# TEI, ALTO and METS. Why we need all of them.

*Günter Mühlberger, University of Innsbruck / Austria, Department for German Language and Literature*

## Problem statement

Linguists who are dealing with historical corpora benefit from large scale projects where millions of books are digitised and made available to the public. In order to get full-text from digital images Optical Character Recognition (OCR) is applied – a technology that is mature and cheap but that produces especially in the case of historical documents erroneous data. Whereas books from the 20th century may be recognized with a word accuracy of up to 95% or even more, in the 19th and especially in the 16th to 18th century word accuracy rates go down to 80, 70 or even 50%.[1] The problem therefore is how to deal with this uncertainty in the TEI document edition process?

## Discussion of formats

Whereas TEI is “the” format for editorial annotation it plays a much less prominent role in the digitisation and library community. Here the Analysed Layout and Text Object (ALTO) format is widely spread. Introduced by Stehno et al[2] in 2003 it is now hosted by the Library of Congress[3] and was adopted not only by a number of libraries but is also supported as a native output format by the text recognition and conversion industry.[4] ALTO was introduced due to the fact that a simple format was needed that serves two main requirements: Firstly to deal with text produced by automated procedures (OCR, Optical Layout Recognition - OLR) including the word accuracy rate or word variants suggested by the OCR engine and secondly to keep the link to the source image by encoding the layout information (coordinates) on block, line, and even word level. Furthermore ALTO was intended to accomplish the Metadata Encoding and Transmission Standard (METS) which is a container format for digital objects originating from digitisation projects, a standard that is also hosted by the Library of Congress.[5]

---

1C.f. Tanner, Simon (2009): Measuring the OCR Accuracy across The British Library’s 2 Million Page Newspaper Archive. Presentation at the IMPACT Conferene on IMPACT Conference: Optical Character Recognition in Mass Digitisation. The Hague 6 and 7 April 2009. Online at: http://www.kdcs.kcl.ac.uk/fileadmin/documents/pubs/IMPACT_Simon%20Tanner%20OCR%20Accuracy%20BL%20April%202009.pdf

2Stehno, Birgit; Egger, Alexander & Retti, Gregor (2003): METAe - Automated Encoding of Digitized Texts. In: Literary and Linguistic Computing, Vol. 18 No. 1 p. 77-88.

3Library of Congress‘ website: http://www.loc.gov/standards/alto/

4The German company CCS GmbH used ALTO at several newspaper and book digitisation projects worldwide. The Russion company ABBYY which is one of the main OCR software vendors supports native ALTO output since 2011.

5 Library of Congress‘ website: http://www.loc.gov/standards/mets/

## Proposed solution

Our contribution to the workshop will be to discuss in more detail the strengths and weaknesses of the TEI, ALTO and METS format always with regard to textual data stemming from automated processing such as OCR or Handwritten Text Recognition (HTR). As already indicated these formats are full-filling specific purposes within the digitisation process. Since millions and millions of ALTO files are stored in libraries all around the world a practical solution has to be found to deal with the transition from images to automatically recognized text and further on to manually annotated and edited text. We will discuss and propose simple criteria where and when which standards may be applied and which transition rules could be used.

## Background

We are dealing with the problem of OCR data since more than 10 years and are currently involved in two EU projects where this question plays a crucial role: In the EU Newspaper project[6] we are responsible for OCR processing of 8 mill. newspaper pages from 11 European libraries. The final output are METS/ALTO packages. In the FP7 project tranScriptorium[7] DEA is setting up a Transcription Platform for semi-automated transcription of historical handwritten texts. The platform shall not only serve researchers (transcribers) but also involve archives (as content providers) and has therefore again to deal with the right choice of formats for each community.

## Presenter

Dr. Günter Mühlberger is Senior Project Manager at the University of Innsbruck (Austria), Department for German Language and Literature. He received his Dr. for a dissertation on Johann Wolfgang von Goethe (1749-1832) and worked from 1991 to 1997 as research assistant and lecturer. From 1998 onwards he initiated and co-ordinated several EU R&D projects in the digital library domain with a focus on the development of Optical Character Recognition (OCR), as well as information extraction and annotation of historical documents. From 2002 to 2012 he was head of the Department for Digitisation and Digital Preservation of the University Innsbruck Library. In 2008 he founded the library network E-Books on Demand (EOD) which currently comprises more than 30 libraries from 14 European countries and provides on-demand digitisation of historical books. In 2012 he returned to the Department of German Language and Literature and is now focusing on Digital Humanities issues. His group is involved in three EU projects among them tranScriptorium, a project that deals with the Handwritten Text Recognition (HTR) of historical documents.

---

6 ICT-PSP project Europeana Newspaper: http://www.europeana-newspapers.eu/

7 EU FP7 project tranScriptorium : http://www.transcriptorium.eu/