IFLA News Media Conference *Reviving the past and keeping up with the future – the libraries' role in preserving and providing access to newspapers and news media*, Hamburg, Staats- und Universitätsbibliothek Carl von Ossietzky, 20.-22.4.2016

Mary Feeney, The University of Arizona Libraries
Ulrich Hagenah, Staats- und Universitätsbibliothek Hamburg Carl von Ossietzky

The IFLA News Media Section „has to lead what libraries ought to be doing in the field of news media", as L. Suzanne Kellerman, Pennsylvania State University Libraries, describes the mission of this section of the International Federation of Library Associations and Institutions (IFLA). It changed its name in 2014 from Newspapers Section to News Media Section and thus expanded considerably its field of activity.[1] One of its main objectives is to disseminate knowledge, interest, and best practices worldwide in the field of newspapers and online news media. Therefore, almost every year, the section holds an open session during the IFLA World Library and Information Congress in August, a satellite conference the week before, and, in spring, a midterm conference usually in a different part of the world in order to reach interested colleagues as efficiently as possible.

The midterm conference 2016 in Hamburg, entitled *Reviving the past and keeping up with the future – the libraries' role in preserving and providing access to newspapers and news media*,[2] explicitly focused on both themes, the digital transformation of historic newspapers and the completely new demands and challenges libraries have to accept if they want to remain stakeholders in the business of archiving and providing access to the whole range of news media.

Newspaper digitization projects

The first part of the conference focused on the theme of digitization of newspapers, with speakers providing details of their projects, along with challenges and lessons.

The Swiss National Library and project partners have digitized millions of newspaper pages over the past several years and made them available at Swiss Press Online (http://www.swisspressarchives.ch). Florian Steffen spoke about lessons learned and how to establish guidelines and models to avoid future issues. Steffen noted that these projects are usually carried out in partnerships. He pointed out mistakes they encountered during preparatory and scanning phases and identified lessons for avoiding repeated errors. Specifically, Steffen emphasized the importance of knowing your collection, from defining it to clarifying rights, as well as the need to clearly define expectations with the digitization company. Lessons learned and solutions have been shared at www.digicoord.ch for exchange of experience between memory institutions.

Regina Wanger, ETH-Bibliothek in Zurich, described the open access platform *E-Periodica*, which is run in collaboration with the Swiss National Library. *E-Periodica* includes Swiss journals in history, science, and culture that have been digitized. Wanger discussed cost models for journal digitization, the workflows, and a redesign in 2016 of *E-Periodica* to include responsive design for mobile devices, a more visual display, and split views of search results or table of contents with page views.

The Bibliothèque Cantonale et Universitaire, Lausanne, Switzerland, provides open access to historical editions of major newspapers of Canton de Vaud on its platform, *Scriptorium*. Silvio Corsini presented about their digitization project, providing context, and like Steffen, addressing difficulties and solutions.

Challenges may include identifying rights holders for the newspapers, determining gaps in holdings, gathering information for metadata, and funding for digitization. Corsini also demonstrated searching and display features of *Scriptorium*.

Erling Kjærbo discussed the National Library of the Faroe Islands' digitization of Faroese newspapers. They have completed seventeen newspapers, about 600,000 pages. Kjærbo noted that, while that may not sound like a large number in comparison to some other projects, it covers nearly all Faroese newspapers published from 1852-1999. There were close to two million page visits in the first year. Faroese publishers have been very willing to include their titles as open access. At the conference, a question was posed about how optical character recognition (OCR) has worked with their particular newspapers. Kjærbo noted that there was no Faroese dictionary to attach to OCR scanning for automatic correction, but it has done well; human correction of OCR tends to take a long time. The library has continued its digitization efforts, currently working on manuscript collections and periodicals from the national bibliography.

Another unique newspaper digitization project is the *Catholic News Archive*, a project of the Catholic Research Resources Alliance. Pat Lawton, Frederick Zarndt, and Jeff Moyer described this collaborative project, which has digitized nine diocesan papers from major United States cities and two national Catholic papers. These newspapers focus on parish life and are unique for local historical information. The newspapers were difficult to access, in scattered locations, and not fully represented in bibliographic sources. The project had thirty „digitizing partners" who helped identify the locations of and best quality source materials and negotiated with copyright holders. Another partner in the project was Reveal Digital, which aims to assist libraries in crowdfunding their open access projects through investment by participating funding libraries.

Tim Russell from NewsBank, in addition to discussing the changing landscape of news, also presented about digitization of unique content and cost models. Using South Africa's *Rand Daily Mail* as an example, Russell described some of the considerations and challenges of this endeavor.

<u>DFG pilot and newspaper digitization projects in Germany</u>

Additional developments in newspaper digitization included a report from Caroline Förster on the German Research Foundation (DFG) Pilot Scheme. Participants included Bayerische Staatsbibliothek, Staatsbibliothek zu Berlin, Staats- und Universitätsbibliothek Bremen, Universitäts- und Landesbibliothek Halle, Sachsen-Anhalt, Deutsche National Bibliothek in Frankfurt, and Sächsische Landesbibliothek – Staats- und Universitätsbibliothek Dresden. The plan was for the libraries to merge their knowledge and experience of newspaper digitization into a masterplan and make available important technical and strategic information.

In relation to the development of the national digitization scheme, Hans-Jörg Lieder, Staatsbibliothek zu Berlin, discussed the data available in the Zeitschriftendatenbank (ZDB), German National Union Catalog of Serials, such as newspaper holdings, publication frequencies, and place of publication, that could inform the scheme. Jessica Hubrich also reported on the ZDB, describing features of the new ZDB platform, including graphical representations and interrelationships between titles.

While joint strategies and a unified plan are being developed, individual libraries in Germany are making progress on their digitization projects. The Staats- und Universitätsbibliothek Bremen has digitized its

collection of seventeenth century German language newspapers. Maria Hermes-Wladarsch presented about the collection, what their project learned about user expectations, which varied by discipline, and what users desire in the presentation of digitized newspapers.

Kerstin Wendt and Michael Luetgen presented about using an open source solution, Goobi (now named Kitodo), in the digitization of newspapers, as well as maps, images, manuscripts, and other materials, at the Staats- und Universitätsbibliothek Hamburg. The presentation included information about how much has been digitized to date, a description of the workflow tool and metadata editor, and new features for newspaper digitization.

Another report, from Olaf Guercke at Bibliothek der Friedrich-Ebert-Stiftung, Bonn, was about the digitization of *Vorwärts - Berliner Volksblatt*. Guercke described the importance of digitizing this political and social periodical, published 1876-1878 and 1891-1933. He also discussed the pros and cons of in-house digitization, what has been completed to date, and objectives going forward for presenting the publication on the web.

The Badische Landesbibliothek in Karlsruhe had been digitizing medieval manuscripts, books, and music scores for five years, and decided to also implement a newspaper digitization project. Dr. Ludger Syré described workflow, OCR processing, and the results, which include about 900,000 images from six newspapers, published between 1784 and 1944. Syré also presented the calendar view of the digitized newspapers, which allows for browsing by different years, months and days. The homepage of the digital collection has information about missing volumes with a request to other institutions with newspaper holdings to help close the gaps in coverage.

Users and usage of digitized newspapers

In addition to presentations about the status of newspaper digitization projects, another theme of the conference centered on the use of these digitized newspapers: how they're being used by scholars, the possibilities for data mining, and more.

Clemens Neudecker, Staatsbibliothek zu Berlin, examined what people are doing with digitized newspapers and presented examples of use cases. For example, there is interest in newspapers for Digital Humanities and text mining, for which large quantities of text and data may be more important than the quality of the scanning. Some examples cited by Neudecker include the *Viral Texts* project (http://viraltexts.org), and the extraction of knitting and crocheting patterns from newspapers in *Trove*, the National Library of Australia database of digitized newspapers. Another use case presented by Neudecker was the development of apps and other creative uses of digitized newspaper collections. For example, the KB National Library of the Netherlands has created an app for mobile devices that enables users to find historical newspaper articles connected to their location. Finally, the use of digitized newspapers for family history research was discussed. In all of these use cases, as Neudecker summarized, OCR of the full-text of newspapers is almost always necessary, open access of content is ideal, and APIs provide the opportunity for innovative uses.

Another discussion about the use of digitized newspapers was presented by Jean-Philippe Moreux, Bibliothèque National de France. Moreux described a project to apply data mining techniques to metadata in the *Europeana Newspapers* collection. A subset of the papers has been refined with Optical Layout Recognition (OLR), which describes the structure of each newspaper issue and article and

classifies the type of content. Moreux's examples of how data mining and visualization tools were used to uncover information in the data included a visualization of the changing importance of illustrations in newspapers; the change in types of content in newspapers over time; and word density per page. The datasets and charts are publicly available at http://altomator.github.io/EN-data_mining/.

Juha Rautiainen, National Library of Finland, talked about how use statistics can tell us about users of digital newspapers. Data collected by the system of the *Historical Newspaper and Journal Library*, which contains millions of pages of digitized newspapers and journals, shows the title, publication day, and language of the title being accessed. Analysis of the data was approached in two ways: by forming a question first and seeing if the data supported it, and studying the data itself for explanations. For example, does the promotion of links in the *Historical Newspaper and Journal Library* through social media increase usage? Another example presented by Rautiainen was looking at the number of page loads in proportion to the pages available by year. This helped reveal which time periods were being used more frequently.

In keeping with the IFLA News Media Section's expansion of scope to news media in any format, Lisa Romero's presentation focused on television news, not digitized newspapers. One goal of the International News Lounge at the University of Illinois at Urbana-Champaign is to increase access to international news perspectives among its students. Users of the news lounge were surveyed, and the findings provided a better understanding of users' channel preferences (Euronews/RAI was most popular, followed by Al-Jazeera/Al-Arabiya), preferences for design of the space, and what they valued about the news lounge.

Keynote speaker Prof. Dr. Jan Christoph Meister, Universität Hamburg, discussed *Libraries as 'Epistemological Agents' of Digital Humanities*. Reflecting on the location of the conference in the historic building of the Staats- und Universitätsbibliothek Hamburg, Prof. Meister talked about the changing role of libraries, inviting librarians to be more active agents in research and not just providers of things useful to researchers. Prof. Meister also discussed what he would like to see from libraries, from his perspective as a digital humanist and researcher: the digitization of relevant corpora, and support for big data – large corpora such as digitized newspaper collections – which can be used by digital humanists for automatic exploration such as pattern recognition and topic modeling. Prof. Meister presented examples of tools that scholars use for „interactive close reading" of a corpus, such as Voyant (http://voyant-tools.org) and „distant reading" of large corpora, including the topic modeling tool, Mallet (http://mallet.cs.umass.edu/about.php). He posed the question of what is the library's role in such a process and other Digital Humanities (DH) research. The quality of a corpus is important to researchers, and this is part of libraries' roles. Prof. Meister provided examples of large data sets – „the good, the not-so good, and the fake and useless" for his purposes – such as Deutsches Textarchiv, Google Books, and Deutsche Digitale Bibliothek. Libraries have done their own digitization of collections or provided digitized content from other sources and put it out there, and researchers want to be able to rely on what libraries provide to be used in technologies like topic modeling. Libraries need to provide the skills and support for scholars who want to use digital materials. Prof. Meister proposed that libraries can be agents in research by offering a DH consultative service, through which librarians would provide data analysis support, consultation for scholars on practices and methods that are particular to a corpus, and an understanding of best practices. Librarians could also maintain „tried and tested" analytical tools and provide a DH help desk. These would be new, but natural, roles for libraries. His talk was motivating to many attendees to take on these new exciting roles.

Digital News Media

While retrodigitization strategies are more or less consistently outlined and driven by libraries, archives, or some service providers, and this often happens in close contact with the researchers' communities, taking into account their needs and DH requirements, for example, things are quite different in respect to born-digital news media.

Anticipating the conclusions of the Hamburg papers, we could say: „Digitization was just faster than the news industry", as Prof. Dr. Stephan Weichert, Hamburg Media School, put it in his keynote speech *The Quiet Death of the Paperboy, or how the World of News Media is Trying to Catch up with Digitization.* Media evolution follows young users' demands at an accelerating rate. This is also going on more quickly than memory organizations can keep pace. Researchers who are aware of what will be lost act in their own way to secure traces of cultural news heritage, as far as they are interested in it. They creatively develop strategies and conduct research in small segments, defined by certain sources, topics, regional or time slots, media types (e.g., apps), or a professional teaching purpose (media schools). But they do not necessarily think about libraries or archives, neither do they notice official collection policies of memory institutions, nor do they think of any systematic preservation approach – or are bewildered about how little usable outcome derives from activities of these players. On the other hand, national libraries try to cope with the extension of legal deposit collecting to web publications, but it seems to be mere technical, administrative activism with relatively poor indexing and use conditions, so that website archiving as we know it currently does not meet much interest on the scientific users' side.

Some of the findings that were reported and discussed during the conference: Prof. Weichert insisted on the disruption, replacement, and revolutionary character of digital innovation. „We do not need traditional news publishers anymore", because they miss anticipating new trends, and they cannot cope with emerging new formats, new storytelling patterns, new socially embedded expectations as fast as needed: „Millennials want a different news experience." Some facts are already well known: selective, volatile, contingent media usage, talkativeness, arbitrarily combined usage of „multiple devices all throughout the day".[3] Latest news usage research offers insights into the daily life-cycle of news. Millennials appreciate short news espressos in the morning, long-reads during the commute, videos in the evening. Prof. Weichert described all these trends and facts, but did not propose ways to keep this current media environment reconstructable for the future.

Frederick Zarndt evaluated metadata formats that are feasible for making born-digital news searchable in the future. He first reviewed some international comparisons of information behaviour, stating that only in Germany, France, Great Britain, and Japan, TV is still the first source of news for the majority of the population; elsewhere social media have taken the lead. Looking at the main source of news by age clearly shows that the young prefer online to print newspapers, and Facebook is comparatively the strongest provider of online news. As for metadata standards, Zarndt stressed the importance of the International Press Telecommunications Council (IPTC) protocols (https://iptc.org/), and especially of Schema.org: „The IPTC brought openness, flexibility, tenacity and deep domain expertise to their participation in the Schema.org effort." He recommended „marking up your articles using the … properties of the https://schema.org/Articletype. Be sure to mark up your page with the most specific applicable schema.org type. For example, a news article should be marked up as https://schema.org/NewsArticle and a blog post as https://schema.org/BlogPosting". Zarndt noted that digital news harvested by libraries is mostly at the level of websites, not individual articles, with the exception of the National Library of Sweden, which collects RSS feeds.

Individual strategies for preserving their own products are being taught at the Missouri School of Journalism. Edward McCain and Dorothy Carner described the integration of personal digital archiving in the journalism curriculum to raise awareness for preservation needs in the news business, by teaching future journalists about the importance of preserving their work. This is only one element of the larger strategic initiative called *Dodging the Memory Hole* which aims at developing ways of preservation for e-only news resources for the future.[4]

Tobias Steinke presented the Deutsche Nationalbibliothek's (DNB) web archiving activities originating from the new legal deposit regulations of 2006. Steinke described their experiences from a .de-domain crawl and discussed restrictions that did not draft a very positive vision for the future: websites are being crawled every six months selectively, the snapshots taken cannot keep dynamic content displayable, archive copies can only be shown inside the library's buildings because of copyright law, and DNB cannot gain individual permissions for free display on the web from all producers. As far as the news industry is concerned, German press publishing houses did not permit news crawls behind their paywalls in a general conference in 2014. Only *Spiegel Online* is said to make an exception soon for a second crawl after a previous one which failed because of technical reasons. Publishers' political constraints have again and again been subject of international reports in the IFLA News Media Sections' meetings; Scandinavian or Baltic news publishers somehow seem to be more inclined to conceive pragmatic solutions.

The changing modes of news media, as discussed by Prof. Weichert, was further described by Hasna Askhita, Syrian Computer Society, in relation to the crisis in Syria. Askhita talked about different news media sources in Syria, from government-controlled websites to individuals who document and show events through social media, which allows for different views to be seen. Askhita also presented about digital local Arabic content, including a media blogging site, *eSyria*, that has served an important role in documenting local cultural heritage.

Methodological reflections about news apps were discussed by Meredith Broussard and Katherine Boss, New York University, who discussed complex products of data journalism, which they regard as „the future of journalism". Their starting point was the observation: „Born-digital news content is increasingly becoming the format of the first draft of history. Archiving and preserving this history is of paramount importance to the future of scholarly research, but many technical, legal, financial, and logistical challenges stand in the way of these efforts. This is especially true for news applications, or interactive databases." None of the U.S. programs for news media archiving has up to now dealt with the most innovative journalistic products: news apps, or interactive news applications that are often reliant on external APIs and custom-built software. Archiving faces huge technical issues, that's why „it is a scary presentation", as Katherine Boss underlined – you deal with dynamic, not static, digital objects, complex software dependencies, including the operating system, programming languages, etc. Emulation techniques themselves need to be migrated. A performance model framework for the preservation of the software system is needed, as well as a metadata schema. An alternative idea would be to just capture the database behind the app and sacrifice the interactive features: user interface, visualization, analysis tools. But „this bare-bones preservation approach would strip the news app of most of its purpose and functionality", and prevent any reconstruction of usage modes in the future.

Finally, an insight into a researcher's perspective should be mentioned: Thomas Risse's report on *Improving the Freshness of Web Collections by Integrating Social Web and Focused Web Crawling.* He focused on the day-to-day process of public communication. Event-following and discussion of certain topics provide valuable insights into the thoughts and communication styles of individuals, groups, and

organizations. Risse's research group, based at the Leibniz Universität Hannover, applies flexible crawling strategies. Most important is authenticity, meaning: „Seeing a web page as the user saw the page (e.g. including ads and tweets at that time point)", capturing the full context in social web crawls, contextualizing twitter information, including short-term website changes during an event. The current approach consists of three steps: 1. Crawling of Social Media content, 2. Extraction of Links, 3. Crawling of Web Pages. An integrated approach would include using a real-time stream from Twitter as the source for the web crawl. Social media URLs follow changes in topic and enable early insertion into the crawling queue. A suitable prioritization of crawls of social media URLs is crucial. The envisaged model of „integrated crawling" is destined to maximize the freshness of content as an element of a SocBigData Infrastructure by limiting the gap between social media gathering, semantic extraction, and web crawl.

Despite all individual attempts of gathering and conserving born-digital news, have most of us already accepted the loss of important parts of our cultural heritage? It seems so – dramatic, but inevitable. Institutions that have maximized digital collections over the past decade, as the Denmark State and University Library in Århus, note that collecting, indexing, and making accessible to users what has been harvested from the web is very complex.[5] Also in Germany, the National Library and the regional libraries, as the institutions responsible for the preservation of German physical and electronic publications, more and more enter into a discussion process about their respective responsibilities and possibilities. We will have to compromise, and in certain segments such as news media, representative collecting in a methodologically transparent way will have to replace the traditional idea of completeness.

For the moment, solutions and perspectives for bridging the gaps seem to be out of reach for the nearer future. Things have been developing independently, not interconnected and without systematic interrelations. So, the outlook has to be pessimistic to a certain extent, and there remain huge themes and challenges for future conferences of the IFLA News Media Section, and elsewhere. These conversations continued at the section's Open Session of the IFLA World Library and Information Congress in August 2016, *Here Today, Gone Tomorrow: The Current State of Born Digital News*.

---

[1] Hagenah, Ulrich: Archivieren – aufbereiten – digitale Lebenswelten für die Forschung verfügbar halten: was können, was sollten Bibliotheken angesichts der Umwälzungen des Nachrichtenmarktes leisten? Die IFLA News Media Section und ihre Fachtagungen 2015. In: Bibliotheksdienst 50.2016, H. 3/4, S. 300–317; Hagenah, Ulrich: Nachrichtenmedien im digitalen Wandel. Die IFLA Newspapers Section und ihre Fachtagungen 2014. In: Bibliotheksdienst 49.2015, H. 2, S. 119–133.

[2] All papers of the conference are available at http://blogs.sub.uni-hamburg.de/ifla-newsmedia/?page_id=242 and will soon be publicized in the IFLA Library (http://library.ifla.org/).

[3] See also Cowan, Chris: Media Insight Project 2014, http://www.mediainsight.org/Pages/default.aspx (Zugriff: 13.9.2016).

[4] https://www.rjionline.org/projects/dodging-the-memory-hole (Zugriff: 6.9.2016).

[5] See for example Skovgård Jensen, Tonny ; Schostag, Sabine; Bønding, Niels: Chasing the news. Report from 10 years of digital legal deposit in Denmark, http://www.kb.se/aktuellt/utbildningar/2015/IFLA-International-News-Media-Conference-/IFLA-International-News-Media-Conference---documentation/. And moreover, it is merely impossible to gather complete national heritage samples, because only a small part of the .com, .org etc. websites from a certain geographical zone can be reliably and comprehensively defined.