News Media Section

Retrodigitization of historical newspapers – workflow, archiving and presentation

Ludger Syré
Head of the Digitization Department, Baden State Library, Karlsruhe, Germany
syre@blb-karlsruhe.de

Copyright © 2016 by Ludger Syré. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License: http://creativecommons.org/licenses/by/3.0/

Abstract:

Historical newspapers belong to the printed cultural heritage. Unfortunately former generations of librarians did not pay very much attention to these ephemeral materials. For that reason many newspapers which appeared in Germany between the 17th and the 19th centuries did not come down to us. They perished because nobody felt responsible to preserve and to archive them properly.

Nowadays the situation has changed completely. Academics and students as well as private researchers have discovered historical newspapers as a valuable source of research for several disciplines in the humanities. Librarians are now deeply engaged in the conservation and preservation of fragile holdings. As old newspapers are easily damaged by intensive reading libraries have for some time produced microfilms for the most important titles and offered them to their users instead of the original paper issues. But everybody knows that to use newspapers with the help of a microfilm reader or a reader printer is rather uncomfortable and tiring.

The digitization of newspapers opens new possibilities. The publication of retrospective digitized materials in the World Wide Web means that everybody can use the documents without restriction of time and space; they are open access and free of fees (CC-BY-SA). The viewer used by the library to present the digitized documents offers a wide range of functions, for example downloading (PDF), full text retrieval (OCR) or stepless zooming.

The Baden State Library (BLB) at Karlsruhe, Germany, which has been committed, since about five years, to the digitization of medieval manuscripts, old and precious books, and music scores, has now expanded its activities into the field of historical newspapers. The library has already published more than 950.000 newspaper pages in the World Wide Web.

The presentation describes the process of digitization and the online-presentation of newspapers within the Digital Collections of the library. Like some other large collections the Baden State Library offers various approaches to the digitized newspapers. Most important is the calendar overview which allows to choose between different years, months and days and the newspapers which appeared at a special date.

Keywords: Digitization, newspaper, cultural heritage, segmentation workflow, calendar presentation

Introduction

Historical newspapers belong to the printed cultural heritage. Unfortunately former generations of librarians did not pay very much attention to these ephemeral materials. For that reason many newspapers which appeared in Germany between the 17th and the 19th century did not come down to us. They perished because nobody felt responsible to preserve and to archive them properly.

Nowadays the situation has changed completely. Academics and students as well as private researchers have discovered historical newspapers as a valuable source of research for several disciplines in the humanities. Librarians are now deeply engaged in the conservation and preservation of fragile holdings. As old newspapers are easily damaged by intensive reading, libraries have for some time produced microfilms of the most important titles and offered them to their users instead of the original paper issues. But everybody knows that to use newspapers with the help of a microfilm reader or a reader printer is rather uncomfortable and tiring.
Further more it is no longer possible to give the paper volumes to our library users. The paper is highly fragile and many volumes are damaged as a result of intensive use and paper degradation due to acidification.

The digitization of newspapers opens new possibilities. The publication of retrospective digitized materials in the World Wide Web means that everybody can use the documents without restrictions of time and space. They are free of charge and available to everybody; like many other institutions we use the conditions by Creative Commons, for example Attribution-ShareAlike (CC BY-SA). The viewer implemented by the library to present the digitized documents offers a wide range of functions, for example downloading as printable PDF (Portable Document Format) from Adobe Systems, full-text indexing and retrieval by OCR (Optical Character Recognition) or stepless zooming.

Although the digitization department of the Baden State Library has gained extensive experience in scanning all types of books the digitization of newspapers meant an unknown challenge. On the one hand newspapers are a mass product which cannot be scanned by a small workshop consisting of two scanners and two permanent employees. On the other hand newspapers show some special features which we do not find with books and journals.

Digitization activities of the Baden State Library

In 2014 the Baden State Library at Karlsruhe, Germany, unexpectedly received considerable funding from a state program which aimed at the improvement of the conditions of education at the universities of Baden. Because the university libraries are part of the university they participated in the funding. The Baden State Library decided to use half of the money for the digitization of newspapers. The decision was based on the observation that students like to use newspapers as sources for their examination papers. In the OPAC of our libraries you may find titles like “The Russian Miracle in the mirror of the international press” (1965) or “Soviet Jewry in the mirror of the Yiddish press in Poland” (1975) or recently “The Luther anniversary and the year 1933 in the mirror of U.S. church press reports” (2013).

Since five years the Baden State Library has been committed to the digitization of medieval manuscripts, old and precious books, and music scores and now has expanded its activities into the field of historical newspapers. The library has already published innumerable books, journals and other printed materials on the history of Baden because it feels responsible for providing important source editions in either printed and digital form for researchers inside and outside the academic world and for the interested public. Newspapers as regional sources are part of the digitization policy of the Baden State Library.

The German Research Foundation (DFG) has been funding digitization projects for many years. It released a strategy paper on "Innovative Information Infrastructures for Research" called "Taking Digital Transformation to the next Level". But national solutions and consistent strategies for medieval manuscripts and for newspapers will be expected not before 2016. We could not wait so long!

We decided to implement a newspaper retro-digitization project on our own expense. At first we had a look at our collection of historical newspapers. It is a small collection because during the Second World War the library was hit by bombs and completely destroyed. The historical newspapers we possess, from the 18th century onwards, were transported into the main library building, sorted in the shelves, recorded and catalogued. For cataloguing the electronic version of the paper, we used the nationwide database for periodicals known as Zeitschriftendatenbank.

As we intended to scan more than 800.000 sheets, we outsourced the process of pure scanning to private service companies which were selected by means of an advertisement. We invited offers and to our surprise some companies offered a price of about 10 Cent per image. That is very low priced. The price included the transport of the newspaper volumes and the delivering of the image data on hard discs.

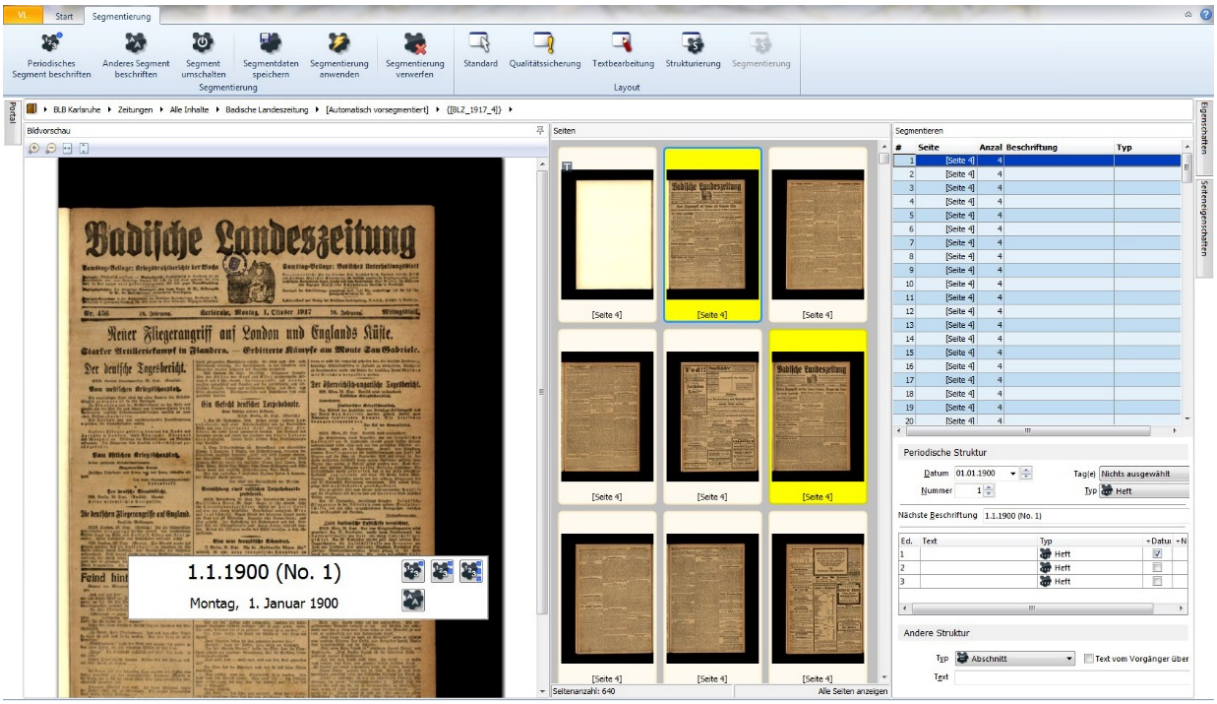
Workflow

The data delivered by some contracting companies were uploaded on the Visual Library server. Ever since the beginning of digitization, five years ago, the Baden State Library has been using the commercial German software solution Visual Library. It automatically controls the whole processing workflow, from the production of the scans and the merging of captured catalogue entries with the associated digital images up to the online presentation within the digital collections on the library homepage.

Moreover, URN (Uniform Resource Names) generation is possible, so that long-term addressing is assured. For providing best online visibility and searchability the OAI (Open Archives Initiative) interface allows the meta data exchange with other portals and library systems. Finally, the software produces display and download formats for the web presentation. JPEGs in different sizes and printable PDFs are derived from TIFF masters.

The structuring of the newspapers by date is supported by a special software module of Visual Library which was developed for the efficient segmentation of periodical publications. When the development was finished end of 2014 we had been the first user. Because a daily newspaper appears very regularly the software system knows the logical continuation of dates and it writes automatically the next day. The tool is based on the capacity of the system to analyze certain layout structures and calculated frequencies. So it knows when a new daily

issue begins. The system knows that Monday 1. October is followed by Tuesday 2. October. It writes exactly this date in the data field of the segmentation table. The operator has only to verify the proposal of the system and the syntax of the date.



Segmentation of newspaper issues

The operators were restricted to control the proposal of the system; they don't have to write the full date with day, month and year. The system is clever enough to know that a particular date was a Sunday or Monday or another day of the week in that year. A group of half a dozen student assistants led by a senior librarian managed the structuring.

If there is a striking difference, for example an enclosure between the newspaper pages, the operator has to intervene manually. Other circumstances may arouse by deviations of frequency or by irregular supplements.

Approximately the complete work of structuring newspaper volumes is done half automatically. It can be said that the segmentation software is a time-saving and efficient tool because it accelerates the structuring process significantly.

The in-house post-processing of bad or missing images and the Optical Character Recognition process were done by students too.

Optical Character Recognition

From the beginning the Baden State Library planned to use Optical Character Recognition. We purchased a lot of ABBYY FineReader licenses for black letter because the major part of the newspapers were printed with Gothic types (Fraktur). Only a relatively small part is printed with the more modern type Antiqua.

In the past all historical newspapers of the library were filmed. Because of the better results by the application of Optical Character Recognition we did not use the microfilm reels as digitization template, but the paper issues.

The full text indexing started with the newspaper volumes from the year 1914 to the year 1918.The anniversary of the outbreak of World War I was the guiding principle. We will continue with the following years from 1919 onwards.

At present the number of full text images amounts to 343.000 images or about 38 percent of the total number of newspaper images.

Results

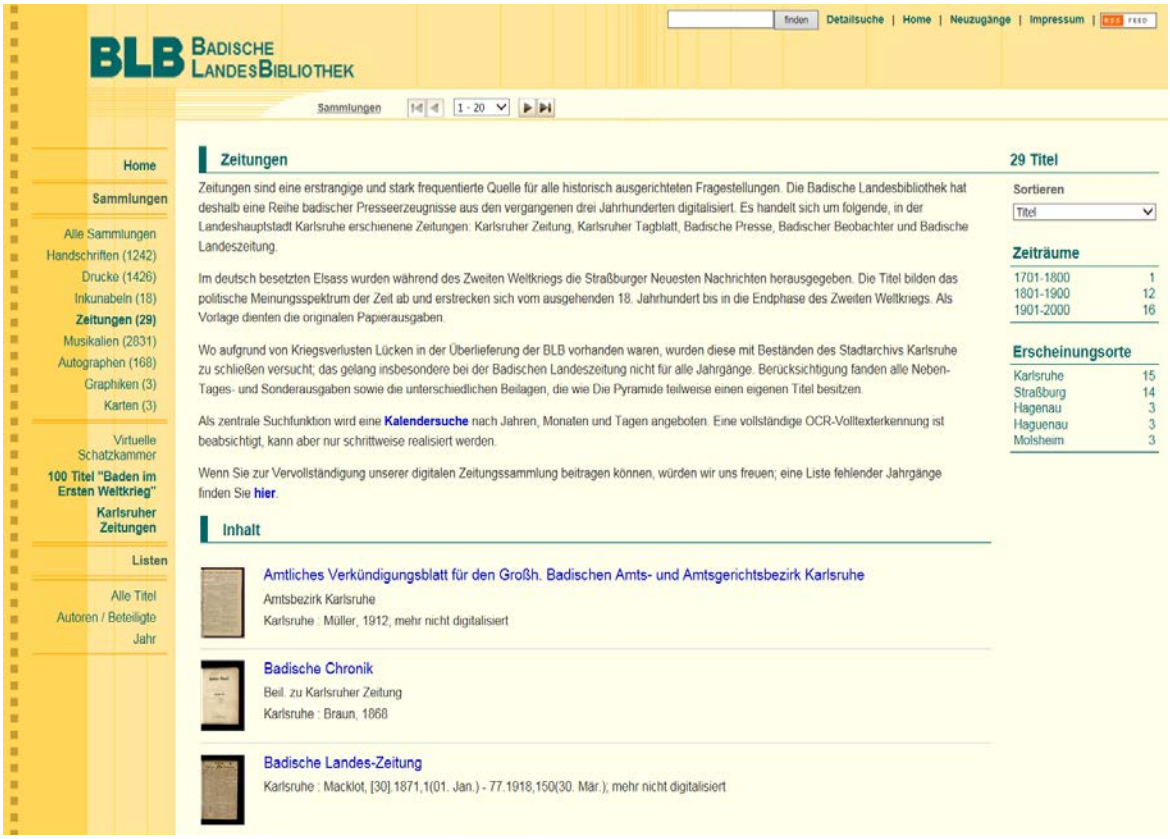
The Baden State Library published six newspapers from the beginning of the historical tradition to the end of the newspaper. Some papers were prohibited after Hitler's rise to power in 1933. In other cases the papers were made subject to a strict control by the Nazi Party. Most newspapers ceased to appear at the end of the Second World War, due to lack of paper.

In total we produced more than 900.000 images to present them online. With one exception the digitized newspapers were published in Karlsruhe, the former capital and the residence of the grand dukes of Baden, before the state became a democratic republic in 1918. In the Third Reich the National Socialist district leader (Gauleiter) became governor of German occupied Alsatia. At Strasburg the Germans edited the Straßburger Neueste Nachrichten as an official paper for their purpose. Nowadays it is an essential historical source for the research on Alsatia under German rule.

The release of the digitized newspapers was on 1th July 2015. It was accompanied by a press conference and by some articles on the successful completion of a huge project. The calendar presentation developed by our contracting software partner seemed to be convincing. The local press of Karlsruhe declared that historical newspapers are not only a treasure trove for the scientific community but for everybody interested in historical events.

Online presentation

The online presentation of newspapers is realized within the Digital Collections of the library. Visual Library supports all meta data transmission standards, postulated by the German Research Foundation, for displaying digitized objects with the so-called DFG-Viewer.

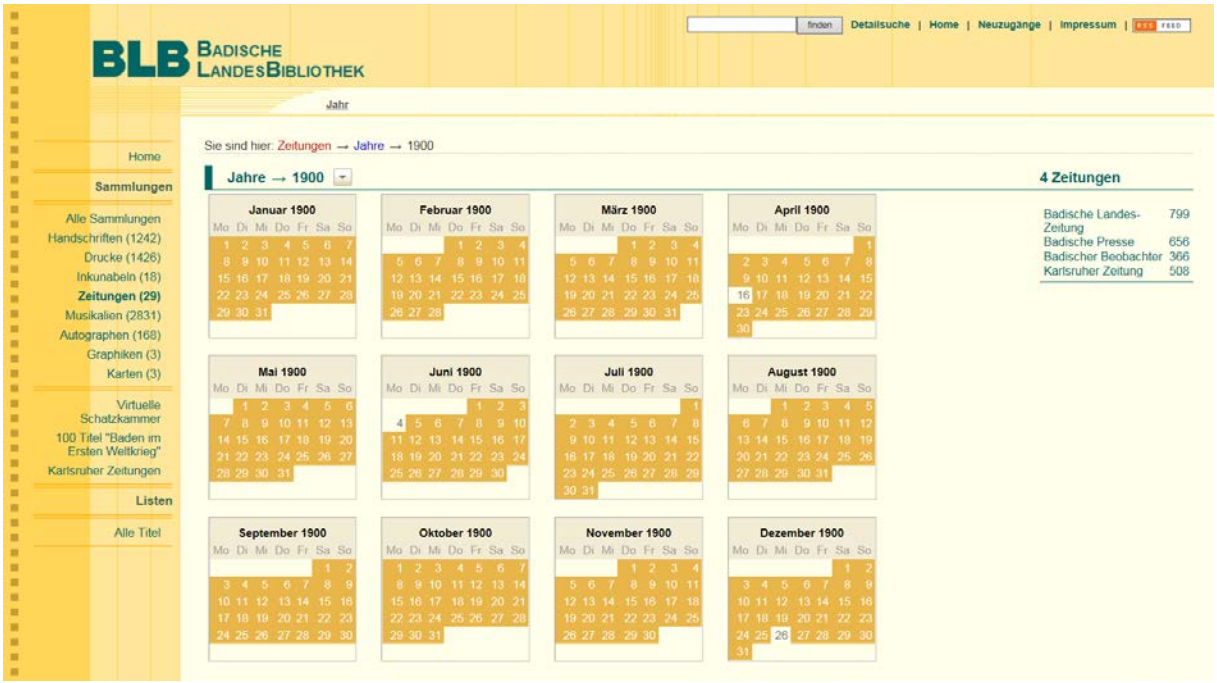


Newspapers in the digital collections

On the left side of the screen display the user may find the classification of all digitized materials which reflects – by the way – the digitization policy of the Baden State Library. It starts with medieval manuscripts, incunabula, autographs and music scores. The group of printed books is very comprehensive. Therefore the newspapers build an independent class.

Like other large libraries, the Baden State Library offers various approaches to the digitized resources. The first view which I would like to call the “classical view” is nothing special. The user enters a newspaper title and may choose between different dates and issues, for example morning or evening edition.

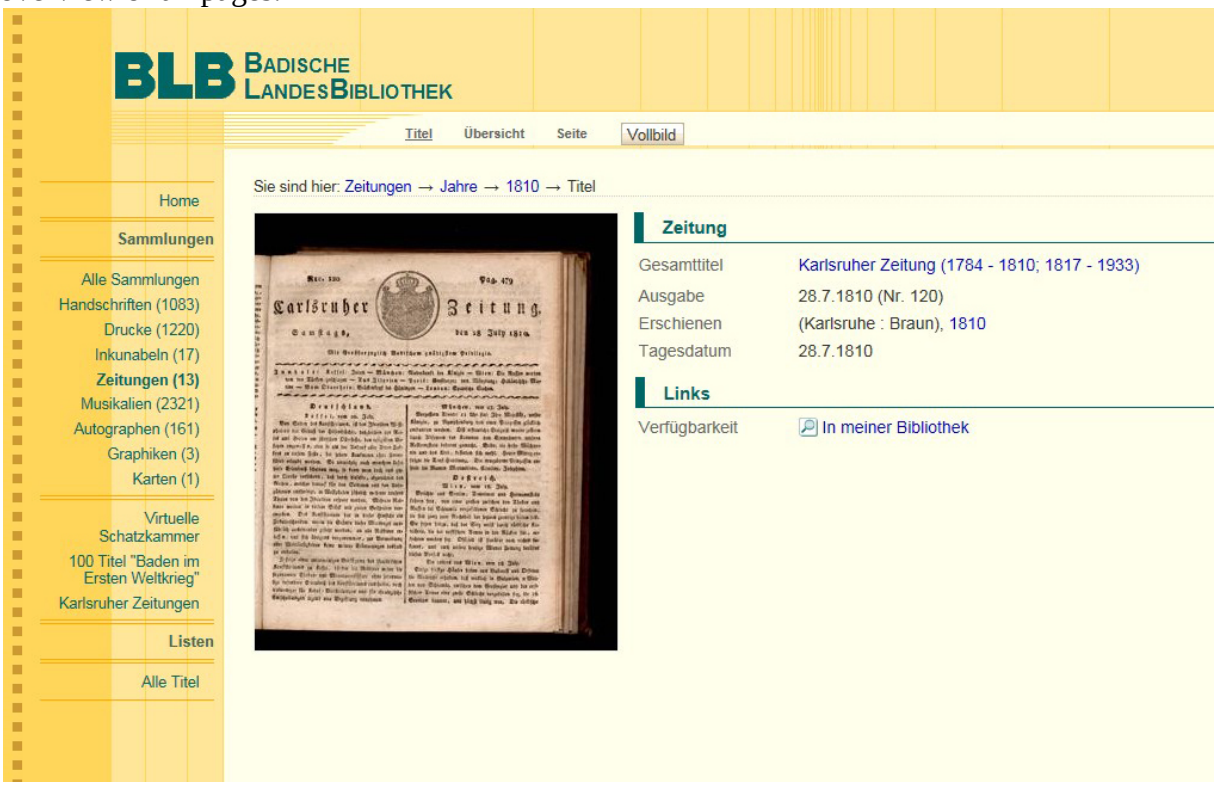
The calendar overview is most important. It allows to select between different years, months and days and the newspapers which appeared on a special date. The first step is to choose the year. Then the calendar opens and gives an overview of that year. The second step is to choose a month and a day. The colour “yellow” shows that for a special day one – or more – newspaper edition is available. The third step is to choose between the newspapers respectively the editions. Finally you arrive at the first page of the paper.



Presentation of the newspapers by calendar

The navigation within the electronic newspapers offers the normal possibilities and functions. Because the relation between the computer display and the newspaper format is very unfavorable, the user will be allowed to zoom.

The Visual Library viewer offers a thumbnail presentation of the first page of any captured document. The navigation bar allows scrolling forward and backward and may present an overview of all pages.



Thumbnail and bibliographical metadata of a newspaper issue

On the right side of the screen display the user will be offered the possibility to limit his search by facets, for example periods of time (centuries) or places of publication. He may choose between several sorting aspects. Concerning newspapers only the limitation by time is reasonable.

Distribution

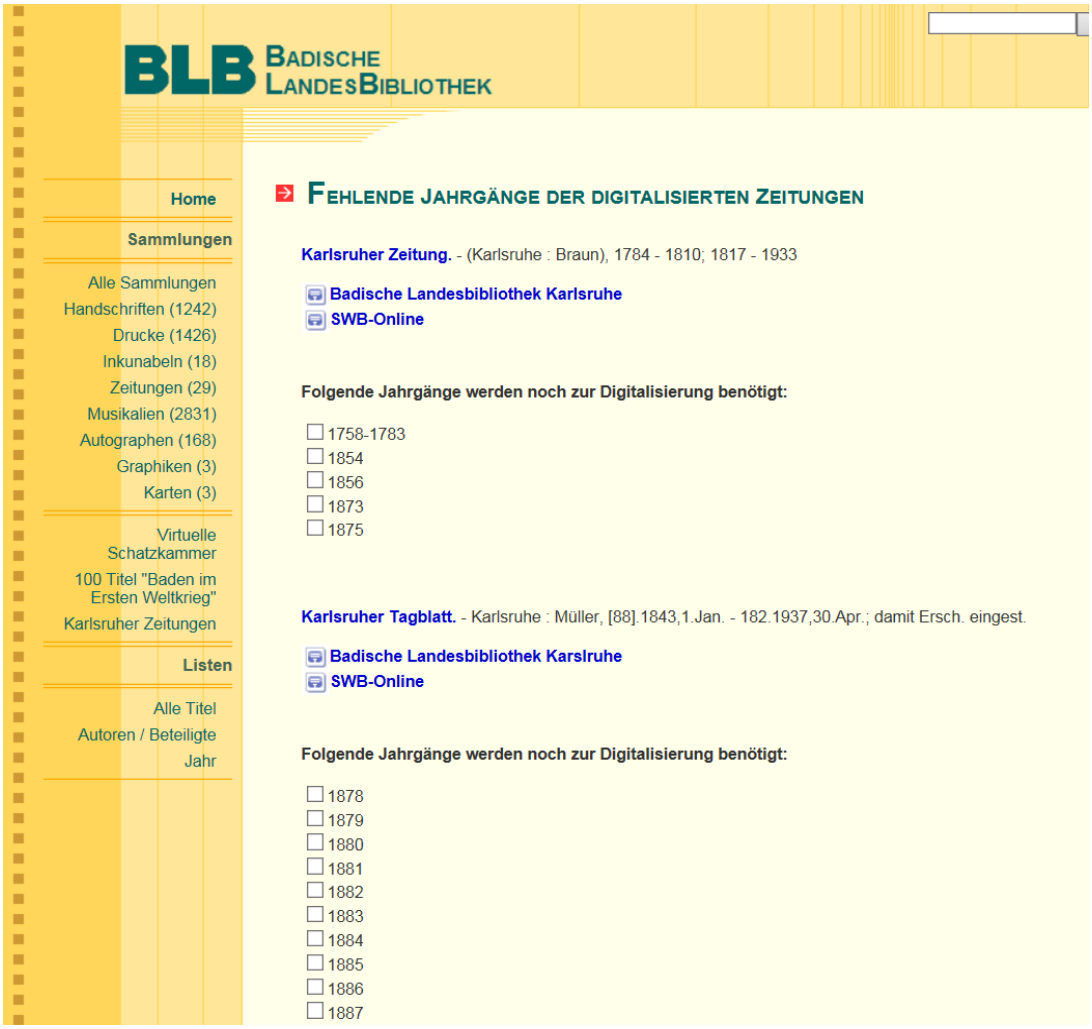
All documents captured digital will not only be presented within the digital collections of the library but exported to other content aggregators. At first I have to mention europeana as the European Digital Library and the German Digital Library. Our digitization software system contains an OAI-interface for the harvesting of our metadata including bibliographical descriptions and thumbnails by using the widespread data formats METS and MODS.

Once a year our digital data became harvested by the regional information system LEO-BW which offers free access to high-quality information from and about Baden-Württemberg. In LEO-BW you can find a variety of regional-related information, multimedia and literature from different resources all in one search.

If there will be build a national newspaper portal in the future the Baden State Library will gladly share its electronic newspapers with other collections.

Closing gaps and further challenges

On the homepage of the digital collections we request our colleagues in other libraries, archives and similar institutions to help us close the gaps in our tradition: “If you could help completing our range of electronic newspapers in the digital collections we would be very happy. Missing volumes you find **here”**. Then the list of the missing volumes follows.



Appeal to help closing gaps

But there are many other challenges and difficulties. Some of them can be traced back to the cooperation between the library and the contracting companies. Some private services had no experience with the digitization of cultural heritages; they worked for insurances or administrations. The scan equipment was unqualified, for example, for large formats, nor was the scan technology state of the art, for example concerning the colour profile.

The exact designation of all data files and the identification of the pages by using a correct header is very important. Another problem was the scanning of newspaper supplements with a deviating paper format. In these cases the scan operator had to use an underlay, mostly a grey paper sheet.

Other typical difficulties were caused by the material features. Sometime there are contradictions between the date printed and the date calculated by the Visual Library segmentation tool. In former times the newspapers were manually printed, so mistakes happened.

Many papers changed their name in the course of time. For example, from 1811 to 1816 the Karlsruher Zeitung was called Großherzoglich-Badische Staatszeitung, but basically remained the same paper.

In former days many papers had no page numbers. So we had to give up the idea of taking the page number into account while structuring the newspapers.

It turned out to be sufficient to distinguish main editions and auxiliary or additional editions, for example for another district or another town. Some newspapers appeared twice or even three times a day.

The scanning, cataloguing and segmentation of newspaper supplements was very laborious. They may have been published only once, for example with an announcement or proclamation by the government. But they may also have been published regularly over many years, for example as a literary supplement. Sometimes they are called simply “supplement”, sometimes they have a title of their own. The frequency ranges from weekly to sporadic. Every newspaper and every supplement needs a new structure to control efficiently and automatically the process of segmentation.

Follow-up project

The large digitization project ended in 2015. At present we continue the scanning of newspapers in a project which is funded by the Kulturstiftung Baden-Württemberg. To complete the political range of our newspaper offers we now digitize the daily paper of the National Socialist Party of Baden called “The Führer”. This paper is an indispensable source for studying the era of the Third Reich in southwest Germany including the years of the Second World War. It was published between 1927 and 1944 and had additional editions for several districts of Baden. Once again we will be able to fill gaps with the help of the holdings of the municipal archive of Karlsruhe.

The workflow will be the same as last year: Scanning the newspaper pages was outsourced to contracting companies. They produced about 65.000 images sent to us on hard discs. They had been included into our digitization system Visual Library. Structuring or segmentation of the papers as well as OCR-processing is managed in-house by student assistants. The release date will be in the autumn of this year. Together with another small project we optimistically hope to reach one million newspaper images at the end of 2016.