Contributors/Presenters:

Dr. Juergen Warmbrunn (Acting Director and Head Librarian),
Herder Institute for Historical Research on East Central Europe, Marburg
Co-presenter: Rolf Rasche, CEO of ImageWare Components GmbH, Bonn

Title:

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

Abstract:

The Herder Institute possesses the biggest newspaper cutting collection referring to Eastern Europe in the German-speaking countries. More than 6 million newspaper cuttings from mostly the 1950s to 1990s are a treasure for research on the “socialist experiment” in East Central Europe. The collection is arranged by themes, persons, and location. Newspapers from about 15 East and West European countries and in more than 10 languages were regularly analyzed and stored.

The press archive collection is now in the process of being digitized. A major question in this context is how to make the content easily accessible in a convenient way (i.e. in accordance with the conditions of the “digitized world”). The very strict German copyright law is an additional hurdle in the process of tapping the full potential offered by OCR.

Therefore some experienced software houses were asked to develop a flexible software tool to generate meta-data from various source materials and to make the digitized content easily accessible.

The basic problems encountered before and beyond the mere process of digitization turned out to be

- the different sources (newspapers, journals, periodicals, newspaper clippings/dossiers),
- different paper types and qualities,
- different paper sizes, individual graphic designs and compositions of each newspaper or journal,
- semantic aspects (content): texts in one language, bi- or multilingual texts
- different typefaces as well as
- the direction of reading and writing (e.g. Hebrew texts).

The software tool needed should therefore cover the following key functions:

- preparation of the images to fit a specific internal standard (scans, epaper);
- analysis and automatic marking of the different elements such as headlines, text, photos with text underlines, charts, diagrams;
- data output in different formats like PDF/XML, Mets, Alto and others.

The paper will give a first insight into these difficulties as well as the results achieved.

Biographies:

Dr. Juergen Warmbrunn: Born in Westfalia, study of Slavistics/Anglistics/Eastern European History and Finno-Ugrian Literature; 1992 graduation to dr. phil at university of Münster; 1993-95 education as librarian; since 1999 head of scientific research library of Herder-Institute, Vice-Director of Herder-Institute.

Rolf Rasche: Born in Westfalia; study of informatics in Bonn, diploma in 1987, employment as software engineer and consultant, since 1995 managing director of ImageWare Components GmbH (development of e.g. capturing client software BCS-2 and document delivery system MyBib eDoc, since 2005 catalogue enrichment with German and Swiss libraries, development of MyBib eL for copyright protected content;