Date submitted: 15/06/2010

Past publishing and future digital developments for news and newspapers

Wolfgang Novak and
Stephan Tratter
TREVENTUS Mechatronics GmbH
Vienna, Austria

Meeting: **102. Newspapers**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Abstract:

Newspaper collections are a very good and important information carrier. One of these is the newspaper "Burgenländische Freiheit"[1] from Burgenland – an Austrian federal state. To make the access to the "bound" information easier a digitizing project was set up.
Within this project several requirements had to be met and problems had to be solved. On the one hand quality was an important issue. The aim of the project was not only to digitize the newspaper but also to get a digital archive. The identified problems were different sized editions, different amount of pages and supplements within several editions. Also the question how the digitized images have to be presented on the internet and which keywords will be used for search routines had to be deliberated.

The Newspaper

As the Austrian federal state "Burgenland" became part of the country Austria, in 1921 the weekly newspaper "Burgenländische Freiheit" was founded by the „SPÖ Burgenland"[2]. In 2006 politicians in Burgenland decided to change the name in "Burgenländische Freizeit", to publish the newspaper every 2 weeks and to provide it to the citizens for free. Due to cost reasons in January 2009 the publication of the newspaper was stopped.
All in all these books have about 200,000 pages of two different formats. From November 1921 until January 1967 the weekly newspaper was published in the format 266 mm × 421 mm and from February 1967 up the newspaper was published in the small format of 196 mm × 265 mm. The weekly editions became bound into 129 books for archiving reasons. 35 of these books are in the big format and 96 books in the small format.
A complete and very good preserved collection of the bound editions is archived in the "Landesarchiv Burgenland"[3]. A 2nd collection can be found in the archives of the national party

[1] http://de.wikipedia.org/wiki/Burgenland_Freizeit, 30.04.2010, further information to the newspaper Burgenländische Freiheit".
[2] Social Democratic Party of Austria (German: Sozialdemokratische Partei Österreichs, or SPÖ).
[3] http://www.burgenland.at/landesarchiv, 30.04.2010, contact data of the "Landesarchiv Burgenland".

SPÖ in Vienna[4]. This collection is not as complete as the collection from the Landesarchiv Burgenland, but it is considered as a well prepared source for cross checks in case of missing supplements, pages or week editions.
The newspaper was a very well known information channel in Burgenland and it was requested several times to have the possibility to do research in the archived bound books. Often the requests came from interested people who tried to research for historical events but even for birthday edition requests came in.

The project

Historical events capture the interest of people and therefore they often do searching for events in the past. Especially the demand for "birthday editions" is quiet enormous. As the search process in bound books can only be done in the Landesarchiv Burgenland it is rather time consuming and cost intensive. Furthermore, people often could not manage to do researching alone and asked the Landesarchiv for help. Therefore the idea of digitizing newspapers was born. The aim behind this project was to save both time and personnel resources needed for searching within printed editions and to provide interested people easy open access to one part of the cultural heritage of the Austrian federal state Burgenland.

Partners of this project were the Landesarchiv Burgenland, Medienhaus, scharf.net and TREVENTUS Mechatronics GmbH. Medienhaus is a company that does many research projects on press and media and as coordinator was responsible for organisation and communication. scharf.net, a programming agency for new media, was responsible for the search module and the internet display. Treventus acted as service provider for digitization and logistics around the book and scanned digital objects.

The project was divided into three phases which were testing and specification, digitization and content enrichment (adding additional historical content) and homepage launch. During all phases there have been meetings to discuss next steps and further procedures.
The first phase started with test scans to check if the bound books were scannable and to get a feeling for the expected image quality. Very important in this phase was the defining of the problems that were expected to appear. In addition the idea to generate additional value to the content of the newspaper was born which consisted of adding editorial articles related to selected cultural events.

During the second phase books were digitized and additional content was created. Editors in cooperation with some important people from the time when the newspaper was produced wrote the additional articles. All this to enrich the content with the context of the former time and to show the historical context. By that an all-embracing view to historical events was possible combining local events and global ones due to this content enrichment step. The development of an adequate search module of easy use was also an important aim in the project and part of the second phase. Here the discussions about the requirements of the modules and the defining of the keywords and the design were very important prior to that implementation.
In the third and last phase the homepage was released. First the beta version was tested by all the participating partners, regarding usability, content performance, etc. over a certain time. After the beta tests the homepage was launched. A central person was responsible for collecting all the inputs coming in from the users to improve new functions.

[4] http://www.wien.spoe.at/, 30.04.2010, SPÖ-Archiv.

Challenges to digitization

The newspaper was digitized using the automatic book scanner "ScanRobot® SR301"[5] [see Fig. 7] and with the manual book scanner "Bookeye 3 A2"[6] [see Fig. 9].

One problem of digitizing had been the binding of the books itself. The newspapers had been bound to annual books. Within the process of binding often the first letters or just some part of it in each row are not visible on a scanned image [see Fig. 3]. This was an important problem for the OCR (Optical Character Recognition) as the words with the missing letters could not be detected. If a person reads such an article he can assume the missing letters and knows the word in combination with the context. For the OCR software this step is not possible at the moment. As a result these words cannot be found by the search module.
A lot of pages had wavy margins [see Fig. 2]. Due to the innovative scanning technology of the automatic bookscanner this challenge did not cause any inconvenience to the scanning process.
Another issue were creased pages [see Fig. 5] and image artefacts in the middle of the page due to folded newspaper [see Fig. 6] which on the one side make the automated digitizing more difficult and on the other side has an effect on the accuracy of the OCR results.
Images in newspaper often already contained Moiré effects [see Fig. 4] could not be improved during the scanning process. Therefore some of the scanned images contained Moiré patterns.
In post processing the renaming was an important issue. Not only every single weekly edition had to get its own name but also the editions differed in the number of pages because of e.g. supplements. Furthermore, these supplements were also a challenge for digitizing as they were of smaller size than the normal newspaper was. So the renaming process could not be done fully automatically but only with manual support.
To complete the list the data transfer has to be mentioned. As all images had to be transferred to another company within the project a simple way for the data transfer was needed. An upload of all images was not useful as this would have needed too much time. Therefore the simplest way were changeable hard discs as they are just pulled out of the Computer or PC and plugged in again.
For the quality control during the project the challenge of physically missing pages was very important. Although the two different formats had been homogeneously and could therefore be scanned automatically for the most part, physically missing pages had an important impact on the quality control. As in nearly every single newspaper pages were missing it had to be ensured that all pages were digitized because during the quality control it was not possible to say if these pages physically existed or not without taking the book physically and checking it. For economical reasons this would have meant that the quality control would have taken much more time and would therefore be more expensive. The challenge to ensure that every existing page was scanned was solved by a quality control during scanning.
During the development of the homepage the framework was the most important point that had to be decided. Therefore some existing homepages with similar content were compared and analysed regarding the advantages and disadvantages. Through that comparison and identified requirements the framework for the homepages was set up.

The OCR was planned for evaluation reasons only for the period between 1984 and 2007. Therefore the search engine had to be developed in a way that you can search for words and for other criteria that are independent from the text. For an easy and comfortable use of the search engine and the whole homepage the display had to be defined too. Here the questions how big the images shall be and which kinds of displaying and page turning are available were important questions as they have a huge impact if the homepage is accepted and used by the users.

[5] http://www.treventus.com/download/ScanRobot_SR301_product_folder_20100122.pdf, 30.04.2010, product description
[6] http://www.imageware.de/de/systeme/buchscanner/bookeye3-a2/, 30.04.2010, product description

As OCR was used it was interesting to know how high the hits will be and which kind of word will not be detected. Therefore OCR tests were made. The experience of other digitized archives showed that names and terms are very common keywords for search routines. To increase the hits especially for the expected names and terms a dictionary was made and implemented into the OCR software.

Realization

The books were digitized in the Scan-Center[7] of Treventus [see Fig. 8]. The transport of the books from the two archives to the Scan-Center was organized in three shipments to minimize the amount of trips on the one hand side and to ensure that the larger amount of books is available in the archive. In this way the interested people had a restricted access to the archived books. But at the same time several parts were shortly blocked as "not available" for the researcher.
For the digitizing process a workflow was developed to accelerate the whole digitizing process and to minimise the time the books were needed in the digitizing chain. For this the existing workflow software ScanFlow™ from Treventus was used. An individual workflow was adapted for all the books of the described project.
After the quality check, where every page was proved manually, the digitized pages were transferred to the partner scharf.net for uploading the images in the prepared database.
Even the implementation of the homepage was organized in several steps. After the last specifications were defined and the beta version finished the three partners of the project made the quality and usability check of the web platform. The focus in these tests was to find out if the requirements that have been defined at the beginning of the project were fulfilled and if these requirements were anticipated correctly. Through this quality check some mistakes were found and corrected. Additionally the user friendliness was improved to be sure that the homepage is accepted by the population and many people will use the homepage in the future.

Outlook

For the future three steps are considered. Now as the homepage is released it is expected that some more mistakes will occur that have to be corrected. It is also planned to enlarge the OCR to the rest of the years as soon as additional budget is available. Also the third part aims to the usability and the value of the available information. Here it is planned to link the information from the newspaper to historical data and events. The first part of this was already done by the additional articles that have been written. Here the aim is to link more broadly to get the link to information from other homepages.
With this pilot project it was planned to make a use case and to create a kind of standard for further digitizing and web design projects of that consortium.

Summary - experience

Within this project many different kinds of challenges were detected that can occur within digitizing projects. The challenges that occurred within this newspaper project were described and the solution presented.

The result of this project is an open access digital archive with searchable content (from the years 1984 up to 2007) that is enriched with editorial articles. Within the presentation the whole workflow of the digitizing process will be shown, as for example the handling of the books, the automatic and the manual digitization steps, the content extraction (via OCR) and the content enrichment. The interface will be presented and also some search examples will be given to show the combination of content from the newspaper and the editorial articles.

[7] http://scanservice.treventus.com, 30.04.2010, Treventus Scan-Center in Vienna

Fig. 1: Book thickness

Fig. 2: Wavy pages

Fig. 3: No inner margin in the spine

Fig. 4: Moiré effect

Fig. 5: Creased, buckled page

Fig. 6: Text background, folded pages

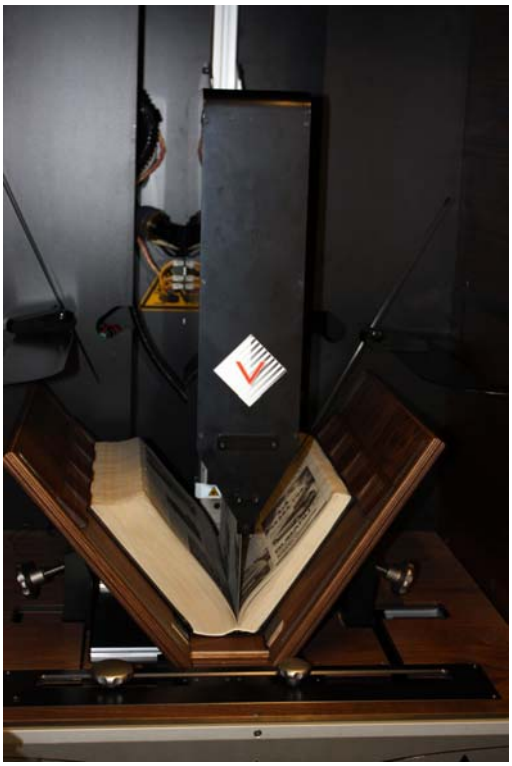

Fig. 7: Automatic bookscanner ScanRobot® SR301

Fig. 8: Shelf with small books and in the lower part the big size books

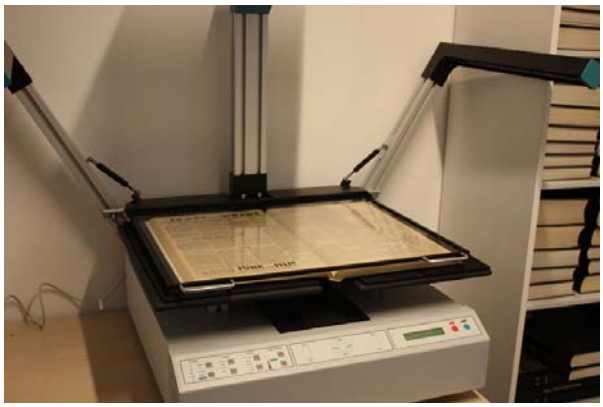

Fig. 9: Manual book scanner