CONSORTIUM OF EUROPEAN RESEARCH LIBRARIES

The application of text encoding facilities to digital version of European early books

Marie-Luce Demonet, Professor of French literature and senior member of the Institut Universitaire de France, taught at the Université de Provence, then in Clermont-Ferrand, Poitiers and Tours (since 2001). Specializing in connections among literature, languages and semiotic theories during the Renaissance (*Les Voix du signe*, 1992), she has published several articles on Rabelais, coordinated the transcription of his novels on the Internet (1995), and published conference proceedings (2006, 2008, 2010). She has also published books, and essays on Montaigne (1985, 1992, 1999, 2002, 2003). She takes interest in transformations of philosophical, juridical and social vocabulary and in questions of theory and literary aesthetics, and about the status of fiction and sign theory in Early Modern times. She was the Dean of the Center for Renaissance Studies in Tours (2003-2007), the Director of the Maison des Sciences de l'Homme (Institute for Social Sciences and Humanities, Universities of Tours and Orleans, 2012-2014), and the Head of the "Bibliothèques Virtuelles Humanistes" team: the program is devoted to conducting research consisting of indexing text in image mode, extracting images from scanned pages, acquiring significant corpora of transcribed texts, classifying and indexing them. Her major project in the field of digital humanities is the exhaustive digital edition of the "Montaigne library".

Graphical analysis and graphematics for the BVH corpus (Bibliothèques Virtuelles Humanistes)

OCRing French Renaissance documents needs not only an accurate tool to analyze the page layout and the shapes of the fonts, but also a systematic library of these fonts supported by a scholarly knowledge of graphemes and of historical languages. Examples will show how textometry and statistics can help the automatic detection of parts of speech, and the enforcement of n-grams methods, in order to reduce the spelling variation of Renaissance texts, and to help character recognition.

Martin de la Iglesia works for the project "Genetic-critical and annotated hybrid-edition of Theodor Fontane's notebooks based on a virtual research environment" as a librarian in the Metadata and Data Conversion group at Göttingen State and University Library. He earned a Magister Artium degree in Art History and Library Science from Humboldt University Berlin in 2007 and has been working as a metadata specialist since then.

Metadata and other data in TEI

Since their first publication in 1990, the Text Encoding Initiative Guidelines (TEI) have emerged as the most widespread standard for XML-encoding transcriptions in digital editions. However, TEI code is much more than just a wrapper around full text. Depending on the depth of encoding, TEI documents may contain diverse and plentiful metadata, e.g. bibliographic data pertaining to the encoded source material, or named entity references linking to authority data. Such data, once identified in a TEI source, can be easily aggregated, processed, analyzed, and visualized.

Klaus Kempf is the Head of the department for collection development and cataloguing at the Bayerische Staatsbibliothek, Munich. He has a special interest in library management affairs, in particular reorganisation and reengineering , personnel management, library building (consulting projects on regional, national and international level).

OCR-ing at the Bayerische Staatsbibliothek: projects and experiences

Bente Lavold, research librarian/philologist at the National Library in Norway. Bente works with history and theology.

Ellen Nessheim Wiger, research librarian/philologist at the National Library in Norway. Ellen works with scholarly editing, text encoding and digital publishing.

From manuscript to epub

«'Moer Korens' dagbøger» (Mother Koren's diaries) are journals written by Christiane Koren from august 1808 to February 1815. All together it is about 2200 pages. The beginning of the 19th Century is a very important period in Norway's political history. This year, 2014, we celebrate the 200-anniversary for our 1814 Constitution. The Constitution was written in Eidsvoll not far from Hovind, the home of Christiane Koren and her family. She was very engaged in the political situation and was well orientated by friends that took part in the meetings at Eidsvoll.

Koren's diaries were not written as private journals, but were circulated and read by everyone in her large network. These journals are kept here at the National Library in Oslo and can be seen and studied here, or in facsimiles online. Parts of the journals were published by Sofie Aubert Lindbæk in 2 volumes in 1915 (available for reading and wordsearch at www.nb.no, and also available as PDF and EPUB downloads). However, we discovered that about 600 pages of the journals were not included in these books, and we therefore now aim to publish the rest of Christiane Koren's journals as ebooks. The manuscripts are transcribed, proof-read, and encoded in TEI P5 XML before being transformed to different digital formats such as HTML and EPUB. All programs, tools and standards used are open and free. At the seminar we will show you details of the process as well as the final products.

Andrea Mazzei and Fouad Slimane are Post-docs in the Digital Humanities Lab at the Swiss Federal Institute of Technology in Lausanne (EPFL).

Andrea Mazzei received a Bachelor and a Master in Computer Engineer from the University "La Sapienza" in Rome. Subsequently he obtained a PhD in Computer Science in the laboratory "Computer Human Interaction in Learning and Instruction" at EPFL. During his PhD he created new eye tracking and social technologies. In particular, he investigated how they could lead to transformations in education, by empowering the way learners read and connect to each other. His research interests now focus on methods for intelligent offline character and handwriting recognition and on modelling uncertain spatio-temporal data.

Fouad Slimane, was a Postdoc for one year (June 2013 - June 2014) at the Institute for Communications Technology (IfN), Department of Signal Processing for Mobile Information Systems, Braunschweig, Germany. He graduated in computer sciences engineering in 2004 from the National Engineering School of Sfax, Tunisia, He received his master degree in computer sciences in 2005 from the University of Rouen, France. In June 2013, he received the PhD degree in sciences from the University of Fribourg, Switzerland in join collaboration with National Engineering School of Sfax, Tunisia. His research interests include image processing and analysis, neural networks, pattern recognition, Arabic text recognition and Hidden Markov Models. He developed the APTI Database (used by more than 50 research groups all over the world) in 2009. He organized the Arabic Recognition Competition: Multi-font Multi-size

Digitally Represented Text at ICDAR 2011 and ICDAR 2013 and the Arabic Writer Identification Competition using AHTID/MW and KHATT Databases at ICFHR 2014. He published more than 20 papers including conference papers, one journal paper and one book chapter. He is a member of program committees of many conferences. He is a frequent reviewer for international journals and conferences.

Incremental and Offline Handwriting Recognition for the Venice Time Machine

The Venice Time Machine aims at unlocking the contents of the Venice State Archive through innovative digital methods. The automatic recognition of handwritten text from several periods and documentary series is an open field of research.

The Digital Humanities Laboratory challenges the problem with two overlapping methodologies: 1) offline handwriting recognition based on state of the art algorithms (HMM, Recurrent Neural Networks); 2) the incremental transcription paradigm, where we provide a transcription online facility which benefits from the input of users to enhance the quality of the outcomes incrementally. Both approaches can be integrated as services into other tools or procedures, such as the semi-automatic extraction of keywords and other metadata for the classification of digitised documents.

We will present the current state of our research with concrete up to date outcomes, reliability assessments, and critically discuss the role of transcription tools and services into a multi-project collaborative endeavour as the Venice Time Machine.

Tomasz Parkoła, M.Sc. Eng. in computer science, is a member of Digital Libraries Team at Poznan Supercomputing and Networking Center. His research interests and expertise cover various aspects of digitisation, including software for digital libraries and museums, digitisation workflow management systems as well as long-term preservation tools. He works on several R&D projects in this field: dLibra, dMuseion, dLab, dArceo, VTL, Cutouts. He is also involved in the IMPACT Centre of Competence in digitisation and Open Planets Foundation. He is the author and co-author of over 20 papers in conference proceedings.

The Impact Centre of Competence: tools for text digitisation and transcription

The talk will introduce the Impact Centre of Competence in Digitisation, its main assets (data sets for historical languages and tools for digitisation) and activities (events, support for the community and knowledge dissemination). Some tools of special interest for the community of research libraries will be also presented, for example, the Virtual Transcription Laboratory, post-correction tools and tools for the evaluation of the quality of transcriptions.

Michael Popham is Head of Digital Collections and Preservation Services at the Bodleian Libraries, University of Oxford. He has worked on the creation of electronic textual resources for the Digital Humanities for more than 20 years, and has previously led the work of the UK's Arts & Humanities Data Centre for Literature and Linguistics, and managed Oxford University's e-Science Centre.

When not to OCR -- the experience of EEBO-TCP

The Early English Books Online Text Creation Partnership (EEBO-TCP)is a collaboration between ProQuest and more than 150 libraries to generate highly accurate, fully-searchable, SGML/XML-encoded texts corresponding to books from the Early English Books Online Database. From the outset, EEBO-TCP chose to have material rekeyed rather than OCR'd, and this presentation will outline the considerations which underpin this decision and reflect upon its consequences.

Lambert Schomaker is full professor in artificial intelligence at Groningen University, The Netherlands and the director of the AI institute at the faculty of Mathematics and Natural Sciences. His main and general interest is in pattern recognition and machine learning: How can machines be made more

intelligent by training them to perform cognitive tasks? In his country, he is active in program committees on e-Humanities for the national science foundation and the royal academy of sciences. His work on pattern recognition and neural networks is a precursor to modern handwriting and gesture-recognition methods on tablet computers such as the iPad. The work of prof. Schomaker is cited in 13 US and 10 international patents. In 2004 he started the study of handwriting recognition in historical collections and is currently active in using mass-storage, high-performance computing in order to build a general search engine for handwritten historical archives: 'Monk'. Apart from handwriting recognition he has work on forensic and paleographic computer-based identification of the hand and on style-based manuscript dating.

The birth of a massive search engine for historical and multi-cultural handwritten collections

In recent years, important progress has been made in the area of computer-based processing of handwritten-manuscript images, both in terms of identification of 'the hand' and in retrieval of keywords or text passages. The MONK system at the University of Groningen, The Netherlands, is a large-scale ongoing effort for continuous training and word spotting in large handwritten collections, ranging from the Qumran Scrolls to medieval, historical and, recently, also Chinese handwritten and wood-block printed text. The diversity of handwriting styles, the contractions and styled abbreviations in historical material make traditional character-based ('OCR') approaches very cumbersome. We have shown that word-image based approaches can yield very interesting results, and the 'bootstrapping' of collections, i.e., starting from zero knowledge to a keyword search engine for a historical collection is entirely feasible with limited human effort. The reason for this success is the vastly increased computational power, paired with 'Big Data' methodologies. The current system contains about 65 books, many with ~1000 pages. There is an increased interest from international research groups and archives. Apart from keyword search, the digital availability of historical manuscripts also allows for specialized functions such as writer (hand) identification and dating. These fast developments illustrate that this application domain is currently enjoying great strides forward.

Zane Vitolina is a passionate IT product marketer. A marketing expert. A manager of important social IT products and IT projects in Latvia. A hobby photographer. Zane's achievements in Latvian IT industry have contributed to a lasting value for society. Many social projects are implemented under her leadership, with all parties — the customer, the supplier, and society — coming on top as the winners. The school management system "E-klase". Nowadays, approximately 80 % of all schools in Latvia use an online grade book and an e-diary. Now, parents have full access to their children's achievements in school by using a web-based school management system and teachers have the opportunity to prepare various reports generated by this system. Every forth municipality in Latvia uses the learning support system for children and young adults who are facing the risk of social exclusion. The system was developed within the framework of a project called "Hand in hand for child's support". The main goal of this system is to help municipalities gather and share information with other institutions about children, who need some help, whether social or financial. More about the work experience and job positions on LinkedIn at: lv.linkedin.com/in/zanevitolina/

How to generate society's interest in digitized books and periodicals?

Can old periodicals and books be interesting for society? Why should society be interested in digitized heritage? Who is interested in this information? And, if people are not interested in it, how can we make this information more appealing? How can we get society back to the library? Is it possible to return people back to libraries? It is possible. There are some simple, but old rules, which work in every community.

The world over, people are talking about changes in libraries. The customer aspect in every library is crucial — the only visitors to libraries are loyal readers, librarians and researchers. Libraries are trying to be innovative by using everything that could be interesting for society, such as social events, book fairs, online libraries, digital libraries and much more. In some countries, libraries serve as the community's culture centre. Therefore, a question arises: not 'what' but 'who' is a library? A library is not only a building full with books, periodicals and artefacts. A library is a community.

Communication always occurs between two parties (at least), and there is always a certain topic — people are always talking about something. The topic must be interesting, and if it is not, the one, who is leading the communication, must make the topic interesting. In order to get the readers back and to attract new ones, libraries must to talk to society in a very user-oriented language and it must to use modern tools to lead that communication. Libraries have a role of the communication leader. And that is the key word for a modern library.

ICT tools are just a tools, which can help people in many ways: to organize work, to reduce waste, to increase income etc. I would like to talk about communication, ICT tools and the modern reader, because I believe, that a good ICT tool, developed to serve the users instead of customers, is the shortest way to success. For customers, too.