

An automatic analysis of international newspapers, periodicals and press dossiers – A case study from Germany

IFLA Newspapers Section Pre-Conference Geneva, Switzerland, ITU, 13-14 August 2014

Author/presenter: Dr. Jürgen Warmbrunn

Copresenter: Rolf Rasche





The Newspaper Collection as an integral part of the Research Library





Research Library

ca. 500.000 volumes (OPAC), ca. 1.800 journal subscriptions, newspapers until 1944; around 10,5 shelf kilometres

Newspaper Collection

715 titles since 1945 from East Central Europe (236 unique holdings according to ZDB) in ca. 12.800 units on 1 shelf kilometre (bound volumes, bundles, films, CDs), ZDB records, transregional and local dailies and weeklies,

ca. 30 subscriptions

Newspaper Cuttings Archives

ca. 5,3 million cuttings files, microfiches, films; around 1,6 shelf kilometres





Newspaper Cuttings Archives – I

Period, scope: 1920/1952-1999; ca. 5,3 million cuttings in 16.300 folders, 640.000 microfiches, 160 microfilms; 40% Poland / 40% Czechoslovakia / 15% Baltic countries:

- Topical 60%

Biographical33%

Geographical7%





Newspaper Cuttings Archives – II (Special collections)

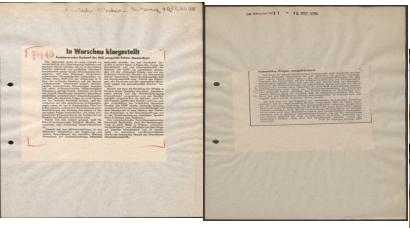
- special collection on the Inter-war Period
- Newspaper cuttings Archives of the German Institute for Contemporary
 History of the GDR (later Institute for Internationale Politics and
 Economics of the GDR) (1952 -1990; USSR, South Eastern Europe)
- Private Archives for Contemporary History and Politics / Wolfram Gabriel (born in Breslau/Wrocław), Rheinstetten-Neuburgweier (800 folders with 30.000 cuttings; 1982-1994)
- radio, press and television reports from Radio Free Europe (1957-1991;
 ca. 1260 archival boxes (600 on Poland, 570 on Czechoslovakia and 90 on South Eastern Europe) on more than 115 shelf metres

Some examples















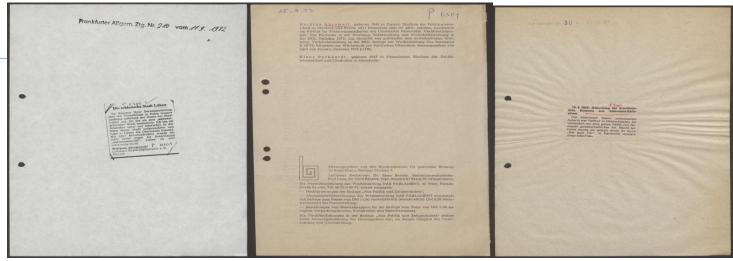


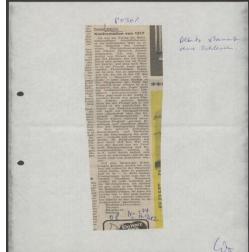








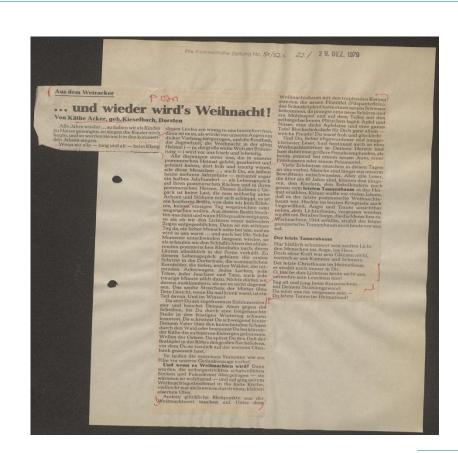












Challenges and opportunities – I



- uniqueness of the material from a historical and regional perspective (transnational for the whole of East Central Europe)
- lack of physical space, conservation problem: paper deterioration
- deliberate "disposal of a disagreeable past"
- support for research through more intensive indexing and digital availability
- increasing online use: worldwide, comfortable, free
- digitization makes new research topics viable
- copyright vs. digitization with open access: is a pragmatic solution possible?
- funding

Challenges and opportunities – II



meshing of research, documentation, presentation and infrastructure/service

 preservation, maintenance and extension of collections, digitization with multimedia applications, improved serviced







Mechanical process

- automatic indexing with minimal manual effort
- heterogeneous sources (paper, language, condition)

Automatic segmentation of the categories of newspaper cuttings

- heading, text of article, author, photograph
- photos, graphics and so on
- signets, hand-written dates

k.o.-criteria

- permissibility of segment related OCR (no fulltext-OCR)
- processing and workflow management
- copyright clearance
- activation or blocking of parts in the presentation system

Project specific requirements – II



Workflow system

- efficiency
- duplicates check
- cooperative work with other institutions
- Management information, project overview, status

Sustainable solution

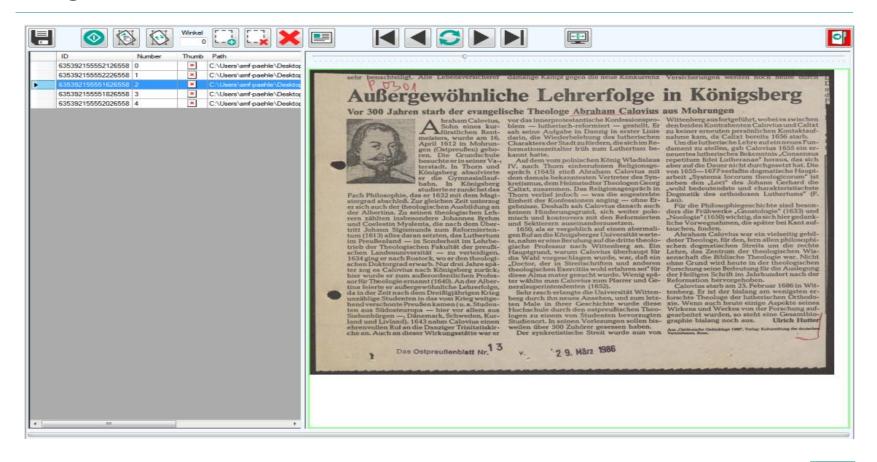
- Support for 10 + years
- Possibility to effectuate technical and legal changes

Legally sound solution

- copyright
- data collection / personal rights

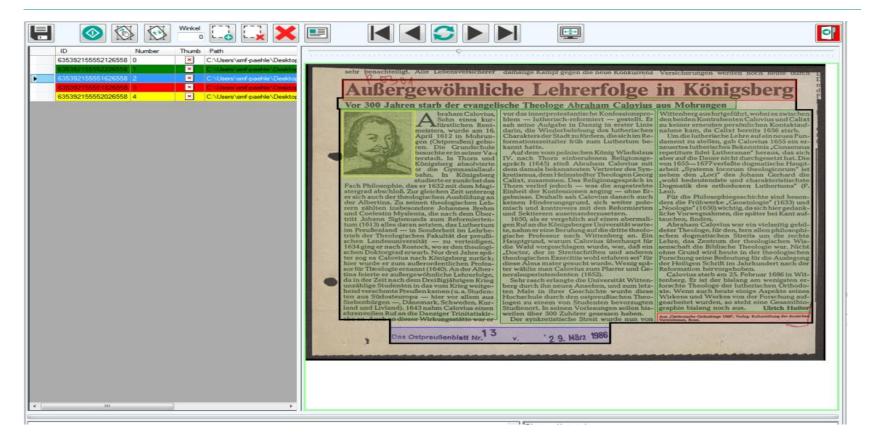
Segmentation - Start





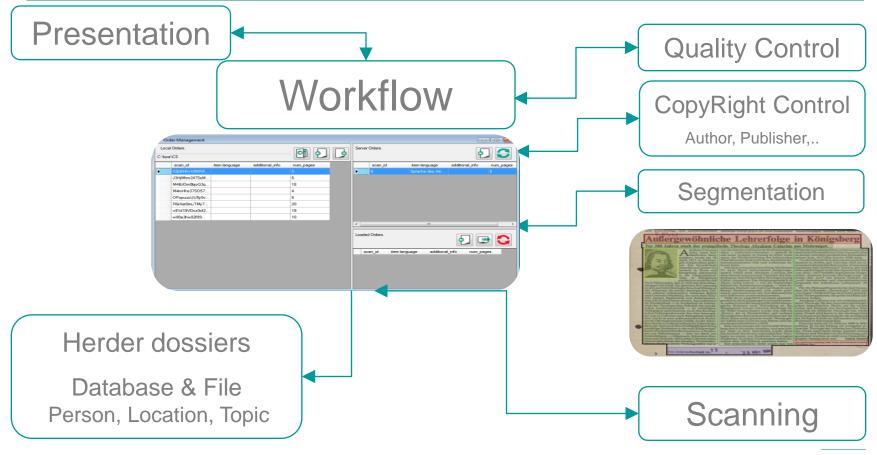
Segmentation – Result





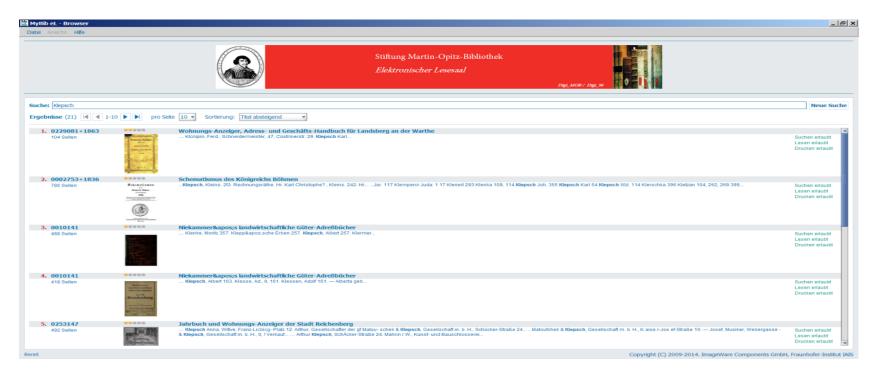
System overview





Hit-list including copyright-control





Presentation including copyright-control





Results and look ahead



Piloting

- test series with 9.000 newspaper cuttings
- automatic segmentation fulfills the subject-specifics and legal requirements
- difficulties with copyright and their effect on the workflow and the presentation system were underestimated

Look ahead

- introduction of the solution in 2014
- objectives
 - 2014: start with about 10.000 newspaper cuttings
 - 2015: standard operation with about 100.000 newspaper cuttings p.a.
 - throughput: 50 newspaper cuttgings scanned & processed per hourStd.
- feedback through a workshop in the summer of 2015 (after completion of our present construction work)

Thank you for your attention







Herder-Institut Marburg e.V., Gisonenweg 5-7 35037 Marburg, Germany ImageWare Components Am Hofgarten 20 53113 Bonn, Germany

juergen.warmbrunn@herder-institut.de

rasche@imageware.de

