Quality Control in Mass Digitisation with ADOCO

Max Kaiser
Austrian National Library, Josefsplatz 1, 1015 Vienna, Austria. Email: max.kaiser@onb.ac.at

Jeanna Nikolov-Ramírez
Austrian National Library, Josefsplatz 1, 1015 Vienna, Austria. Email: jeanna.nikolov@onb.ac.at

Georg Petz
Austrian National Library, Josefsplatz 1, 1015 Vienna, Austria. Email: georg.petz@onb.ac.at

Christa Müller
Austrian National Library, Josefsplatz 1, 1015 Vienna, Austria. Email: christa.mueller@onb.ac.at

Martin Reisacher
Austrian National Library, Josefsplatz 1, 1015 Vienna, Austria.

Abstract
In a Public Private Partnership with Google the Austrian National Library (project “Austrian Books Online, ABO”, operational start was December 2010) is systematically digitizing 600,000 public-domain books/~200 Mio pages and gradually making them available via its Digital Library, Google Books and Europeana. When dealing with such large amounts of data quality assurance is a challenge. The quality assurance work in ABO was started in June 2011.

The presentation will explain the approach chosen by the Austrian National Library and describe the evolvement of the quality chain over the past year. It will trace the processes established in the library for downloading the digitized files and checking their quality, both with semi-automated methods (such as combining text field-size and words per page to detect suspicious cropping) and in manual audits.

Special software (ADOCO – ABO Download and Control) was implemented and is continuously developed to meet the needs of the quality auditing process. ADOCO enables simultaneous, multithreaded downloads. It is based on Primefaces and Spring Webflow, using Linux command line tools wrapped in JAVA (wget, tar, gpg, exiftool for image metadata, md5sum,..) and uses a MySQL-Database for technical and bibliographic metadata. It allows for various searches and views on the relevant volumes.

The talk will cover considerations for the establishing of quality metrics and quality indicators, possible error phenomena, error severity scales, clustering of errors, tools that were evaluated and used in the project so far and the strategic approach to quality assurance in mass digitization.

Additionally we will outline future quality perspectives in mass digitization, such as collaborations of ABO with the IMPACT Centre of Competence for Digitisation and the FP7 project SCAPE and experiments and tools developed there.

Requirements, challenges and opportunities from the first year of quality assurance will be shared as Lessons Learned.

Keywords: quality assurance, mass digitization, metadata management, error metrics.

Curriculum Vitae

Max Kaiser is Head of the Research and Development Department at the Austrian National Library, Vienna. He has many years of experience in R&D in the areas of digital preservation, digitisation, and digital libraries. He has been involved in a large number of EU R&D projects in different research programmes of the European Commission, having been coordinator of the EuropeanaConnect project. He is also member of the Management Board of “Europeana V.1.0” and is leading the “Innovation” work package of the “Europeana V.2.0” project. Max Kaiser is member of the Board of Directors of the Open Planets Foundation, a not-for-profit organisation which provides expertise and practical solutions in the area of digital preservation. He acts as programme lead for “Austrian Books
Online”, the Austrian National Library’s large scale digitisation cooperation with Google. In this project, the largest public-private-partnership in Austria’s cultural sector so far, the Austrian National Library’s entire historical book collection will be digitised and put online.

**Jeanna Nikolov-Ramírez** is the operational project manager of Austrian Books Online - the Public-Private Partnership between Google and the Austrian National Library in which 600,000 books are being digitised and made accessible online. Before joining the Department for Research and Development at the Austrian National Library she worked as director of postgraduate courses at Danube University Krems, where she taught classes in digital collections management, iconography and media art histories bringing together experts from informatics, business administration, libraries, museums and archives. She was involved in the EU projects DM2E, IMPACT, PLANETS and EuropeanaConnect, in the latter as project manager. She is also an active member of the Google Quality Working Group of the Google Books project.

**Georg Petz** is a Software Developer at the Austrian National Library. Currently he is working in the Austrian Books Online project - the Public-Private Partnership between Google and the Austrian National Library and implemented ADOCO, the quality control application. He holds a master in business informatics from the Vienna University of Technology. Georg Petz joined the Department of Research and Development in 2007 and was involved in the EU projects TELplus, PLANETS and EuropeanaConnect.