Die Grenzboten on its way to virtual research environments and infrastructures

Manfred Nölte, State and University Library Bremen, noelte@suub.uni-bremen.de
Martin Blenkle, State and University Library Bremen, blenkle@suub.uni-bremen.de

Abstract
The State and University Library Bremen (SuUB) is dedicated to the digitisation of its historical collections. Digitisation is an important instrument in improving the accessibility of the valuable information contained in fragile historical documents. It facilitates academic research and teaching and is indispensable to the digital humanities.

Usually, digitisation projects produce digital images, metadata for cataloging and web-navigation purposes and OCR full text for searching. This information is made available through the library's web portal for digital collections. However, digital humanists need high-quality full texts enriched with metadata in the right format to analyse them with powerful software tools.

The historical journal "*Die Grenzboten*" serves as an exemplary model to bridge the gap between digitisation projects in libraries and research infrastructures. "*Die Grenzboten*" is a long running serial publication (1841 – 1922). It can be classified as a literary journal that also covered politics and arts. We demonstrate that OCR post correction and a page-wise structuring are prerequisites for the creation of a high-quality TEI version of a full text. The TEI version was created in cooperation with the Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). A fully automated OCR post correction was developed at the SuUB Bremen.

To enable scientists to work with powerful software tools the transfer of high-quality full texts to research infrastructures is a necessary step. The question now is: What has to be done to prepare raw OCR output for this purpose in a reasonable and cost-effective manner? What quality is needed or expected? Which metadata and file formats are needed? Shouldn't there be closer cooperation between research infrastructures and digitizing libraries? OCR full texts, even post corrected, are not perfect but character recognition rates around 99% certainly provide more options than just being used as a search index. There is a vast amount of textual resources out there ready to be made fully accessible for scientific research!