"EUROPEANA NEWSPAPERS" PROJECT WORKSHOP ON TEXTS REFINEMENT AND QUALITY ASSESSMENT

Aleksandra Trtovac, aleksandra@unilib.bg.ac.rs
University of Belgrade, University Library „Svetozar Markovic", Belgrade

University Library "Svetozar Markovic" in Belgrade is one of the 18 full partners in the project "Europeana newspapers" along with the most important national and university libraries of Europe and LIBER, German company CCS[1], and 11 associated partners. The users of Europeana portal will have access to the full text search and information retrieval on over 18 million digitized pages of newspapers. The project "Europeana newspapers" lasts from 2012 until 2014.

Within the project, at the University Library in Belgrade on 13 and 14 June 2013 the workshop on the assessment of the quality and refinement of digitized newspaper articles was held. At the workshop about fifty librarians and information specialists from different parts of Europe took part (from Finland, Estonia, Iceland, Denmark, than France, UK, Germany, the Netherlands, Switzerland, Spain, Poland, Czech Republic to Slovenia, Croatia, Bosnia and Herzegovina, Bulgaria and Macedonia). The workshop was attended by colleagues from the National Library of Serbia, the Nikola Tesla Museum, University Library of the University of Nis.

After a welcoming speech by prof. Dr. Aleksandar Jerkov, director of the University Library "Svetozar Markovic" and basic information about the workshops presented

1 Content Conversion Specialists

by Marike Wilms, the main organizer of the workshop and the representative of LIBER, we had the opportunity to hear presentations and helpful suggestions of colleagues working on developing of applications, testing and analysing of delivered digitized content. Also, we have actively participated in small group discussions and left comments on the so-called "Democracy wall" where Marike Wilms had prompted several topics: "I discovered that...", "I noticed that...", "I feel that...", "I have learnt that...", "I would like to suggest... ".

Hans-Jörg Lieder from the State Library in Berlin, the project coordinator, spoke about technologies used during assessments of quality and refinement of newspaper articles - optical character recognition (OCR), optical layout recognition (OLR), named entity recognition (NER) and class recognition. Also, he emphasized importance of the project in terms of newspaper articles metadata - in addition to the full-text search, it is possible to search through the metadata of which the texts that describe the photos which newspaper articles are illustrated with are particulary interesting. Also, Lieder spoke about the challenges and difficulties encountered during the refinement of old newspaper texts. First of all, old newspapers are printed by poor quality ink, which makes optical character recognition more difficult. Benefits for users of Europeana, and the partners in the project include the ability to perform keyword search and phrase search, image browsing, and information about them, text mining, user sourced correction and enrichment of contents, access to contents via mobile applications...

Clemens Neudecker from the National Library of the Netherlands in The Hague spoke about the process of refining newspapers text and developing tools for better data organization of contents provided by project partners. Two partner institutions are engaged in these tasks - the University of Innsbruck and the German company CCS. The accompanying steps of refinement process are: newspaper selection, determination of usage rights, metadata collection (Master list), binarization, file renaming and folder structure, final check of data and metadata, OCR/OLR, METS[2]file production, and finally NER. Problems in these processes are mainly related to supplying contents with different digitization and access policies and large variety of contents in term of file formats, fonts, languages etc. The tools that have been developed for better refinement of text and data organization are: BCT - Binarization and Colour Reduction Tool, FRT - File Rename Tool, FAT - File Analyzer Tool. ABBYY FineReader SDK and software State-of-the-art OCR are used for optical character recognition, docWorks technology for optical layout recognition, and technology Stanford CRF-NER for named entity recognition, supporting German, English and Dutch as well as French and Lithuanian for named entity recognition of persons, locations and organizations.

Stefan Pletschacher and Christian Clausner from the University of Salford presented difficulties in OCR of historical newspapers. Quality assessment is important because of the use of fonts that are difficult to read which often causes errors in the digital version. By comparing the two versions, the accuracy of the OCR process could be measured and this was in turn helpful when setting requirements for the outsourcing of OCR work.

After these lectures and analyses, small groups were formed with representatives of various institutions who had the opportunity to become more familiar with each other, discuss and share experiences related to the organization of work in the project, the digitization of newspapers and make suggestions for improving the project.

The first day of the workshop ended with

2 *Metadata* Encoding and Transmission Standard

a visit to the Nikola Tesla Museum and a joint dinner of all participants.

Second day of the workshop started with presentation about NER by Lotte Wilms from the National Library of the Netherlands. She explained that the project anticipates named entity recognition to persons, locations and organizations and showed examples of the mentioned named entities. During the practical part of the workshop, participants divided into two groups watched the presentation of software Named Entity Attestation Tool. The software is easy to use and allows easy editing of recognized entities, but there was a small problem that the workshop participants were not familiar with the Dutch language, and they were not able to participate more actively. So far, this tool is available for named entity recognition in German, Dutch and French.

During the workshop, prof. Dr. Cvetana Krstev and prof. Dr. Dusko Vitas, representatives of the Group of Human Language Technologies, Faculty of Mathematics, University of Belgrade, reached agreement with the coordinator of the project, Hans-Jörg Lieder, to do named entity recognition for Serbian language of processed texts of Serbian historical newspapers too.

Then Klaus Gravenhorst presented the results of the CCS related to optical layout and metadata recognition using the software docWorks. To achieve the best results it is necessary to structure unstructured texts first. General rule system enables recognition of words, text lines, text blocks, columns, ilustrations, advertisements, tables and the following page types: title page, contents page, illustration page (a page that has at least one illustration), a page that contents advertisements only. Structural analyses through classification of headlines including article continuation is also possible. The software docWorks enables conversion of recognized articles metadata into METS XML format.

Last presentation was given by Ginter Müxlberger from the University of Innsbruck and he talked about metadata formats used in the project - in the first place METS and ALTO[3]. The focus was primarily on structural metadata, and following structural elements were suggested: title section, headline, advertisment, illustration, column title, page number... Also, suggestion was made to organize the articles according to the following types: breaking news, short news, book review, theatre review, obituary, advertisment, job announcement, weather forecast, novel, poem etc. Participants in the project were encouraged to make proposals for structural metadata, considering the needs of users, as well as to analyze the formal content of the paper.

In addition to new information on the project and the development of applications and tools, the workshop was good opportunity to meet many librarians and IT professionals who came from different parts of Europe in nice and friendly atmosphere. It was a good opportunity to learn something new, to share experiences and to create new possibilities. As librarian of the University Library, I'm extremely glad that, as it seems to me, we were very successful in organizing of «Europeana Newspapers» project workshop as well as various social events for our dear guests and colleagues.[4]

Accepted: 16th August 2013

3 Analyzed Layout and Text Object

4 Presentations and videos from the workshop are available at "Europeana Newspapers" project web site - http://www.europeana-newspapers.eu/focus-on-newspaper-refinement-quality-assessment-in-belgrade/ and the YouTube channel of the University Library "Svetozar Markovic" - http://www.youtube.com/user/UBSMBeograd