News Media Section

A medium is formed: user experiences and challenges digitizing historic newspapers in State- and University library Bremen

First Author (full first- and surname, no title)
Maria Hermes-Wladarsch, Handschriften & Rara, State and University Library Bremen, Bremen, Germany.
E-mail address: hermes@suub.uni-bremen.de

Abstract:

Between 2013 and 2015, the State and University Library Bremen was part of a newspaper digitization pilot scheme, founded be the German Research foundation. In this project, the State and University library Bremen digitized its whole collection of German language 17th century newspapers. To successfully perform the project, we closely worked together with scientists.The presentation focusses on user expectations concerning the digitization and online presentation of historic newspapers in libraries as well as the challenges for the librarian praxis.
It divides into three parts: After the named DFG-project on newspaper digitization and the challenges resulting from the difficult material are shown, user expectations on digitization of historic newspapers will be presented. Afterwards user requirements on the presentation of digitized newspapers will be discussed. The conclusion shows experiences with a user orientated newspaper digitization.

Keywords: Newspaper digitization; 17th century newspapers; German Research Foundation Pilot Scheme; user expectations on newspaper digitization.

Introduction

Newspapers, as natural they seem today, developed only about 400 years ago. With the first printed newspaper of the world in 1605, not only one publication among others developed: For the first time there existed a medium which published information on world affairs for everyone. Without newspapers, the changes of the early modern age cannot be understood. Therefore, newspaper digitization must have started with 17th century-newspapers.

Between May 2013 and December 2015, the state and university library Bremen was part of the project "digitization of historic newspapers", funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG). The project had, supplementary to the development of structures and standards in newspaper digitization, three aims: First, to raise the digitized historic newspapers in Germany up to 1.5 million pages; second, to improve the so-called Viewer of the German Research Foundation as the central presentation system for historic newspapers; and third the development and improvement of the German union catalogue of serials (Zeitschriftendatenbank) for newspapers.

First I would like to introduce you to the digitization project of the State and University Library Bremen. In this project, the material's complexity required a close cooperation with scientists. This presentation is based on the experience and results obtained here, as well as on pursuing discussions and questionings with scientists working with historic newspapers. – Afterwards, I would like to present the requirements on newspaper digitization that scientists formulated in our DFG-project. Finally, I will present scientist's expectations on the *presentation* of digitized historic newspapers. I am referring (additionally to our experiences in the digitization project) on the one hand to results of a scientist's workshop on newspaper digitization in the named project, funded by the German research foundation; on the other hand on guided interviews with scientists working with historic newspapers, which I realized in 2015.

1. The German Research Foundation project on newspaper digitization

In the pilot scheme of the German research foundation, the state and university library Bremen digitized its complete inventory of German language 17th century newspapers. The specifics of newspapers (which are publicity, periodicity, currency and universality) developed during the 17th century. This was a challenge for us: to satisfy the user expectations as well as meet the specifics of the early medium newspaper.

The library in Bremen holds a more complete inventory of German language 17th century newspapers as any other library. The originals have been gathered within quite a few decades by the Institute for German press research of university of Bremen. Initially, the originals lay scattered over more than one hundred libraries and archives all over Europe. In Bremen, the single newspaper editions were researched, assigned to newspaper titles and finally arranged alphanumerical. The originals were microfilmed and later on re-printed. As the originals lay scattered all over Europe, we decided to digitize those reproductions of

microforms, for the quality was better that the films' quality. Our intention was to have the best possible digitization result and to include the results of scientific research. This way it was possible to include the scientific results of the classification of single newspaper editions to newspaper titles and by the same time to satisfy user expectations of digitization of a complete century of newspapers.

In our project, we digitized 500 newspapers which were published between 1605 und 1700 with roundabout 375.000 pages and about 750 titles. It was always clear to us that, as the 17th century newspapers are in the public domain, they would have to be presented under creative commons license public domain mark 1.0. To enable an easy access, the titles were catalogued in the German union catalogue of serials. In this process, all changing titles of the newspapers have been recorded: Some 17th century newspapers had up to 20 titles. By now, all titles can be researched including divergent spellings.

To present our newspapers, we use the digitization software visual library. There are different access points for our users: an alphabetically sorted list of newspaper titles, of editions, of places of printing and a list of dates of publication. Alternative, when using the search field our users can research in metadata.

A calendar function we considered necessary. This was confirmed again and again by scientists. It is the first time the early newspapers are made accessible on newspaper edition level. But there were significant difficulties: The early newspapers consist of reports on events, which took place on various locations in Germany and Europe. Those reports were sent to a newspaper's publisher, then collected and printed, often without censorship. Additionally to the dates of publishing, the first newspapers therefore include reporting dates and reporting periods. Those two types of dates appear in two calendars. The German territorial landscape has always been fragmented. At different locations in Germany, up to the 18th century different calendars were in force, the – older – Julian calendar and the – nowadays used – Gregorian calendar. Those calendars differentiate by 10 days. That means, each date in our newspapers could possibly refer to the one or to the other calendar system.

We could not expect of our users to know which calendar was meant in a specific newspaper edition. Therefore, we verbalized the dates to reflect the historic character – and we converted the dates into the nowadays used Gregorian calendar. We recorded three types of dates: First the reporting periods due to the Julian calendar; second, the reporting periods due to the Gregorian calendar; third, the dates of publishing.

The calendar on our digitized newspapers first shows an overview of decades, afterwards of years, finally in a monthly view. The user can now choose concrete dates of

publishing. It is possible to see a list of all newspaper editions that were published on a special date. This way, users can research for concrete events such as events in the context of the Westphalian peace, as well as they can research for special dates. But it's a characteristic for 17th century newspapers that not every newspaper edition had a date of publishing. Therefor, the calendar system does not show every existing and digitized newspaper edition. – With our calendar, all newspaper editions are interpreted for the first time. To enable such a presentation, the dates of all newspaper editions – about 80.000 dates – have been manually collected. Our aim was to generate a foundation for a computer based analysis.

Quite a few times scientists asked us to additionally provide full text. Our newspapers, though, show a lot of irregularities: extent, title, font and dates change often, occasionally from one to another newspaper edition. Changing fonts in the Gothic typeface family sometimes in one and the same newspaper edition have to be emphasized. Our tests on optical character recognition therefor produced only inadequate results: The optical character recognition had a quality of 60 to 80 percent which is not acceptable for scientific research. Therefor we decided to do without full text.

But what about the concrete cooperation with scientists in our project? We digitized roundabout 500 newspapers. At least 100 titles have not yet been described by scientists. To prioritize the titles for digitization, we closely worked together with scientists. Many times scientists asked us for titles they worked with, when digitizing we gave priority to those titles. By this means we were able to ensure that our material was presented to the users as soon and as complete as possible. – Also the title format was agreed upon with scientists. Finally, with specific issues we were continually in contact with scientists. This close contact enabled us to closely relate our digitization with scientist's requirements. – In our project, we had the strategy of a controlled pragmatism. This way it was possible to transfer a complete century of newspapers in digital transformation.

2. User expectations on digitizing historic newspapers

Before starting the pilot scheme on newspaper digitization, in 2009 there was a round table discussion on newspaper digitization in the State and University Library Bremen. Scientists formulated the urgent needs of newspaper digitization in Germany. In 2014, there was another scientist workshop in the State and University Library Bremen. We invited scientists from linguistics, literary studies, cultural studies, aesthetics, media studies, history (political history, social- and economic history). They were supposed first to present their

research topics; secondly, to formulate their desires concerning a content-related selection of historic newspapers for digitization; third to name requirements on the presentation of digitized newspapers. Those questions we also asked scientists in mid-level academic positions while a focus group interview when our project was finished.

Asking the scientists about their desires on newspaper digitization, we sometimes got answers like this: "Once all 17th and 18th century newspapers are digitized and we have search engines, I am a happy person."[1] But what concretely do they want from newspaper digitization?

After having asked scientists about a content related selection, we had to notice: The expectations on newspaper digitization highly depend on the discipline and the questions. For example, linguistics want to include all dialectal and geographical regions and different types of newspapers in digitization; first of all, they are interested in full text. Additionally to figuring the typological spectrum, media scientists are interested in digitization of newspapers that shaped the newspaper landscape in one or the other way (e.g. evergreens, leading mediums and innovative mediums). It is important to digitize whole newspaper editions, not only singular pages or articles. Polities scientists want to have different reference levels of exercise of power considered (national, regional, local), they are looking for types of communicating policy with text and picture. Historians emphasized, that digitization should always include complete periods and historic eras. Full text is desirable. To evaluate historic newspapers with digital humanities methods and of big data point of view, it is necessary to provide big corpora of full text. Finally, visual culture scientists need to have high quality illustrations with high resolution.

When summarizing the scientist's desires we found: Important is first to digitize a typological spectrum, which means newspapers of big centers as well as small regions and news sheets (so-called "Intelligenzblätter"); second to consider evergreens, leading mediums and innovative mediums; third to digitize a wide range of regional and political newspapers; fourth to digitize leading exponents; and finally to provide thematic collections. Because a selection of newspapers for digitization highly depends on the questions asked it is even more important to ask the scientists themselves what they need. Every selection of newspapers for digitization has to be linked with scientist's desires.

While the scientists differed when naming criteria for a selection of newspapers, there was a good agreement on the ways historic newspapers should be digitized.

[1] *„Wenn einmal alle Zeitungen des 17. und 18.Jhs. digitalisiert werden und man verfügt über die Suchmaschinen, bin ich ein glücklicher Mensch.“*

1. First of all: what do we mean when we say „newspaper digitization“? A number of scientists commented, he or she would not understand the librarian distinction of newspapers, pamphlets, flyers and single sheets, especially when the early publications are meant. This distinction would only be relevant for really special questions. They would prefer to see all types of publication in one and the same portal. When researching for a special type of publication they would like to be informed about the distinguishing criteria.
2. When digitizing historic newspapers, legal issues always matter. Copyright (when it comes to the digitization of non-public domain newspapers) and data protection have to be mentioned. Those topics are highly relevant for providers like us. When digitizing newspapers from the period of German National Socialism, terms of access are relevant also. For scientists, these kinds of questions are less relevant – more precise, not relevant at all. For them, the unrestricted access to digital newspapers is necessary. There would be no using comfort in looking at newspapers, e.g. of the period of German National Socialism, only at defined reading places within a library. Scientific users see the main benefits of digitization in an open access, independent of time and place.
3. In 2014, scientists discussed in Bremen about the objectives of digitization: Should newspapers be digitized en masse, would it therefor make sense to digitize only microfilms – or should we prefer high quality scans from originals with optical character recognition that would take longer to be produced? Once again we had to realize that the expectations on newspaper digitization highly depend on the discipline and the questions.
4. But independent from the discipline, many scientists emphasized the significance and value of optical character recognition. A statement might illustrate: „Facsimiles are nice and everything, but frustrating when you are looking for something in a newspaper and do not exactly know where you can find it. When an optical character recognition is too expensive or too difficult, keywording or semantic indexing could help. At least names and other identities could be found this way and link directly to the digitized page. Starting from this page, one could read further on. A lot to do, I know.”[2]
5. Art historians, on the other hand, require high definition scans, not only microfilm scans.

[2] *„Faksimile-Ansichten sind hübsch und alles, aber frustrierend, wenn man etwas in den Zeitungen sucht, von dem man nicht schon genau weiß, wo es steht. Wo eine Volltexterfassung zu teuer oder schwierig ist, da könnte eine Verschlagwortung/semantische Erschließung helfen, so dass zumindest Personennamen und andere Entitäten erfasst sind und direkt auf die digitalisierte Seite verweisen, so dass man von dort aus lesend den Inhalt weiter erschließen kann. Viel Arbeit, ich weiss.”*

6. All scientists emphasized, it would be necessary to provide a funded sample of digitized newspapers as soon as possible as critical mass. The stock of The Microfilm Archives of the German Language Press (MFA) should be used for digitization.

3. User expectations on the presentation of digitized historic newspapers.

Afterwards I asked the scientists about their desires on the *presentation* of digitized historic newspapers. First I wanted to know how scientists evaluate the present situation on finding digitized newspapers. The answers had a wide range. It started with: „Really bad, because it depends on accidences if you find or don't find the digital reproductions. A central evidence – similar to ANNO in Austria – would be necessary. Even if I know that there are digital reproductions in Bremen and navigate to the website, often I don't know where to look next."[3] On the other hand, another user said: "I'm just happy about everything that is available digitized. I enjoy the Bremen project very much."[4]

But what exactly are they missing? There is one topic that outshines it all: The access to digitized historic newspapers could be better. The scientists desire one access point for all digitized newspapers: „Central, as a virtual reading room, following ANNO's example, with reporting obligation and at least links for all German institutions as well as systematic research for digitization of German press in other countries."[5] Another user said: „It would be ideal if archives listed and digitized their stocks of newspapers. But for there is no complete bibliography of newspapers in German archives, such a registration and digitization will medium-term not take place."[6] By the way: Quite a few times I addressed the scientists that what they are looking for is included in the new version of the German union catalogue of serials. But no one did know about the new functions. One important aim of newspaper digitization therefor has to be the mediation of the great new options and functions e.g. of the German union catalogue of serials.

The desire of an uniform access is standard; additionally, some scientists expressed something which can be called gold standard: "One single overall bibliography or data basis

[3] *"Sehr schlecht, weil immer wieder von Zufällen abhängig ist, ob man die Digitalisate findet. Es wäre ein zentraler Nachweis - vergleichbar dem österreichischen Beispiel - nötig. Selbst wenn ich beispielsweise weiß, dass in Bremen Digitalisate vorhanden sein sollen und gehe auf die Homepage, dann stehe ich völlig ratlos da, wo ich nun weitersuchen soll."*

[4] *„Ich bin einfach froh über alles, was mir digitalisiert zur Verfügung steht. Das Bremer Projekt genieße ich dabei schon sehr."*

[5] *„Zentral, als virtuellen Zeitungslesesaal nach dem Beispiel von ANNO, mit Meldepflicht und mindestens Verlinkung für alle deutschen Institutionen sowie systematischer Suche nach Digitalisaten deutschsprachiger Presse in anderen Ländern."*

[6] *„Es wäre ideal, wenn Archive ihre Zeitungsbestände listen und digitalisieren würden. Da aber zu diesem Zeitpunkt keine komplette Bibliographie der Zeitungsbestände in Archiven vorliegt, wird eine solche Erfassung und Digitalisierung auf kurz- bis mittelfristige Sicht allerdings nicht stattfinden."*

where you can search in different ways (for metadata, periods, geographical areas, language areas, newspaper titles..), preferably with the option on full text search. – Additionally: semantic annotation of the full text: names of persons, locations (standardized and linked to the GND etc.), dates."[7]

Summarizing the scientist's expectations on the access to digitized newspapers one can say:

1. In Germany, there are many different newspaper portals. Scientists often do not use the German union catalogue of serials to find out where there are digitized newspapers. Instead, for example, they ask their student assistants to look at all newspaper portals in Germany. The scientist's desires differ from our librarian perfectionism. One scientist said a simple list of links would be enough.
2. With every new portal the users have to get familiar with a new portal logic before he or she can even start a research for newspapers. The scientists I asked said they wanted to start researching content immediately when looking at a website.
3. The scientists also asked for the possibility of a metasearch on all digitized newspapers: „Systematic researches are complicated, users have to rely on single suppliers (such as the State and University Library Bremen) or on chance discoveries (often via Europeana)."[8]
4. Quite e few scientists desired an obligation to report new digitized newspapers, alternatively a button to get informed on new digitized newspapers in the German union catalogue of serials: „Information on which titles have been digitized recently with which extent and intensity, would be great. Digitizing newspapers on demand would be a good way to directly provide source material for scientists."[9]

After finding the newspaper one was looking for, the research within the newspaper can begin. I asked the scientists in which ways they use digitized newspapers. The most frequented search access point was the full text research. In 2013, Bob Nicholson called this the „bottom-up-method" of scientific research in newspapers: We can now access the bottom

[7] *„Eine EINZIGE übergreifende Bibliographie/Datenbank, in der man vielfältig (Metadaten, Zeiträume, geographische Räume, Sprachräume, Zeitungstitel…) suchen kann, vorzugsweise mit der Möglichkeit der Volltextsuche. – Ausserdem: semantische Annotation der Volltexte: Personen-, Ortsnamen (standardisiert und mit GND etc. verlinkt), Datumsangaben"*

[8] *"Systematische Recherchen werden dadurch erschwert, man ist mehr oder weniger auf Einzelanbieter (wie die SUB Bremen) angewiesen oder auf Zufallsfunde (oft via Europeana)."*

[9] *„Eine Information darüber, welche Titel in welchem Umfang und welcher Digitalisierungstiefe gerade zum Korpus hinzugefügt worden sind, wäre allgemein überaus hilfreich. Die Funktion „Wunschzeitung" bei Digitalisierungsmaßnahmen halte ich für eine gute Strategie, die Forschung direkt mit Quellenmaterial zu versorgen"*

level of text directly and navigate to each individual word or phrase directly, provided it has been captured accurately by OCR software. In this perspective, OCR is not simply a desire of scientists working on newspapers. OCR is seen as standard of newspaper digitization because without OCR the high quantities of newspaper pages cannot be researched. A problem is the low quality of OCR full text especially when talking about fracture fonts: The scientists I talked with did not know about the problems with optical character recognition. Additionally to high-quality full text with high recognition rates, a few scientists asked for information on the full text quality on library's websites to be able to evaluate the quality of their research results.

One benefit of full text is the highlighting function; some scientists said it would be great if also the whole article could be highlighted. This assumes that users can look at the whole digitized page, not only at a single article: Contextualizing an article is only possible if you can see the whole page. Therefore, digitization should aim on complete newspaper editions– and it should be possible to download the complete edition as well as PDF, as many users mentioned.

Scientists also use the calendar function as research access point. A common practice for historic works is the research for historic dates and events. But there also can be problems: For example, in 17th century newspapers dates are often missing. When using a calendar function, those newspapers are therefore missed. Concerning the project of the State and University Library Bremen, this could be a problem for users being unfamiliar with the material. This example shows: Search functions should be self-descriptive. Scientific users do not want to analyze a website's logic. Calendar functions are seen as natural.

It is important to notice that search access points also depend on the questions of the researcher: "I need to have the possibility, according to my needs, to select according to location, year or title. Only afterwards I would want to change to another – of course comfortable – view. At least one, if not all newspapers should be full text researchable."[10] On the other hand, recording metadata must not be ignored: „When looking for special newspapers or titles, I want to have a perfectly structured metadata research (year, date, number, title, extent, publisher etc.)."[11]

[10] *„Ich muss sofort die Möglichkeit haben, entsprechend meinen Bedürfnissen nach Ort, Jahr oder Zeitungstitel auszuwählen, je nach Anliegen. Erst danach würde ich einen Wechsel in eine Ansicht vornehmen, die natürlich möglichst komfortabel sein sollte. Wenigsten eine, möglichst mehrere Zeitungen sollten volltextdurchsuchbar sein."*

[11] *„Suche ich Infos zu bestimmten Zeitungen/Zeitungstiteln, dann möchte ich eine gut strukturierte Metadatensuche (Jahr/Datum, Nummer, Titel, Umfang, Hrsg. etc.)."*

Conclusion

So what is the conclusion after a comprehensive cooperation with scientists in a digitization project? Each scientific work with newspapers aims – and has to aim – on the most possible accuracy. When digitizing historic newspapers in libraries, we have to translate those claims and desires into realistic facts. But this only is possible in close association with scientists. First, scientist's interests will have to be considered when selecting newspapers for digitization. But the project of State and University Library Bremen showed that it is really important to remain in intensive contact with scientists that are working with the digitized newspapers during a complete project.

Finally, our experiences show: Scientific users do not want endless debates about the presentation of historic newspapers. They want a clearly structured, easily understandable presentation – and we librarians are the ones to develop such simple and functional systems. The detailed questions that are important for us are less interesting for scientists. But the scientific users want to be considered and be involved in the selection of newspapers for digitization and their presentation. When asked, what they want for future newspaper digitization, one scientist got to the heart of it: "In general, more investigations like this to get to know the scientists desires."[12]

[12] "*Generell häufiger Befragungen wie diese, um die Wünsche der Wissenschaftsdisziplinen zu erfragen*"