issue

29.03.2016

Olaf Guercke - Retrodigitization and Electronic Presentation of the "Vorwärts"
What should be done from now on

Research project: Possibilities and challenges for the future

• Option to block content at a page level
  – Extremely important requirement to cope with copyright issues
  – Has to be developed on base of the access management functions of MyBib el

29.03.2016
Olaf Guercke - Retrodigitization and Electronic Presentation of the "Vorwärts"
In wunderschönem Mai
Wenn alle Knospen sprießen
Hol ich mein schönstes Kleid herbei
Den Lenz geschmückt zu grüssen.

Und setz ich mich ins grüne Gras
Und kriegt mein Kleidlein Flecken
Und pfleht mich der Osterhase
Ich lache – mache keinen Muckes
Und springe auf mein Zimmer
Drei Löffel Seifenflocken Luxe
Schön hat's den alten Schimmer

511 mal wurde ein Seidenstück
mit Persil gewaschen unter fachamtlicher Aufsicht! Das ge-
waschene Stück wurde auf seine
festigkeit untersucht, und es ergab sich
im Hinblick auf ein so altmägisches Waschen

sein ausserordentlich günstiges Er-
gebnis für die Schonung des Gewebes!

Gibt es einen besseren Beweis für die hochwertige Beschaffen-
extheit eines Waschmittels?

Denken Sie: 511 mal gewaschen!

Das entspricht bei regelmässigem viertägentlichem Waschen
einer Gebrauchsduer von mehreren Jahrzehnten!

Persil schont und pflegt die
Wäsche in ausserordentlichem Masse!
1. Beilage des Vorwärts
Sonntag, 3. August 1930

Wälgeister in Bibliotheken

Fabrik für Reisereichen

Buddha und der Marsfall

Stadtbibliothek

29.03.2016
Olaf Guercke - Retrodigitization and Electronic Presentation of the "Vorwärts"
What should be done from now on

Conclusion:

• We are deep inside the project but there still is a long way to go.
• Now is not the time to make conclusions but to ask questions to be answered in the future:
  
  – How good is the result of the OCR precisely and are there possibilities to optimize it?
  – Is it possible, to make use of the OCR-generated text in other research-projects?
  – Could the function of flexible content blocking be implemented successfully
  – Which desirable user features could be implemented at what expense?
  – What aspects of the now imminent workflow conversion brought real benefits?
  – How and to what extend has the user interest evolved?
Thank you for your kind attention!

29.03.2016

Olaf Guercke - Retrodigitization and Electronic Presentation of the "Vorwärts"