ONB Labs – An Open Digital Hub of Cultural Heritage

Monika Kovarova-Simecek[1], Sophie-Carolin Wagner[2], and Stefan Karner[2]

[1] University of Applied Sciences St. Pölten, Matthias Corvinus-Str. 15, 3100 St. Pölten, Austria
[2] ONB Labs, Heldenplatz 1, 1015 Vienna, Austria
monika.kovarova-simecek@fhstp.ac.at

Abstract. Historical research practices are being gradually transformed by the digitization of historic sources on the one hand, and the usage of digital methods and semantic technologies on the other. We present two digital initiatives of the Austrian National Library, ANNO and ONB Labs, which, as a digital hub of cultural heritage, enhances accessibility to and knowledge discovery within historical datasets considerably. To do so, ÖNB provides resources such raw data, metadata, or Linked Open Datasets which can be accessed as data dump or via a SPARQL API, allowing for live querying of RDF datasets. ONB Labs, additionally, offers services, tools, and APIs to enrich this data such as IIIF, Open Annotations, SACHA, or Jupyter Notebooks which allow to create and share documents including live code, visualizations and narrative texts. In this demo-paper, we explore the question how digital methods and semantic technologies can be used in the context of historical research and illustrate their application through the example of a project investigating the cultural history of financial news in Vienna 1771-1914. We conclude with an outlook and call for action to encourage the use of ANNO and ONB Labs among digital historians.

Keywords: Digital Archives, Knowledge Discovery, Services and Tools

1   Introduction

Historical research practices are being gradually transformed by the digitization of historic sources and the usage of digital methods and semantic technologies [1]. This demo-paper presents two digital initiatives of the Austrian National Library (ÖNB), **A**ustria**N** **N**ewspapers **O**nline - ANNO (anno.onb.ac.at) and ONB Labs (labs.onb.ac.at), as digital hubs of cultural and scientific heritage that maintain a repository of public knowledge (RPK), enhance knowledge discovery, and support historical research [2]. By providing digitized historic sources and technical tools for their exploration, ÖNB contributes considerably to promoting digital humanities in general and digital history in particular. Against this background, we try to answer the question, how knowledge technologies provided by ANNO and ONB Labs can be used in the context of historical research discussing their application and potentials through the example of a research project which explores the cultural history of financial news in Vienna 1771-1914. We conclude with an outlook on further possible developments and their impact on collaborative historical research.

2   Theoretical Framework

The advantages of integrating quantitative approaches against the background of digitization are widely shared in historical studies and the possibilities of digitization are increasingly being exploited. These initiatives can be summarized by the term Digital History, a movement that understands itself as a branch of Digital Humanities with overlaps on many levels [3]. Digital History goes, however, far beyond purely quantitative approaches known especially among economic and social historians as Cliometrics. Rather, through the integration of quantitative-based but creative approaches and promoting institutional and cross-system collaboration by the utilization of technological innovation, Digital History seeks to extend its sphere of influence to other areas of historical research such as cultural or communication history and to enhance the potentials of cultural heritage preservation. Nicholson [4] speaks of the digital turn in humanitarian scholarship, which is based on two pillars: "the creative use of online archives and a willingness to imagine a new kind of research". He sees the roots of this development in the cultural turn, which brought the formative power of language and culture in social developments into the focus of research, and in culturomics investigating those phenomena with computer-aided methods. For cultural scientists, the added value of digitization lies especially in the ability to recognize hitherto hidden structures and patterns in the sense of Moretti's distant reading [5] allowing to analyse political, social and cultural discourses [6] in a new way.

3   ANNO & ONB Labs – A Hub of Historical Data Sources

With ANNO and ONB Labs, researchers are provided with a rich set of data, tools and services to explore historical data and contribute to digital history research.

3.1   ANNO – AustriaN Newspapers Online

In 2003, the Austrian National Library launched a digitization initiative focusing on the transformation of historical newspaper and magazine collections into digital copies. The result of this project is today available as ANNO (http://anno.onb.ac.at/anno), a digital portal that makes the digitized records available free of charge and without any time limit. The digitization of the historic material takes place continuously, whereby the ranking of the titles to be digitized takes into account the interests of researchers and prioritizes holdings with the highest intensity of use. ANNO pursues the goal of offering collections that are as integrated and complete as possible by collecting holdings from different archives and libraries. Currently, the ratio of fully complete holdings is 22% (271 titles of 1,258), with the holdings being continually supplemented and completed. As of April 2019, 1,258 titles were available in digital format for the period from 1568 to 1947. The OCR-based full-text search covers more than 90% of the digitized material with focus on the years 1689-1947. This corresponds to 1,466,451 newspaper and magazine editions with nearly 20 million pages, which are searchable at key-word level [7].

The full-text search is based on automatic indexing of source material by means of Optical Character Recognition (OCR). Due to still existing problems with the text recognition, ANNO inventories can also be very faulty in some texts. ANNO offers the possibility to specify the search with the help of further search options such as Boolean operators, wildcards, phrase search or distance search. The search results are displayed as a list of newspaper issues in which the given term occurs. The search results are sorted according to different criteria (relevance, date, title or the time of their inclusion into the index) and classified according to several categories (title, location, language, period, and topic). This display also provides information about the date of publication and the number of hits in the issue.

Thanks to its accessibility, scope, and quality, ANNO has become an important digital repository of historical newspapers. However, ANNO is still fraught with a few weaknesses, especially with respect to information extraction and semantic richness. The validity and usefulness of the analysis tools in form of filters are limited due to the lack of transparency regarding category definitions and search algorithms underlying the filter criteria. Some of these points can be compensated by ONB Labs which offer collaborative tools for enriching, processing and visualizing of datasets generated by ANNO.

3.2   ONB Labs–Interfaces

The Austrian National Library conserves, archives and presents Austrian cultural heritage. The new methods to scan and provide printed materials led to an enormous increase in data, requiring new technologies to work with these masses of data, such as data mining, data linking, machine learning, artificial intelligence or algorithmic analysis. ONB Labs seeks to foster and support the application of these technologies.

Analogue to the conservation of memories in a consciousness system, a crucial part of long-term preservation of cultural heritage is its continuous usage and actualization. To ensure that the Austrian National Library establishes services and infrastructure dedicated to foster the work with its collections. Part of these services is the ONB Labs, a platform dedicated to support and inspire research, experimentation, and creative work with the Austrian National Library's digital collections and data. ONB Labs selects and presents openly licensed collections, tools to support working with them, and sketches to illustrate use cases. Data provided via ONB Labs is licensed under CC0 or published with a public domain mark. Via the IIIF API and interface SACHA (Simple Access to Cultural Heritage Assets) users can create collections including various data sets based on their specific needs and interests. These collections can be shared, cloned, exported to other tools, and downloaded.

One service implemented in ONB Labs is the provision of the Austrian National Library's Linked Open Dataset. Linked Open Data refers to openly and freely available data online, identified via Uniform Resource Identifier (URI) and accessible via HTTP. This allows to link and send requests to various data resources. The Linked Open Data Set of the ONB Labs includes metadata of our historic newspapers, our historic postcards, and the catalogue of the Austrian National Library and can be ac-

cessed via ONB Labs as data dump and via SPARQL Lab, allowing for live querying of RDF datasets.

The focus group of ONB Labs are not solely Digital Humanists, but also digital artists, creatives and data scientists. Whereas this might be a diverse group of people, they are combined by working in fields where technology, research and work inform and guide each other reciprocally. Computational analysis is crucial within all of these areas. The provision of raw data, metadata, possibilities to enrich this data, OCR, tools and APIs tailored to the needs of programmers, are immanent to the work of these users and therefore the focal point of ONB Labs.

Sharing, commenting and forking is a reoccuring element in ONB Labs. Using Jupyter Notebooks allows to create and share documents including live code, visualizations, and narrative text. The ONB Labs team shares various notebooks and their documentation to allow users to use them as a basis for their own needs and further share them via the ONB Labs GitLab.

4 Case Study: Exploring the History of Financial News

ANNO and some features of ONB Labs were key elements in a research project investigating the genesis of financial journalism and the popularization of financial news in Vienna 1771-1914. To answer the research questions, how financial journalism changed with respect to relevant newspapers, formats, and functions over time, and which public should be addressed with financial news in what form, we (1) searched ANNO for relevant key-phrases, (2) extracted the metadata associated with the search term (title, date of issue and number of hits) available on the website in HTML format using Web Scraper (webscraper.io) in a structured, statistically evaluable form usable for further analysis, (3) enriched the data with additional criteria such as main topic and target group, (4) visualized the data in form of interactive graphs using Python and Plotly (plot.ly), and (5) created collections of newspapers the visualizations are based upon using IIIF (iiif.io).

The dataset contains 4,028,974 hits spread across five terms: Börse* (1,685,441), „Wiener Börse" (235,597), Aktie* (1,158,329), Obligation* (594,106), and Pfandbrief* (355,501). The query was made on 2019/02/19 and 2019/03/04. As newspapers are digitized continuously, a query at this point in time would lead to different results. For each term, more than 100 newspapers were identified, with more than 80% of all hits being included in the top ten [8]. The visualizations, as shown in Fig. 1. exemplary, provide information on the scope and intensity of financial news over time as well as the relevance of newspapers and readership groups. They serve as an important guide for deeper research and allow direct access to the historic press material. Here, however, it should be considered that the application of IIIF is limited in time due to copyright regulations. Until 1878 the graphs point to IIIF collections, after that the material is available in ANNO.
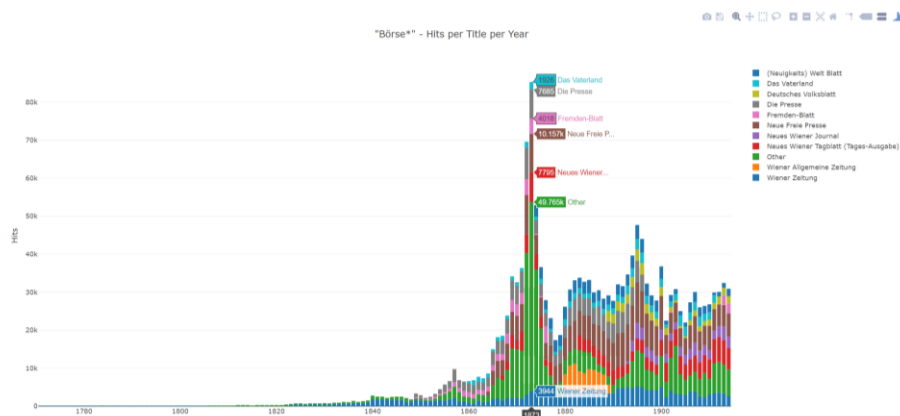
Fig. 1. Interactive graph of results for 'Börse*' displaying hits in the top 10 newspapers.

5   Outlook & Call for Collaboration

The intention of this demo-paper is to draw attention to the datasets, tools, and services provided by ÖNB to encourage collaborative practices in the enrichment and re-utilization of historical datasets among digital historians. Components to annotate user-generated collections based on IIIF and Open Annotations are continuously implemented and in the near future annotations will be sharable – both publicly and privately. We are especially eager to explore the intersections of historical data, linked data, and visualization, and provide support for collaborations via the services of ONB Labs. We invite all interested parties get in touch with OBN Labs and share their ideas.

References

1. Graham, S., Milligan, I., Weingart, S.: Exploring big historical data: the historian's macro-scope. Imperial College Press, London (2016).
2. Neuroth, H.: Bibliothek, Archiv, Museum. In: Jannidis, F., Kohle, H., Rehbein, M. (dds.), Digital Humanities: Eine Einführung. J.B. Metzler Verlag, Stuttgart, pp. 213–222 (2017), p. 213.
3. Cohen, D.J., Rosenzweig, R.: Digital history: a guide to gathering, preserving, and presenting the past on the Web. University of Pennsylvania Press, Philadelphia (2006).
4. Nicholson, B.: The Digital Turn: Exploring the methodological possibilities of digital newspaper archives. Media History 19, 59–73 (2013), p. 63.
5. Moretti, F.: Distant reading. Konstanz University Press, Konstanz (2016).
6. Koenen, E.: Digitale Perspektiven in der Kommunikations- und Mediengeschichte: Erkenntnispotentiale und Forschungsszenarien für die historische Presseforschung. Publizistik 63 (4), 535–556 (2018), p. 8.
7. ANNO Homepage: http://anno.onb.ac.at/suchhilfe.htm, last accessed 2019/06/19.
8. ONB Labs Homepage: https://labs.onb.ac.at/en/topic/financial-news/, last accessed 2019/06/19.