News Media Section

Retrodigitization and Electronic Representation of the "Vorwärts" A Workshop-Report

Olaf Guercke
Bibliothek, Friedrich-Ebert-Stiftung, Bonn, Deutschland
olaf.guercke@fes.de

Abstract:

During the time of the German Kaiserreich and the Weimarer Republik the "Vorwärts – Berliner Volksblatt" has been the most important periodical of the organized social democracy in Germany. Today it is an essential historical source for German political, social and cultural life within that timeframe. In January 2015 the Library of the Friedrich-Ebert-Stiftung started the digitization of the "Vorwärts", aiming to make it accessible for research via a web-presentation. The corpus contains about 200.000 newspaper-pages in about 19.000 issues, which were published from 1876 until 1878 and 1891 until 1933. While currently the digitization from the paper-original proceeds, the web-presentation gets prepared in cooperation with the software-provider ImageWare Components GmbH.

The submitted lecture is a workshop-report, which responds to the multiple aspects of this project from the beginnings to the first presentable results. It starts with the project-planning, focusing on the pros and cons of in-house digitization. It then describes the challenges of the complex internal structure of a newspaper-corpus, trying to frame standards for the quality of the images and for the workflow of the scan-process. The focal point of the lecture is nevertheless the discussion of questions connected to the web-presentation of this complex source. In this context the demands of the visualization of newspaper-pages and the requirements of retrieval-capabilities are considered from a user's point of view. Moreover the lecture describes the importance of Optical Character Recognition and the challenges of OCR with German-gothic font. On top of this, copyright issues concerning the web presentation of a historical newspaper are addressed.

The aim of the lecture is, to encourage an open discussion about the several aspects of a comprehensive newspaper digitization-project.

Keywords: Vorwärts, newspaper, digitization, OCR, social democracy

Inhalt

Introduction

The aim of the project "Vorwärts bis 1933", implemented by the Library of the Friedrich-Ebert-Stiftung since January 2015, is the complete digitization of the "Vorwärts - Berliner Volksblatt" from its foundation 1876 until the ban by the National Socialists 1933. The digitized "Vorwärts" then shall be provided to the public through a web presentation, which is searchable on base of an OCR-generated index. A total of about 200.000 newspaper-pages in some 19.000 issues will be digitized in high quality from the paper-originals. The text at hand is a workshop-report, in which the project is described in all its current aspects and perspectives of future development, so that other practitioners in the area of newspaper-digitization can share our experiences.

It should be prepended, that the project is currently at a crucial stage. So far, the work has mainly been focused on producing high-quality images, while the questions regarding the web-presentation were postponed. The workflow was therefore designed under the premise that - in terms of image-quality, data-formats and filename structure - the output generated in this phase would have to be usable within a broad variety of possible presentation-scenarios. Recently, with the aim of developing and realizing the presentation, a research project was initiated together with the company ImageWare Components GmbH[1]. Purpose of this project is, in short, to create a researchable representation of the "Vorwärts" in the web on basis of the presentation-software MyBib eL[2] and thereby cooperatively sharing expertize with the aim to improve the software with regard to the requirements of a huge newspaper digitization. The web presentation created in this project shall be hosted by the Hochschulbibliothekszentrum des Landes Nordrhein-Westfalen (hbz).

In the light of this decision, the "Vorwärts" project is currently in transition. According to this development, the present paper is divided into a chapter focused on the various aspects of the work done so far in the project, followed by a chapter focused on the first experiences, results and prospects with regard to the presentation. Nevertheless, before addressing these subjects an explanation regarding the motivation of the Vorwärts-digitization shall be given.

1. Why do we digitize the "Vorwärts"

At the beginning of every large digitization project the question arises, in what extend such an undertaking makes sense for the executing library. Which groups of library clients might be interested in the result? Is the source to be digitized really relevant for research? Does the project correspond to the objectives, which the library has set in terms of its profile? In short: Is the public and institutional demand for the results large enough to justify the expenditure of manpower and financial resources implicated by the project. The following text is an attempt to answer these questions with regard to the "Vorwärts bis 1933" as a digitization-project of the Library of the Friedrich-Ebert-Stiftung.

[1] Hereinafter referred to as: "IWC". Cf.: http://www.imageware.de/ [17.03.2016]

[2] Cf..: http://www.imageware.de/produkte/mybib-el/ [17.03.2016]. The acronym "eL" refers to the German term "elektronischer Lesesaal".

Ever since its establishment the "Vorwärts" was the central press organ of German Social Democracy. Its first period of appearance went from 1876 to 1878, when it was banned due to the German Kaiserreichs "law against the dangerous activities of social democracy"[3]. Newly founded in 1891 it was carried forward until the renewed ban by the Nazis on 02/28/1933. During the 3rd Reich it continued to exist in exile until 1940 under the name of "Neuer Vorwärts". The history of the “Vorwärts” in the Federal Republic of Germany ranges from the founding in 1948 to the present day. Within the publication period relevant to this digitization project, the 19th century-“Vorwärts” started with four issues per week and a circulation of 12,000 copies. From 1900 it appeared daily and from 1920 two times a day, with the circulation steadily increasing to finally about 300,000 copies in 1932.[4]

Regarding the content and it’s general aims the “Vorwärts” tried to be both the official mouthpiece of the SPD executive committee and an ordinary Berlin local paper. So, as a "central organ" of an organized political movement, which was very influential in the German Kaiserreich and the Weimar Republic, the Vorwärts is a highly relevant source for historical research focusing on social democracy, the labor movement or related forces and their impact on political disputes and developments in this timeframe. At the same time the Vorwärts is a treasure chest for cultural research, which relates to the social democratic milieu in this period. The growing interest in the digitized “Vorwärts”, which is currently generated by the ongoing project, manifests itself in many user requests, that come both from the academic sphere (e.g.: conception of a seminar to the Weimar Republic, resources for Bachelor- and Master-Thesis, Doctorate, research-projects) as well as from the general public (Genealogy, literary projects, exhibitions). Beyond this, the editors of today's “Vorwärts” are highly interested in getting access to their own history.
At the moment, the “Vorwärts” can be researched almost exclusively via the microfilm edition.[5] Due to this fact this important source is often not taken into account in research projects for which it would be highly relevant, since the effort of finding suitable texts is simply too large. In addition, given the size of the corpus, a thematically oriented search, which extends over several volumes is de facto impossible. A digital presentation of the “Vorwärts” with OCR-based keyword-search would therefore make an important historical source, which at the moment can only be used with great expenditure of time, conveniently accessible for research and the public.
The Library of the Friedrich-Ebert-Stiftung is the world's most important library specialized on the subject of German labor movement. It almost speaks for itself that the digitization of a source with such central importance for the social democratic part of the German labor movement correspondents to the narrowest mission of such a library. Not only will the project meet the library’s mission as a highly specialized provider of information for teaching, research and public. It also will enable today's social democracy for taking a detailed look into its own history.

[3] Cf.: http://www.documentarchiv.de/ksr/soz_ges.html [17.03.2016]
[4] Cf.. Walter G. Olischewski: Eine Zeitung schreibt Geschichte. 100 Jahre Vorwärts. In: Neue Gesellschaft, 23(1976), H. 10, S. 788-791.
[5] The paper-originals are, in general, very sensitive, partly damaged and therefore banned for public usage.

2. What has been done so far

2.1. The scan-process: digitization as an in-house-project

In the past there have been approaches in the Library of the Friedrich-Ebert-Stiftung to perform the Vorwärts-digitization using external service providers. For the current project this approach has been altered for various reasons. One important point was, that the Library was reluctant to rely on an external service provider in handling the highly complex, sensitive and sometimes strongly damaged newspaper corpus. Also, it wanted to have control over the production process at any given time. Against this background, the decision to arrange for an in-house-digitization came to a fairly early stage in the project. In the following, challenges and requirements of such an in-house digitization shall be outlined with the aim of providing decision support for other digitization projects.

Challenge 1: Suitable devices

Due to reasons of proper conservation and to aesthetical aspects, the achieved reproduction of the original newspaper pages should as authentically as possible. Therefore a relatively inexpensive and time-saving digitization based on the microfilm edition of the "Vorwärts" was ruled right out from the beginning. With regard to the partly damaged and sensitive documents the use of a scanner with page feed was also excluded. Therefore a large overhead-scanner with the following capacities had to be acquired for the project:

- Format DIN A1
- Book cradle, suitable for heavy and up to 15 cm thick folios
- Production of high quality color scans
- User-friendly and capable software
- Fast scanning
- Workstation with high computing power able to process large amounts of data
- Universal applicability in- and outside the project

Devices that comply with these requirements, are offered by several German and international manufacturers. Before a new purchase an intensive market survey[6] is recommended, also a public call for tender is required in most cases. For the "Vorwärts"-project a scanner of the type "Zeutschel OS 12000"[7] was finally purchased. It should be noted here that both the cost for the device and the workload for market survey and tendering should not be underestimated in advance.

[6] A suitable alternative to the demonstration of devices by the manufacturers may be the consultation of practitioners in other libraries who often are willing to share their experiences and show their devices.
[7] Cf.: https://www.zeutschel.de/de/produkte/scanner/farbscanner/os-12000-din-a1.html. [17.03.2016]

Challenge 2: Manpower

A significant disadvantage of overhead scanners lies in the relative slowness of the production process progress. Each newspaper-page has to be scanned individually in the correct order and thereby must be treated carefully. Each issue has to be augmented with metadata. Post-processing, quality control and various sideline-tasks take at least as much time as the scanning itself. In order to obtain a realistic estimation of the projects duration, results of extrapolations based on the pure scanning speed should at least be doubled.

Regarding the "Vorwärts" digitization, one librarian (B.A., 25 hours per week) and two scan-operators (with a total of 23 hours per week) work exclusively for the project. So far, a production rate of approximately 6,000 images per month was achieved incl. post-processing, quality control, upload and project-management. So obviously a long term approach and a lot of patience is needed. Of course one also needs dedicated, responsible staff, be it the librarians or also the scan-operators, who above all have to bring along the ability to perform a basically quite monotonous activity over a long time with great care.

Challenge 3: Work environment

The scanning process should take place in a space which can be shielded from direct sunlight. It should be considered that the scan-operator performs physical work under high concentration and should be supported therein by an appropriate working environment. Ergonomic devices such as height-adjustable tables are useful in any case and should be considered in financial planning. As heavy and bulky scanners might be used it is also necessary to consider, whether ceiling height and building-statics allow the implementation of such device.

Conclusion

An in-house newspaper digitization encompassing the scope of the "Vorwärts"-project is a costly and labor-intensive task which requires a lot of commitment over a longer period of time. In early stage of planning the pros and cons of outsourcing the scan-process vs. in-house digitization therefore should be carefully assessed in consideration of the projects objectives.

2.2. The internal structure of a newspaper-corpus as a challenge

To represent a newspaper-corpus via the web in a way, that allows the user an intuitive experience of its wealth of content and structure is a major challenge. It should be aimed at a form of presentation, in which any user can see at any time, in which issue of which volume he currently is browsing and in which he can switch easily to day and year overviews. Since the "Vorwärts" project was launched without the support of a workflow system above the scanner-software, and since it was not yet clear how the presentation should be ultimately developed, the first important task was to capture the structural data of the corpus during the scanning process in a preferably general way. This was achieved through clear rules regarding the allocation of the file names for the images and the definition of a logical folder structure. The

overall structure of this concept is briefly described in the following paragraph. Additionally the file naming scheme designed to illustrate this structure, is discussed:

As mentioned above the “Vorwärts” started with four issues per week, later moved on to daily appearance and finally was published two times a day. The issues are numbered sequentially within each volume. In the period of two issues per day, the morning and evening editions are additionally counted with numbers starting with an “A” (morning) and “B” (evening), so that in this timeframe every issue has a general number and an additional number as morning/evening edition[8].
During the appearance of the “Vorwärts”, numerous supplements[9] can be identified, which are considered as independent newspapers because of their independent numbering and should therefore be presented independently in the project. There are also numerous extra- and special-editions, variations of one and the same issue with different content as well as variations in the number assignment.[10]

The scanning software OS 12[11] allows a configuration that enables the creation of filenames for the produced images during the scanning process according to the metadata entered by the scan operator. Also it automatically creates the folder structure to which the images are exported. The following metadata gets identified and inserted by the operator during the scanning process:

- Acronyms for newspaper- or supplement titles. E.g. "vw" for "Vorwärts", "fs" for "Frauenstimme"
- Publication date: year-month-day, entry per calendar function
- Number of Issue
- If necessary: Evening- or morning-edition number

With this information the software produces a unique file-name for each image according to the following scheme: "Newspaper-year-month-day-number_ additional number if necessary-number of page".
A file named like e.g. "vw-1923-12-25-325_b165-004" could, on the basis of this structure, always be identified as "page 4 of issue 325, evening edition B 165 of the “Vorwärts”, published on the 25/12/1923". For each of the special cases described above separate arrangements have been established.

[8] E.g. the issue nr “50; A25“ would be the 50th issue and 25th morning-issue of its volume. It would be followed by the issue nr. „51; B25“ (51th issue, 25th evening-issue)
[9] Supplements of the „Vorwärts are upon others: „Jugend-Vorwärts“, „Frauenstimme“, „Blick in die Bücherwelt“, „Sonntag“, „Heimwelt“, „Unterhaltungsblatt des Vorwärts“, „Volk und Zeit“, „Die Wählerin“, „Die Neue Welt“.
[10] Another “special feature” is, that between the 15/02/1928 and the 30/09/1932 the Berlin evening-edition of the “Vorwärts” was called "Der Abend", while the “Vorwärts” outside Berlin appeared as "Vorwärts-Spätausgabe” Besides the head on the front page both versions probably had a similar content. This problem is not yet solved in view of the presentation.
[11] Cf.:https://www.zeutschel.de/de/produkte/capturing-software/os-12.html. [17.03.2016]

The files are then automatically stored in a folder system with four layers (newspaper title / year / date / number), so that the images of each issue are packed in one given subfolder.
This outcome of the preliminary work currently has to prove its usefulness for the workflow-development using the presentation software MyBib eL. It has already proved its helpfulness for the management of several user requests for images from the already digitized editions of the "Vorwärts".

2.3. File-formats and quality standards

The images produced in the project are 24-bit color-scans with a resolution of 300 dpi. They are exported as uncompressed TIFF files, which complies with the standard set in the DFG guidelines for digitization projects. [12] In addition to these master files JPEG files of both 80% and 100% quality as well as PDFs of entire editions are produced. Master files and derivatives are automatically produced by the software during the scanning-process and then stored in different main folders, each within the file structure described above.

A problem in this context may be the huge amount of memory space which is needed to store the TIFF master files. On the one hand, the TIFF-format is essential as it is the lossless and sustainable standard data format, which is required for reliable long-term archiving. On the other hand a single DIN-A3 color image at a resolution of 300 dpi needs about 55 MB of memory space, which means that for the "Vorwärts" project a total of 11 TB will be needed. The sustainable storage of such masses of data is a challenge.

Quality control takes place on the basis of the images stored in the folder system. The assessment includes the correct allocation of images to each issue, the completeness of the pages, the correct alignment and complete display of the text block and the correctness of the scan parameters.

At this point, the lack of a coherent workflow system in the project becomes apparent. It is relatively complicated and time-consuming to rescan and replace incorrectly scanned images. It also would be desirable if minor editing functions, such as deskewing would be supported by the software, in which the quality control is executed. Hopefully the ongoing development of the project will bring along significant benefits in these cases.[13]

2.4. Previous workflow

Having highlighted some important specific aspects, the workflow shall now be described briefly as a whole.

[12] Cf.: Deutsche Forschungsgemeinschaft: DFG Praxisregeln „Digitalisierung" – DFG-Vordruck 12.151 – 02/13. S. 15-17. [17.03.2016]
[13] See also chapter 3.2. of the present paper.

- The Libraries paper-copy of “Vorwärts” is bound in large folios of up to 1500 pages. The first step of the digitization is the complete examination of these folios including the documentation of number, date of publication, number of pages and conservation status in an excel-spreadsheet.
- It has been found that because of a strong unevenness of the templates and partly solid text loss due to a very tight binding it is impossible to obtain high-quality scans from the folios. Therefore these folios are carefully opened by a bookbinder. For every quarter he pages are stored in acid free archive boxes.
- Now the pages are scanned one by one, whereby the scan operator types in the metadata and controls the number of pages on the basis of the previously created excel spreadsheet. He also records any loss of text and fixes damaged pages with special tape.
- After that, the quality control and the upload of the scans take place.
- Missing issues and heavily damaged pages with serious text loss are noted and ordered at certain intervals in the Universitäts- und Landesbibliothek Bonn.[14]
- All these operations proceed parallelly. Current status of the project is the following:
  Digitized are the volumes 1925-1933 (a total of 50,000 Images)
  Opened and liberated from the folio-binding are the volumes 1919-1933
  Checked and recorded in spreadsheets are the volumes 1914-1933

After describing the actual state of the project and the various aspects of the production of images from newspaper pages, the following chapter will be dealing with the provision of these images for the public in a web-presentation.

3. What should be done from now on

3.1. Formulation of objectives: What features should the web presentation have?

In the first place, the web presentation of a newspaper needs a clearly arranged and intuitively manageable viewer which meets the requirements of the originals’ large newspaper format, is usable both via desktop as well as on tablets or smartphones and allows a comfortable reading. Beyond these basic requirements some preferable features of the presentation shall be pointed out in the following.
As the project grows, inquiries and request of clients are increasing and of course are processed. Contemplating about the requirements of the presentation it makes sense to think from a client’s point of view. Who are these clients and what kind of presentation do they need?

[14] The newspaper-department of the ULB Bonn stores a relatively complete paper copy of the “Vorwärts” and has kindly declared its willingness to to support the Vorwärts-project with filling in missing issues. For this, the author gratefully wants to express his appreciation.

Two major groups of clients who have a significant interest in the "Vorwärts" are researchers and students, usually from the field of historical sciences, who require the source as a basis for seminar preparations, theses, dissertations and other research projects. These groups together represent about 2/3 of the requests currently reaching the project. They usually have very specific requests. The following real-life example illustrates this:

For the preparation and design of a multiple-semester seminar at the faculty of history his university, a lecturer needs access to reports from the "Vorwärts" about left-wing labor-unrest in Halle in the years of 1919, 1920 and 1921.
The lecturer provides a detailed list with events of particular interest for the seminar.
To find the relevant articles which are hidden in about 5.000 newspaper pages, the following combination of search entry points would be useful[15]:

- An OCR-based keyword-search which ideally can be narrowed down to the timeframe which is of interest (Here: spring 1919, spring 1920 and spring 1921).
- A possibility to browse via volume and publication date to take a closer look at the issues which were published around the dates of the respective events.

Each search entry point needs to be complemented by the other. The OCR search revaluates the search opportunities considerably by enabling the client to scan large text masses under very specific issues that would otherwise require significant and unaffordable labor[16]. However, such a search is, especially when dealing with templates in Gothic script, never completely reliable, which is one reason, why the classic access via browsing is also necessary. The browsing entry allows the investigation of smaller timeframes in their entirety and the verification of OCR-based search results[17]. Additionally, it is not to be underestimated that the browsing entry enables the client to get a quick overview of the entire project in its structure and size which provides a lot more guidance than a keyword-search on its own.

Apart from these possibilities of search entries the presentation should ensure consistent traceability through permanent links, which are in the best case implemented at a page level. This could be solved by using Uniform Resource Names (URNs) in a defined namespace.
Another important point is the copyright issue, which almost inevitably occurs in large newspaper digitization projects for content published in the 20th century. In connection with these

[15] Since no web presentation is available at this stage, the request requires a considerable amount of work for client and librarian: The client must provide a detailed list of events, the librarian has to search for matching articles in the issues of the indicated timeframe and then produce and send the scans. The client's task now is, to separate the the proper hits from the uninteresting ones. He has no chance to control, whether the research of the librarian is complete or not.
[16] E.g.: when it comes to look for texts of certain authors, who might have written over a longer period in the "Vorwärts". Here, without OCR the detailed checking of a single volume á 6000 pages would take several days of hard work.
[17] E.g.: when it comes to examine the entire coverage around a particular event. A possible question would be: "How was the Kapp-Putsch reported in the various resorts of the "Vorwärts?""

issues a possibility to block content at the page level would be highly useful[18]. Due to the size of the body, such a possibility should be usable in an absolutely straightforward way and with a manageable workload.

3.2. What has been implemented so far

- Viewer / display
  Regarding the viewer MyBib eL already brings along some functionalities which meet the requirements of the representation of newspaper pages in the web. The individual pages can be scaled continuously and variably to a comfortably readable size. The interface for this feature is intuitively usable. The single issues are presented with an information text consisting of the newspaper title ("Vorwärts" or one of the supplements), volume, release date, number and a field for further information. From this display the client can switch to a hit list with search-results (provided that a keyword search was performed before) to the main search entry. The hit list provides an excerpt of OCR-generated text for each hit. Within the images, the searched keywords are highlighted with a colored background.
  All this provides a very good basis for the further development of the web-presentation.

- OCR
  With regard to the OCR processing of the entire body a major challenge is the predominantly Gothic font for most parts of the newspaper. Moreover, due to the huge amount of material to be processed the time factor has to be taken into account. These two factors played a major role in our decision-making for an OCR-engine.
  Since in the field of commercial OCR engines for Gothic font there is currently no serious alternative to the products of the company ABBYY[19] and since looking at free

[18] A blocking option at an issue-level is in general not suitable for newspaper-digitization, since it does not make sense to lock an entire newspaper-issue with otherwise unproblematic content due to a single article with copyright on it.

[19] Cf.: Günther Mühlberger: Digitalisierung historischer Zeitungen aus dem Blickwinkel der automatisierten Text- und Strukturerkennung (OCR). In: ZfBB 58 (2011), 1, S. 13: „*Im Gegensatz zum Markt für Scanner, auf dem sich eine große Anzahl von Firmen tummelt, ist bei der OCR Software in den letzten 10–15 Jahren ein starker Konzentrationsprozess zu beobachten. Im Wesentlichen dominieren nur noch wenige Firmen die Szene: Nuance mit OmniPage, ABBYY mit FineReader und IRISmit Readiris. ABBYY wiederum ist die einzige Firma, die sich seit mittlerweile zehn Jahren auch um historische Schriften und Dokumente bemüht und mit der Frakturerkennung ein Alleinstellungsmerkmal besitzt.*
Einigen Wind aufgewirbelt hat auch die Ankündigung von Google im Jahr 2006, dass man eine freie OCR Engine (Tesseract) zur Verfügung stellen werde.1Unter Leitung von Thomas Breuel vom Deutschen Forschungszentrum für Künstliche Intelligenz (DFKI) in Kaiserslautern wird nunmehr seit 2007 das Programm OCRopus entwickelt, das diesen Ansatz aufgreift und ebenfalls unter einer Open Source Lizenz zugänglich gemacht wird.11 Allerdings befindet sich OCRopus noch in einem sehr frühen Stadium, sodass eine seriöse Abschätzung noch nicht getroffen werden kann, ob damit eine ernsthafte Alternative zu den kommerziellen Softwarepaketen gegeben sein wird. Immerhin handelt es sich sowohl bei Nuance als auch bei ABBYY um weltweit agierende Unternehmen, die auch Kyrillisch, Chinesisch, Hebräisch, Arabisch und knapp 200 weitere Schriften unterstützen. Für ein konkretes Zeitungsdigitalisierungsprojekt im mitteleuropäischen Raum bedeutet dies, dass man kaum eine ernsthafte Alternative zur Frakturerkennung der ABBYY Software finden wird und sich daher mit den Möglichkeiten und Tücken dieses Programms etwas intensiver auseinander setzen sollte – auch um die Angaben von Dienstleistern besser nachprüfen zu können."

software the engine "Tesseract" seemed to be the only promising solution, both engines were tested in the project. For the test several sample pages (JPEG 80%) were processed using Tesseract and ABBYY. We then analyzed the XML-files produced by the software containing the recognized words and their coordinates on the image. In addition, the results were checked by performing keyword-searches and counting the hits in the MyBib eL-test-presentation.
The results of these tests were clear. The queries that were performed with ABBYY OCR regularly generated about twice the number of correct hits compared with the results of searches with the Tesseract engine. Analyzing the XML-files a much better word recognition for ABBYY-OCR was evident even at first glance. Another crucial factor was the time. While it took the ABBYY engine about 11 seconds to process one page, including the generation of XML-files, the Tesseract engine needed about 8.5 minutes (500 seconds) per page. The processing of the whole text-body in a reasonable timeframe would be practically impossible.
Based on these two factors it was a clear decision that the ABBYY-engine will be used in the "Vorwärts" project. More extensive test series were initially postponed but would be certainly useful in context with the further development of the research-project.

- Workflow
  So far the workflow of OCR processing and metadata assignment has only been outlined in the project and still has to prove itself in practice. It soon will become apparent up to which grade it will open up possibilities to improve and speed up the scan production workflow. The development of the workflow in whole will be an interesting subject for research within the project and will be precisely described in further reports about the "Vorwärts" digitization.

3.3. What should still be applied – Possibilities and challenges for the future

The following section describes features that might be useful additions to the MyBib eL functions in terms of newspaper digitization projects. These could be taken into consideration as medium-term implementations of the project.

- Configuration of the "advanced search"
  An important point deriving requirements formulated as expectations of potential clients is the configuration of the already available "advanced search" to narrow results quantities and to support the accuracy of search results.
  By now the MyBib eL supports searches for phrases, truncation of the search terms and search for multiple words at the same time.

These functions provide a solid foundation and should be explained to the clients in detail on the search-entry-site. Apart from that the following features would be really useful to enhance the overall accessibility:

- The ability to narrow the search to specific time periods:
  Since inquiries of researchers from historical sciences generally are focused on their subject in connection with certain periods of time this feature would be of really great value for those clients. As pointed out in chapter 3.1 the researcher looking for labor unrest in the city of Halle in springtime 1919, 1920 and 1921 could use the keyword "Halle" and focus his search on the periods of interest to generate meaningful results for his case. A search for "Halle" performed on the complete stock would certainly generate far too many hits.
- It would also be useful to allow combined searches with the Boolean operators “and”, "or" and "and not". Furthermore, the possibility of a fuzzy search would make sense.

- Search entry "browsing"
  In the MyBib eL the access option "Browse" currently provides a linear string of issues, which can be sorted by release date or media number and then are shown in sections of up to 50 outputs at once. Since the project includes a total of 19,000 issues, this form of display is not sufficient for an easy access to the entire body via browsing. It would be desirable that a list of volumes is shown at the first browsing level and that on subordinated levels up to 650 issues per volume would be made available on basis of a calendar function. At the lowest level, the different issues of any given day (morning and evening edition, supplements) may be displayed. This approach with three browsing-levels would enable the user to comfortably get access to a particular issue and to detect the extent and structure of the whole project.

- Option to block content for the web-frontend at a page level
  As already pointed out in Chapter 3.1 this would be a highly useful feature for many digitization projects of newspapers. In its classical function as in-house application, MyBib eL brings along many functionalities of access management and restriction. To develop these features to a manageable contend-blocking function on page level for the web frontend is a central concern of the research-project.

3.4. An accompanying blog for the “Vorwärts”-project

To meet the already large interest in the digitization project the Library of the Friedrich-Ebert-Stiftung is currently preparing the implementation of an accompanying blog for the project. This blog is intended to serve the interests of the various groups of clients from academia and the public sector and will therefore cover a correspondingly broad range of topics. The blog is intended to be provided to the public simultaneously with the publication of the web presentation. In addition to insights into the technical and methodological issues for the

realization of “Vorwärts” project the blog shall mirror the diversity and importance of the sources’ content.

Conclusion

As this report is published the project "Vorwärts bis 1933" is already quite advanced but still far from completion. Therefore, regarding to the pros and cons of the direction, in which we have chosen to walk, no final conclusion can be reached yet. As this report intends to present the current state of a complex project in its multi-layered structure it likewise documents the starting point for further developments that might lead to other problems and questions that will have to be solved:

- How good is the result of the OCR precisely and are there possibilities to optimize it?
- Was it possible, to make use of the OCR-generated text in other research-projects?
- Could the function of the flexible blocking of content be implemented successfully
- Which desirable user features could be implemented at what expense?
- What aspects of the imminent workflow conversion brought real benefits?
- What experience could be made with the accompanying blog?
- How and to what extend has the user interest evolved?
- Did new problems occur in the course of work and how were they solved?

These and similar questions could be the subject of a further report that will have to be compiled, when the project has undergone further developments towards its completion.