What's Past is Prologue: The NewsEye International Conference

Towards a future of interdisciplinary collaboration between Cultural Heritage, Digital Humanities, Computer Science and Data Science

16-17 March 2021

NEWS EYE

A Digital Investigator for Historical Newspapers

BOOK OF ABSTRACTS

Forward

This publication presents the abstracts of the 2021 International Conference organised by the consortium of the EU Horizon 2020 research and innovation programme project *NewsEye: A Digital Investigator for Historical Newspapers*. This conference, which was held online on the 16th and 17th of March 2021, featured the work of four keynote speakers and over thirty presenters hailing from eleven countries.

Spanning from May 2018 to January 2022, NewsEye is a research project advancing the state of the art and introducing new concepts, methods and tools for digital humanities by providing enhanced access to historical newspapers for a wide range of users. Hence, the conference theme *Towards a future of interdisciplinary collaboration between Cultural Heritage, Digital Humanities, Computer Science and Data Science* reflects certain aspects of the NewsEye project itself. The NewsEye consortium is composed of three national libraries and six universities based in Austria, Finland, France and Germany, which have each contributed specialised knowledge in the aforementioned subjects.

The NewsEye International Conference was divided into six sessions over two days. The contributions of Ann Dooms, Clemens Neudecker, Ian Milligan and Gerben Zaagsma kicked off the first day with talks on the diverse subjects of mathematical approaches for leveraging the information hidden within images of cultural heritage, multimodal perspectives for digitised newspapers, digital history and the politics of cultural heritage digitisation and the transformation of historical scholarship by digitised newspapers, respectively. The order of this publication follows the remaining five sessions.

More information about the NewsEye project can be found on the project's website, newseye.eu.

Day 1: 16 March 2021

Session 1: Paper Presentations from NewsEye Contributors and Colleagues

How Finnish was Finland's First Newspaper?

Antti Kanner | Eetu Mäkelä | Jani Marjanen | Mikko Tolonen | *University of Helsinki*

Although newspapers are generally seen as a seventeenth-century invention, the medium became considerably more common in the eighteenth century and spread to most parts of Europe. Finland gained its first newspaper in *Tidningar Utgifne af et Sällskap i Åbo,* which was first published in January 1771 in Turku (Åbo). As Finland was a part of the Swedish kingdom until 1809, it is reasonable to not only see *Tidningar* as a Finnish newspaper, but as part of the ecosystem of Swedish newspapers published in the capital Stockholm and in some regional towns, such as Turku in Finland. Due to a tradition of writing newspaper history from the perspective of the later nation-states, both Swedish and Finnish historiographies have failed to deal properly with the role of *Tidningar* as part of the newspaper ecosystem of the period. From a Swedish perspective, it has been ignored because of its Finnishness, whereas from a Finnish perspective it has been highlighted as the first without understanding its imperial context. Both perspectives are flawed. Furthermore, previous studies have not been able to address what being Finnish meant to the men producing this paper in the late eighteenth century or how Finnishness could manifest itself in the newspaper.

We seek to understand the role of the *Tidningar* by using new opportunities provided by the digitalization of historical newspapers and by asking on what grounds and to which extent *Tidningar* can be seen as a Finnish newspaper. We argue that different aspects of its Finnishness can be gauged by 1) looking into the normative position of the editors and authors of the paper with regard to Finland and the Finnish language, 2) studying which themes and topics the newspaper described as Finnish, 3) mapping which locations (both within and outside Sweden) were mentioned in the newspaper and analyzing to which extent they come across as particularly related to Finland, and 4) interpreting the relationship of *Tidningar* to other newspapers published in Sweden in particular by comparing their contents. To answer our questions, we deploy a mixed-methods approach in which we combine reading of key texts with statistical analysis of linguistic features and the mapping of Named Entities.

Overall, our analysis sheds light on different aspects of Finnishness and thus provides a more nuanced image of *Tidningar* as a vehicle for Turku-based intellectuals to place themselves on the map of Swedish intellectual life. Part of this staging was that their newspaper targeted particularly Finnish issues, which is also visible in the contents of the newspaper. However, the Finnishness of the newspaper comes across as very uneven, focusing on only some parts of Finland and only on some topics, most prominently language, religious life and the economy. We also highlight that this type of mixed-method approach can be useful also for digital humanities studies that deal with rather small datasets, as it highlights a healthy combination of qualitative interpretations and quantitative results.

Corpus Building Meets Text Mining: The Creation of a Topic-specific Newspaper Corpus on the Topic of Return Migration using LDA and JSD

Sarah Oberbichler | *University of Innsbruck*

This paper addresses topic-specific corpus building for historical research and shows how text mining methods can support corpus building in order to distinguish between relevant and non-relevant articles on the topic of return migration. The goal is to illustrate how simple but creative approaches can lead to good solutions. I first present the motivation and necessity of creating sub-collections for specific research questions, then explain the problems and difficulties encountered in the process of corpus building, and finally show results of a novel approach combining Latent Dirichlet Allocation (LDA) and the Jensen-Shannon Distance (JSD).

While the motivation to create a topic-specific newspaper corpus is based on the lack of material to study return migration (which is defined as 'cross-border migration to the country of origin' (Currle, 2006)) to Austria between 1850 and 1950, the necessity to use text mining methods arose from the difficulty to work with search keywords alone. The topic of return migration is difficult to define conceptually. There are only a few clear terms such as 'Rückkehrer', 'Heimkehrer' or 'Rückwanderer' (all German terms for 'returnee') that lead to more or less relevant articles on return migration. However, they only cover a small amount of the whole spectrum of return migration articles in newspapers. The combination of keywords that occur together in a defined word distance (e.g., return migrants, returning emigrants, returning home, returning families, return of emigrants) is helpful but error prone. Even if it were possible to cover the topic of return migration in its entirety, many of the word combinations would lead to results that are not relevant. On the other hand, expanding the search by avoiding word combinations and using only typical words in a collection, such as 'return' or 'returning' on return migration, leads to a considerable number of irrelevant articles.

In order to use machine learning to support the building of a representative corpus on return migration, this process starts with the creation of a manually annotated training and testing collection, containing relevant as well as non-relevant articles. This step is performed with the beta version of the NewsEye platform, where improved newspaper issues of the ANNO newspaper collection are available (https://anno.onb.ac.at). In the next step, LDA and the JSD method are being used to group words and similar expressions that best characterize relevant or irrelevant documents for the topic of return migration and to measure the similarity between articles. While the combination of LDA and JSD to group similar articles, documents or groups of documents has been described in several research papers (Fothergill et al., 2016; Lu et al., 2019; Niekler & Jähnichen, 2012), here a new approach was tried out due to the necessity to find representative results for further qualitative analysis. For each article in the test corpus, the ten most similar articles from the training corpus are extracted. These articles carry the information about the manually assigned relevancy. If 60 percent of the automatically found similar articles are annotated as relevant, the new article is marked as relevant, too. Otherwise it is marked as irrelevant. As a result, it can be said that corpus building can be successfully supported by text mining methods if manual annotations are part of the method.

Towards Automated Discourse Change Detection

Quan Duong | Lidia Pivovarova | *University of Helsinki*

Large collections of text, such as news archives, reflect valuable information on discourse dynamics: the change in the most discussed topics, opinions and attitudes. Discourse change is not necessarily reflected in language change but nevertheless could be detected in large collections of diachronic textual data.

Time series derived from text are frequently presented in digital humanities papers, though used mostly for illustration rather than as an input for a formal analysis. In some studies, diachronic difference in the data constitutes the main research question, e.g. [1,2].

One of the difficulties in digital humanities is that they often have to investigate humanities research questions and implement computer science methods at the same time. In digital humanities research, questions are generally very complex and involve a lot of uncertainty, thus ground truth needed for numerical evaluation is usually unavailable. Moreover, quite often digital humanities research deals with a specific use case, which means that they deal with a single non-annotated dataset without proper split into training and test subsets.

To overcome this difficulty, we propose an evaluation on multiple synthetic datasets. The idea is to exploit manually assigned categories that are labelling articles in many news collections. Distinct periods and spikes in the data could be mimicked by sampling from a single label according to a certain pattern, while all other categories are sampled randomly. Then the task is to implement a model able to find a subset of documents that are related to the same theme and follow the pattern, without looking at the manually assigned labels. Synthetic data is widely used in lexical semantic change detection ([3,4], among others), but we are unaware of any similar work performed on the discourse level and exploiting news categories for a similar purpose.

We present our preliminary experiments on Finnish news datasets. We build multiple synthetic datasets for the YLE news archive (freely available from the Finnish language bank: http://urn.fi/urn:nbn:fi:lb-2017070501). These datasets are used to evaluate trend detection methods. Then, the best performing methods are applied to the STT dataset to find potential discourse change in the independent dataset (freely available from the Finnish language bank: http://urn.fi/urn:nbn:fi:lb-2019041501).

The basic approach that we use consists of two steps: first, breaking the news collection into smaller datasets and then classifying these datasets as either stable or non-stable. For the first step, we use either clustering or topic modelling, for the second step recurrent or convolutional neural networks and a few simpler baseline methods. Our experiments show that synthetic datasets allow us to rank methods as either more or less suitable for the task. Application to the STT dataset allows us to find some interesting phenomena, though recall is still a problem for our method.

[1] Kestemont, M., Karsdorp, F., & During, M. (2014). Mining the twentieth century's history from the time magazine corpus. In Abstract book of EACL 2014: the 14th Conference of the European Chapter of the Association for Computational Linguistics (p. 62).

[2] Marjanen, J., Zosa, E., Hengchen, S., Pivovarova, L., & Tolonen, M. (2020). Topic modelling discourse dynamics in historical newspapers. arXiv preprint arXiv:2011.10428.

[3] Rosenfeld, A., & Erk, K. (2018, June). Deep neural models of semantic shift. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers) (pp. 474-484).

[4] Schlechtweg, D., & Walde, S. S. I. (2020). Simulating Lexical Semantic Change from Sense-Annotated Data. arXiv preprint arXiv:2001.03216.

Day 2: 17 March 2021

Session 1: Digitised Newspapers and Machine Learning: Extraction and Classification of News Items

Challenges in Extraction and Classification of News Articles from Historical Newspapers

Dilawar Ali | Steven Verstockt | *University of Ghent*

Historical newspapers contain huge amounts of information in the form of articles, illustrations, advertisements and old maps. Manual digitization of these news articles consumes a lot of time, cost and effort. Automatic digitization of these historical newspapers is a feasible solution to extract the information in a relatively faster way. For the digitization of newspapers, the traditional approach is to extract the article information based on document layout analysis. But the digitization of newspapers using just a document layout analysis technique is a challenging task because the layout of these newspapers changes with time and does not remain consistent.

In this research study, we will highlight the challenges that can occur while digitizing newspapers using traditional approaches. Furthermore, we will also present a solution to digitize historical news articles by combining deep learning and traditional document layout analysis approaches. We used the METS/ALTO OCR results combined with a simple computer vision technique for article extraction.

Some results are shown below in Figure 1. For illustration classification, we used the image classification technique using both supervised and unsupervised approaches. We achieved good results of the grouping of content, i.e. title/heading, illustration, text and captions that belong to one news article. There are still many challenges that need to be addressed to improve these results, e.g. grouping news patches from different columns, the title/head of news article detection, joining of multiple titles/subtitles of one news article, linking of an illustration to an article, distinguishing between captions or normal text and false/missing or overlapping detections.

It is challenging to link a piece of a news article to news when a news article is distributed in multiple columns and the news heading is only available in the first or two columns as shown in Figure 2. Our final goal is to combine all these tools in an automatic metadata enrichment tool to improve the searchability of historical contents. Clustering methodologies, for example, will be helpful in finding similar pictures and headings based on visual and textual features.

Figure 1: Article extraction results

Figure 2: Challenge in grouping all patches of a news article with title/heading in only first two columns

Narrating Politics: How Journalists Changed the Way They Cover Political News in France

Étienne Ollion | *CNRS/CREST, École polytechnique*
Rubing Shen | *Sciences Po*

Has the coverage of politics evolved in the media and, if yes, why? Several studies have pointed toward a notable shift since the 1980s. Increasingly, politics is described as a 'horse race', a fierce competition between actors of the political field. This classic story is now well-established. 'Politics' – the coverage of the action – is said to have taken precedence over the description of 'policy' – the content of the debates.

There is less consensus is, nevertheless, on the timing as well as on the causes of such a shift. Across countries, competing narratives exist among historians on these two aspects (for France, see Kaciaf, 2013; Saitta, 2005; for the USA, see Patterson, 1993; for a review of this abundant literature, see Aalberg et al., 2011).

The goal of this presentation is to leverage recent NLP tools to offer new insights into this classic question. More specifically, we use deep learning algorithms to automatically annotate a large corpus of print media from 1950 to the present day. To do so, we first train several classifiers on selected datasets before we assess the quality of their annotations (over 80%: i.e. equal to a human annotator).

We apply them to a corpus of French daily newspapers. Articles from the French newspaper *Le Monde* (since 1945), along with *Libération* and *Le Figaro* (since 1995) are analyzed in detail. Overall, 174,577 articles were automatically annotated.

After a brief presentation of the research question, the paper will introduce the method. It will discuss its merits, its limits and its pitfalls. It will then move on to present the most significant results. Using a set of varied indicators, our research confirms to revisit the classic narrative: the standard account of politics did change drastically, and by the early 2000s politics was predominantly presented as a 'horse race'. Depending on the type of indicator, though, the timing differs, a fact that leads us to nuance and specify the established narratives on the topic.
We then turn to the second question: that of the causes. Our goal is to disentangle several competing explanations, among which that of a generational effect, of a rising competition between newspapers for a narrowing audience and of the rise of a new writing canon borrowed from investigative journalism. Supplementing our annotated data with information about sales and about the career of journalists, we offer some tentative conclusions about the rise of horse race journalism. We conclude restating the potential, and the potential issues, of using NLP for doing historical research.

Aalberg, T., Strömbäck, J., de Vreese, C. H. (2011). The framing of politics as strategy and game: A review of concepts, operationalizations and key findings. *Journalism: Theory, Practice & Criticism*, *13*(2), 162-178.
Kaciaf, N. (2013). *Les pages « Politique »: Histoire du journalisme politique dans la presse française (1945-2006).* Presses universitaires de Rennes.
Patterson, T. E. (1993). *Out of Order*. Vintage Books.
Saitta, E. (2005). Le monde, vingt ans après. *Réseaux*, *131*(3), pp. 189-225.

Semantic Segmentation and Document Layout Recognition - Approaches to Full Text Recognition of Early Chinese Newspapers

Matthias Arnold | *Heidelberg University*

The use of convolutional neural networks in digitizing historical documents has drastically expanded the quality and scope of available sources for digital analysis.[1] The ability to reuse or to refine pre-trained models means that DH practitioners are now tackling sources previously deemed impossible to process automatically.

In light of these developments, we wish to present our own work on adapting tools developed for use with Western materials to the complex layouts of republican Chinese newspapers from the Early Chinese Periodicals Online (ECPO) project.[2] Our results demonstrate how cultural biases towards Latinized scripts and layouts affects models and algorithms. To overcome these shortcomings, we conducted experiments with crowd-sourcing, pattern recognition and machine-learning. We aim to provide a development workflow from image scan to machine-readable full-text, and started to publish our ground truths for re-use.[3]

Because of the dense document layout, sub-optimal image scans, special characters and varying reading directions, current OCR engines still fail to process full pages. A literature survey on Chinese approaches to the material showed that the large libraries mostly create author-title indexes for their systems.[4] We therefore decided to first create ground truth ourselves. We implemented an annotation tool to generate bounding boxes and created semantic groups (e.g. all parts an article consists of). We also applied four high-level labels (advertisement, article, image and marginalia) to boxes and groups, and trained a neural network to detect the respective areas on the pages across a larger data set. In addition, we created a double-blind keying workflow to create high-quality full-text. In a next step we will use this second ground truth data to train the OCR engine and improve recognition for Republican China newspaper texts.

[1] Bernhard Liebl and Manuel Burghardt, 'An Evaluation of DNN Architectures for Page Segmentation of Historical Newspapers,' April 15, 2020, https://arxiv.org.
Sofia Ares Oliveira, Benoit Seguin, and Frederic Kaplan, 'DhSegment: A Generic Deep-Learning Approach for Document Segmentation,' in *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)* (IEEE, 2018), 7–12.
[2] Early Chinese Periodicals Online (ECPO), https://uni-heidelberg.de/ecpo.
[3] Matthias Arnold and Lena Hessel, 'Transforming Data Silos into Knowledge: Early Chinese Periodicals Online (ECPO),' in *Heuveline, Vincent , Gebhart, Fabian Und Mohammadianbisheh, Nina (Hrsg.): E-Science-Tage 2019: Data to Knowledge* (Heidelberg: heiBOOKS, 2020), 95–109, https://doi.org/10.11588/heibooks.598.c8420; Matthias Arnold, 'Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition,' *Digital Studies / Le Champ Numérique* 11 (forthcoming).
[4] A publication about our systematic literature survey is currently in preparation. For an introduction see Fang Zijin 方自金, 'Status Quo, Problems and Suggestions of the Photocopying and Publishing of Newspapers in the Republic of China' 民国报纸影印出版的现状、问题与建议, *Chuban cankao* 出版参考 2019, no. 6 (2019), www.sohu.com. Also, cf. Yang Jiaying 杨佳颖 and Xu Xin 许鑫, 'Semantic Labeling of Advertising Images in Newspapers of the Republic of China Period: Illustrated by Shaoxing Opera Advertisement Published in Xinwen Bao' 民国报纸广告图像资源的语义标注——以《新闻报》所刊的越剧广告为例, *Library Journal* 图书馆杂志 online first (30 June 2020): 1–11. and Xiao Hong 肖红 and Huai Yan 槐燕, 'An Analysis of Quality Checking Problems in the Practice of Digitization of Newspapers of Republic of China' 民国报纸数字化实践中的质检问题探析, *Tushuguan xue yanjiu* 图书馆学研究, no. 07 (2017): 61-78+87.

Ares Oliveira, Sofia, Benoit Seguin, and Frederic Kaplan. 'DhSegment: A Generic Deep-Learning Approach for Document Segmentation.' In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, 7–12. IEEE, 2018. https://doi.org/10.1109/ICFHR-2018.2018.00011.

Arnold, Matthias. 'Multilingual Research Projects: Challenges for Making Use of Standards, Authority Files, and Character Recognition.' *Digital Studies / Le Champ Numérique* 11 (2021).

Arnold, Matthias, and Lena Hessel. 'Transforming Data Silos into Knowledge: Early Chinese Periodicals Online (ECPO).' In *Heuveline, Vincent , Gebhart, Fabian Und Mohammadianbisheh, Nina (Hrsg.): E-Science-Tage 2019: Data to Knowledge*, 95–109. Heidelberg: heiBOOKS, 2020. https://doi.org/10.11588/heibooks.598.c8420.

European Research Council (ERC). 'Guidelines on Implementation of Open Access to Scientific Publications and Research Data,' April 21, 2017. https://ec.europa.eu/research/participants/data/ref/h2020/other/hi/oa-pilot/h2020-hi-erc-oa-guide_en.pdf.

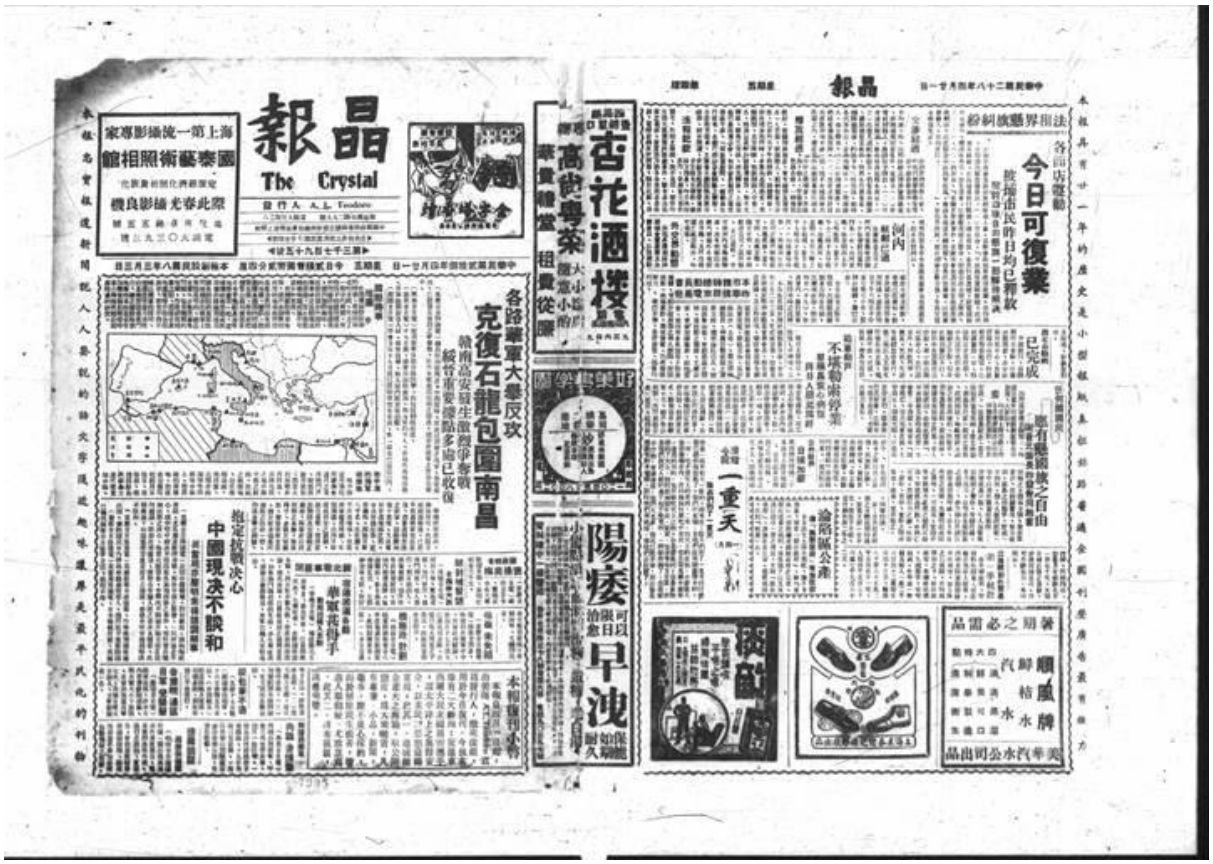

Fig. 1: A typical complex newspaper page layout: Jing bao **晶報** (The Crystal), April 21, 1939, pages 1+4. In ECPO: https://kjc-sv034.kjc.uni-heidelberg.de/ecpo/publications.php?magid=1&isid=20&ispage=1.

Fig. 2: The page fully manually annotated from our ground truth set (orange = article, blue = image, green = advertisement, purple = header and marginalia). Jing bao **晶報** (The Crystal), April 21, 1939, pages 1+4.

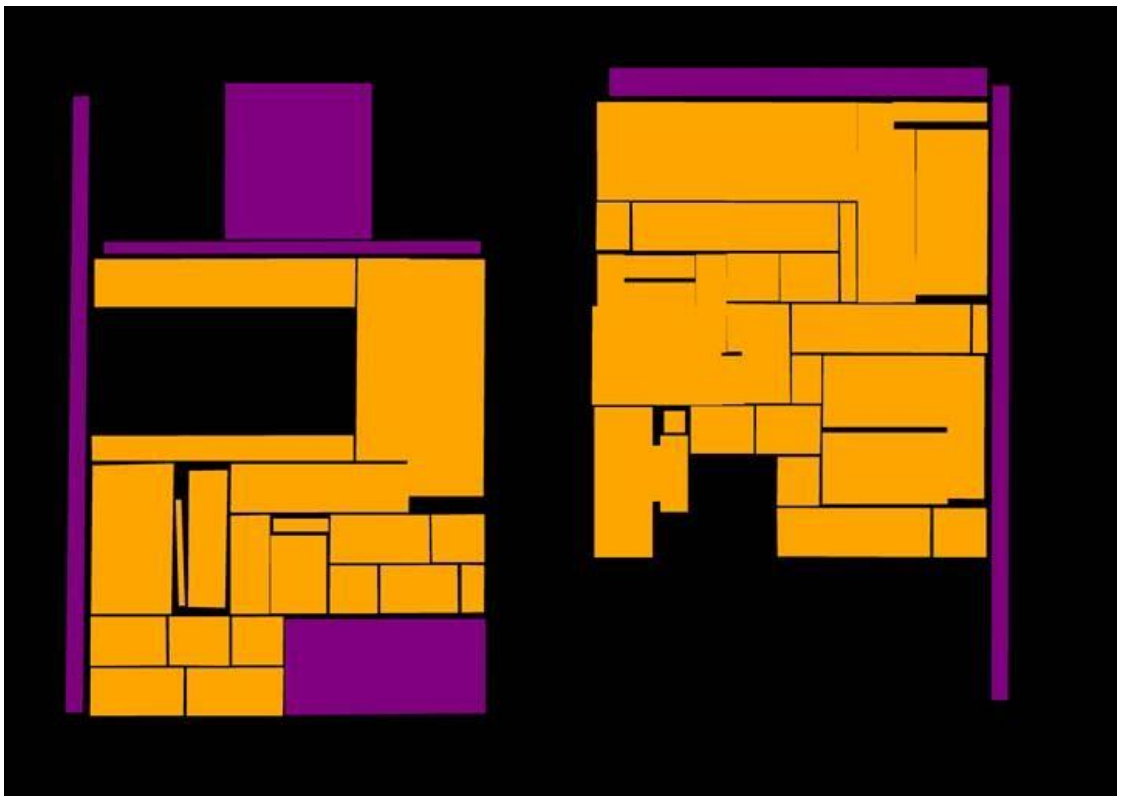

Fig. 3: Detection of content types using dhSegment: advertisements and images ignored, orange = text (i. e. "articles"), purple = header and marginalia. Jing bao **晶報** (The Crystal), April 21, 1939, pages 1+4.

The Dublin Gazette

David Brown | *Beyond 2022, Trinity College Dublin*

The *Dublin Gazette* is the record of official Ireland, published occasionally from the late seventeenth to early eighteenth centuries, and continuously from 1711 to 1922. No single complete run survives. *Beyond 2022*, in collaboration with the Oireachtas Library, the Library of the Honorable Society of King's Inns, the Free Library Company of Philadelphia, the Library of Trinity College Dublin and the National Library of Ireland, has reassembled the *Gazette* to once again deliver the official news, three times a week for two centuries.

To produce tens of thousands of issues, several printers were used, but the format remained remarkably consistent. Even at the end of British rule and the establishment of the Irish Free State in 1922 – a moment of major constitutional upheaval – the same printer continued to produce the same sort of content in the usual two-column format, simply replacing the royal crest with the new title 'Iris Oifigiúil' (Official Gazette, or Magazine) in Gaelic typeface. The subtitle published by authority never changed. *Iris Oifigiúil* continues to be published today, online.

The Public Record Office of Ireland was destroyed in 1922, and with it most records of British rule. The *Dublin Gazette*, therefore, has unique significance as a source for Irish historical records. The journal routinely published most announcements of an official nature, including statutes, proclamations, appointments, military affairs, petitions and judgments of the courts. A digital edition of the *Dublin Gazette* enables many thousands of the original records lost in the destruction of the PROI to be replaced with these substitutes, and new domains of historical research to be opened up for the first time.

Beyond 2022 is a major historical research project based at Trinity College Dublin that aims to reconstruct, in virtual form, the Public Record Office of Ireland. *Beyond 2022* is an international collaborative project funded by the Government of Ireland, whose core partners include the National Archives of Ireland and the UK and the Public Record Office of Northern Ireland. There are currently more than 40 additional archival partners, including 10 in the USA where many Irish records have been acquired by research libraries and other institutions over the years. Transkribus is used within the project as a core technology to make the digital assets provided by numerous archival partners cross-searchable. The project is also developing an NLP pipeline, an RDF framework for entities identified in the texts, an immersive 3D reconstruction of the original record office and a digital assistant that will interact with users using text and speech.

The *Dublin Gazette* was never intended for long term preservation and both print and paper quality can be quite poor. To overcome these shortcomings, HTR models have been created for four separate typefaces, representing the main presses responsible for printing the newspaper over the years. These HTR models produce excellent results from early newsprint, all of which is in English. The *Gazette* follows a columnar layout throughout its run, making it ideal for an NLP approach, as sections are normally defined both clearly and consistently. For example, Proclamations always have this simple heading, military appointments are always below the heading 'Dublin Castle' and movements of shipping will always be found under 'Port News'.

The importance of the *Dublin Gazette* as a unique source for Irish history cannot be overstated. In most jurisdictions, the historical newspapers are complementary to state archives and testamentary records. In Ireland, this newspaper is a surrogate for all of these.

Session 2: Digitised Historical Newspapers: Working with Classified News Items and Information Extraction

The Incredibly Differentiating Labor Market: Evidence from Job Offers

Raven Adam | Saranya Balasubramanian |Vera Maria Charvat | Manfred Füllsack | Jörn Kleinert |Hanna Misera |Nenad Pantelic | Jakob Sonnberger |Georg Vogeler|*University of Graz & The Austrian Center for Digital Humanities and Cultural Heritage*

The ANNO collection of the Austrian National Library contains a comprehensive set of digitized Austrian Newspapers from the years between 1850 and 1949. Using this, we shed new light on the change in labor relationships over one hundred years. We aim for three goals: Firstly, we intend to generate a unique data source for empirical analyses in various fields of humanities, assuming that job advertisements, particularly classified advertisements in newspapers, are a great source of individuals' wills, wishes and offers widely unfiltered by others. Secondly, we want to illustrate the strong change in labor relationships in many dimensions such as extent, regional reach, sector focus, skill requirements, job characteristics, self-perceptions of employers and employees, as well as expected and/or pretended relationships. We expect to detect significant changes in these dimensions over the long-time horizon. Thirdly, we want to test and further develop digital methods for text mining and text analyses on a database of millions of advertisements with fairly heterogeneous content and structure.

The core of our dataset consists of 35 daily newspapers with regional or national reach and long print run in the time period from 1850 to 1949 containing more than 8 million pages. The papers hold job ads in almost every issue but have often particularly many ads on weekends. While a large fraction of the ads is concentrated in special sections of classified advertisements, we find advertisements and job offers or requests in more unstructured parts of the papers as well, necessitating a strong focus on data preprocessing.

For this, the relevant sections from the newspapers are extracted and transformed into machine-readable formats by experimenting with different extraction tools. After correcting for OCR miss-reads and other kinds of noise, texts are tokenized and applied to various analysis methods ranging from frequency and co-occurrence analyses to modes of topic-modeling (SLA, DLA, Doc2Vec, Top2Vec) for extracting semantic aspects. This allows different perspectives on the ads. They can be viewed in isolation, seen as a cohort, or put into a thematic cluster or in a more hierarchical structure. Specific categories, catch words or phrases and the structure of the ad or specific sentences can be analyzed, as well as the tone and transported image of the job offer or request. Absolute and relative frequencies of words, terms and phrases can be compared and put into a time dimension for analyzing job attributes and employment demand. Different methods in this regard shall allow for different cuts through the data.

The labor market example is chosen because changes in the labor market are very pronounced and several changes over the 100 years can be detected. The number of observations is high enough to disentangle changes in the size of industrial labor from the changes in using newspaper ads as a means of facilitating the relationship. We consider the development of newspaper job ads and job requests as a very specific reaction to the differentiation of industrial production.

London Calling: A Methodological Quest throughout Eighteenth-Century London Auction Advertisements

Alessandra De Mulder | *University of Antwerp*

Word embeddings have been proven to achieve interesting results for historical linguistic research and digitised sources such as books and newspapers. The methodological paper at hand explores the possibilities of this DH methodology for the empirical examination of auction advertisements in eighteenth-century London. The central argument is that few eighteenth-century corpora give as many opportunities to apply, test and refine digital methods as digitised newspapers. Word embeddings are extremely useful for first explorations of sources, such as advertisements, and, by extension, newspapers, by providing insight into the average structure, as well as a detailed content analysis of digitised big data sets. Moreover, the paper will show that word embeddings, when combined with a more traditional approach to historical sources, might reveal broader societal changes, developments and patterns that would otherwise escape the historian's gaze.

The United Kingdom, and more precisely London, has been the breeding ground of an extensive historiography surrounding eighteenth-century consumer culture for its central role due to the political turmoil in France and other evolutions such as the transatlantic trade flows of commodities such as coffee. This not only led to an explosion of texts by great thinkers about the possible negative consequences of consumption, but it also meant that London was at the forefront of many retail and taste innovations. However, we urgently need to move beyond well-known Enlightenment discourses and discover anew the day-to-day commercial language used by more common eighteenth-century retailers and salesmen themselves. The mindset of the eighteenth-century consumer will be approached through the multiple and often subtle semantic shifts in referencing the quality of goods in auction advertisements. As I will demonstrate, these references are indispensable to expose and de-construct the way sellers implemented linguistic strategies to attract buyers.

I will contextualise a word embedding analysis of over 5,000 pages of eighteenth-century auction advertisements by performing a close reading of the sources to find motivational drivers revealing eighteenth-century consumption patterns. The auction advertisements will be compared to additional material that will mainly consist of dictionaries, philosophical treatises and furniture manuals. I will propose further fine tuning through other DH methods to illuminate and investigate the impressive but still rather general overview the word embeddings provide. Tracing the pivotal words and stock phrases used to sell products and contextualizing them within auctions advertisements and by extension newspapers as well as contemporary literature will lead to fresh insights into the cultural history of values in eighteenth-century Britain.

Translocalis - A Database of Readers' Letters Published in the 19th Century Finnish-language Press

Heikki Kokko | *Tampere University*

Translocalis is a digital database of readers' letters sent to Finnish-language press in the middle of the 19th century. The database contains all the readers' letters written in the name of local communities and published in the Finnish-language press from the period of 1850–1875, gathered from the fully digitized newspaper collection of the National Library of Finland. A nationwide phenomenon of readers' letters to newspapers developed during this first phase of modernisation. The characteristic of this phenomenon was that the letters were often written in the name of the local community. Thousands of Finnish-speaking people wrote about their experiences and sent letters to newspapers from the 1850s onwards. The writers of the letters came from different strata of society. There were a lot of ordinary people – peasants, crofters, and workers – who wrote letters. The letters were often written behind the shield of anonymity, which enabled the challenging of the power structures in the old society of estates.

The Translocalis database is the digital history project of the Finnish Academy Centre of Excellence in the History of the Experiences (HEX) to collect up this cultural heritage that had largely been unnoticed by historical research before the digital era. The first phase of the project (since 2017) has been the collecting of the newspaper clippings from the other newspaper material with the tool provided by the National Library of Finland. Currently, there are about 27,000 readers' letters collected from the era of 1850-1875. As the material has been collected from the press with this digital tool, the database includes the letters with the name of the newspaper or journal in which they were published and the publication date. The database is fully optical OCR-recognised; thus, it enables the use of the methods of Digital Humanities (DH). During the manual collection work, the letters have been enriched with metadata, such as the place of writing and the named writers of the letters. Currently, about 28 percent of the letters are identified with the full name of the writer and 22 percent of the letters with the occupational title.

The phenomenon of readers' letters written in the name of local communities continued until the first decades of the 20th century. However, the manual collection of the material from the 1880s onwards is difficult, because the text volume of the Finnish-language newspapers and journals increased manifold. Therefore, the second phase of the project will be the finding and collecting of these letters from the newspaper material by the methods of computer and data science. The premise is to use the hand-picked data as sample data in this process. I am currently searching for computer and data science collaboration partners to carry out this task. We plan to publish the database online with a user-friendly interface.

Because of the geographical and societal representativeness of the culture of readers' letters in Finland, the database contains valuable and rare grass-roots experiences from the interface of modernity. This could significantly contribute to the research of civic society, nationalism and the social history of media, among other research fields. To see more, visit: https://research.tuni.fi/hex/https://research.tuni.fi/hexhttps://research.tuni.fi/hex/https:/research.tuni.fi/hex.

The DIGITARIUM as a Research Corpus: New Approaches to Extracting and Linking Named Entities from Historical Newspapers

Claudia Resch |Matthias Schlögl |Nina C. Rastinger | Dario Kampkaspar |*Austrian Centre for Digital Humanities and Cultural Heritage*

For the last decade, researchers and those responsible for archives or libraries have been cooperating intensively with IT experts and digital humanists to advance the digitisation of historical newspapers. Funding initiatives have been pushing for the digital provision of images and full texts, which has significantly increased the number of pages available to researchers. This brings along a decisive extension of possibilities when doing research on digitised newspaper collections.

Based on a concrete newspaper full text digitization project hosted at the Austrian Centre for Digital Humanities and Cultural Heritage, this paper will discuss possible contributions of smaller collections to the development and assessment of new methods and tools. Our ongoing project focusses on one of the oldest newspapers in the world still published today, namely the *Wiener Zeitung*, formerly called *Wien[n]erisches Diarium*, which was founded in 1703. This periodical is a vast resource for extracting information about persons, places, institutions and events. In the project, it serves as a research corpus for (semi)automatically generating structured data from historical newspaper articles. The full-text editions of a selection of more than 300 issues have already been published in a former project and are now available in the prototype "DIGITARIUM" (Resch & Kampkaspar 2020): https://digitarium.acdh.oeaw.ac.at.

After a short introduction into this research corpus, we will concentrate on the lists of the historical *Wiener Zeitung*, especially on the items of obituaries; these lists of deceased persons (inside and outside the city centre of Vienna) contain detailed information on the name, occupation, place of death, age and optionally also title, marital status and cause of death, which will be systematically extracted from the full text and then enriched with other existing sources.

While the extraction of named entities from contemporary articles is already relatively reliable, models suited for historical newspapers are scarce and, if available at all, much less precise. In the case of 18th century obituaries, we therefore make use of their relatively consistent form: First, a basal Named Entity Recognizer is created using some rules to generate suggestions for speeding up the process of annotating training data. This data is used to train a model, which is iteratively improved using Active Learning and evaluated with gold standard data. In a further step, we will test possibilities to externally link the extracted location data. As a corresponding data set, we therefore will use the so-called 'Vienna History Wiki' because it contains a large number of entries on Vienna's city history, as well as maps and further biographical references.

The exemplary data from the *Wiener Zeitung* and our experimental approach will be used to assess the extent to which topographical information can be automatically obtained, combined and reused. The goal of this paper is to explore how digital collections of historical newspapers can be further enriched using data from other digital portals, data collections and knowledge resources to deepen our understanding of the past.

Examining a Multi-layered Approach for the Classification of OCR quality without Ground Truth

Mirjam Cuper | *National Library of the Netherlands*

In the past few decades, more and more heritage institutions have made their collections digitally available. At the same time, the availability of computer driven research tasks is increasing. With this combination, large data sets can be analysed in a fairly short time, which would not be possible by hand. However, there is a pitfall: the Optical Character Recognition (OCR) quality of digitised text is not always of high quality. This leads to several possible problems which can cause bias in the research results, both on the information retrieval and the analysis level. Although most researchers are aware of the presence of OCR errors, often they are not able to quantify these errors or the impact on their research. This leads to uncertainty about whether results can be published or not (Traub, Van Ossenbruggen, & Hardman, 2015). A measure for the (relative) quality of OCR could be very beneficial for the field of Digital Humanities.

For a few collections, a so-called 'Ground Truth' set is available. A Ground Truth set consists of digitised texts that are manually corrected by humans. Therefore, these digitised texts are of high quality and contain little to no errors. They can also be used to determine the quality of the corresponding OCR output. However, the creation of a Ground Truth set is time consuming and expensive, resulting in only a small amount of available Ground Truth sets. The above leads to the following question: How can the quality of OCR-ed texts be determined when Ground Truth data is absent?

This study examines a multi-layered approach for measuring the OCR quality of digitised texts without the presence of Ground Truth. Previous research has mentioned a few possibilities, such as garbage detection (Taghva, Nartker, Condit, & Borsack, 2001; Wudtke, Ringlstetter, & Schulz, 2011), lexicality, and confidence values from the OCR engine (Springmann, Fink, & Schulz, 2016). All of these measures have problems with their accuracy due to the nature of texts. Therefore, we propose a multi-layered approach that combines various different measurements. If relevant, the time period or the categories of a text are taken into account for the criteria of each measurement. Every measurement has its own weight, which determines how much it affects the outcome. After combining the outputs of the different measurements based on their weights, one categorical quality measure is suggested, such as 'weak', 'medium' or 'high' quality.

To substantiate our approach, we will use a data set for which we have the Ground Truth available. First, we measure the quality with the use of the Ground Truth. Then, we compare this result with the predicted quality measure from our proposed approach and determine the precision and recall. Our approach will be tested on a collection of Dutch historical newspapers and books from the 17th century until the 20th century. The total set consists of ~36,000 newspaper articles and 2,055 book pages.

Springmann, U., Fink, F., & Schulz, K. U. (2016). *Automatic quality evaluation and (semi-) automatic improvement of OCR models for historical printings*. arXiv preprint arXiv:1606.05157.

Taghva, K., Nartker, T., Condit, A., & Borsack, J. (2001). *Automatic removal of 'garbage strings' in OCR text: An implementation*. Paper presented at the The 5th World Multi-Conference on Systemics, Cybernetics and Informatics.

Traub, M. C., Van Ossenbruggen, J., & Hardman, L. (2015). *Impact analysis of OCR quality on research tasks in digital archives*. Paper presented at the International Conference on Theory and Practice of Digital Libraries.

Wudtke, R., Ringlstetter, C., & Schulz, K. U. (2011). *Recognizing garbage in OCR output on historical documents*. Paper presented at the Proceedings of the 2011 Joint Workshop on Multilingual OCR and Analytics for Noisy Unstructured Text Data.

Discovering Spatial Relations in Literature: What is the Influence of OCR Noise?

Motasem Alrahabi | Gaël Lejeune |Caroline Parfait | Glenn Roe | *Sorbonne University*

Digital Humanities methods enable the exploration and exploitation of digitized corpora at unprecedented scales. They also allow for refined research at several levels of granularity, from syntactic or hermeneutic perspectives, or through the identification of geographical named-entities, which allows us to observe the evolution of language and its territorial distribution. However, there are notable limitations in the performance of Named Entities Recognition tools for humanities research due to the variability of the input data (linguistic, diachronic, diatopic variability). Moreover, this lack of robustness to variation is particularly striking when dealing with literary corpora, even more so when it involves early modern texts.

The correct recognition of named entities is correlated with the training of the language model implemented in the NER system. Language models are usually trained on so-called 'clean data' – assembled under optimal laboratory conditions – and for application to a specific corpus, which thus limits their generalizability to other corpora. Moreover, language models for early modern texts often require access to large corpora which have previously been transcribed using OCR. The quality of these transcriptions remains the subject of many current research projects [Baledent et al., 2020]. In essence, the malfunctioning of NER tools is attributed, on the one hand, to the level of quality of the transcriptions provided as input and, on the other hand, to the fact that the corpus being processed does not correspond to the corpus on which the language model was trained. To overcome the problem related to the quality of OCR transcripts, users implement a strategy that is costly both in terms of time and finances, consisting of cleaning the transcribed text.

Indeed, any number of errors can exist in OCR transcriptions [Stanislawek et al., 2019] and this search for perfection, though perhaps feasible on very small corpus, can be never-ending and represents a considerable expenditure of time at larger scales. Our project seeks to evaluate out-of-the-box NER tools, specifically Spacy, on minimally-corrected OCR transcriptions. This experiment should allow us to see the capacity of these tools to do their work outside of ideal laboratory conditions, aiming to get closer to a more everyday use of these tools, i.e. as a user who has neither the time, nor money for corrections, but nevertheless seeks actionable results. By way of this tension between ideality and reality, we have eschewed for the moment any ground-truth, which is costly to produce. Nevertheless, we use what we consider to be a reference text.

The reference texts are extracted from ELTeC, a multilingual European Literary Text Collection in which entire novels are available in standardized versions. The texts we use in hypothesis-testing consist of the OCR transcription of the same texts, downloaded in PDF format from the Gallica website. The first novel on which we focus is Marguerite Audoux's *Marie-Claire* (1910), a novel of about 34,500 words. We carried out initial tests on short text extracts of about 200 words and found that the pre-trained Spacy models are capable of recognising a number of terms even when roughly transcribed by the OCR tool. The 'fr core news sm' model finds 79% of entities present in both the reference and the hypothesis text, and 12.5% of entities which are incorrectly spelled in the hypothesis text.

Evaluating the Multilingual Capabilities of PERO-OCR with Digitised Historical Newspapers: A Belgian Case Study

Julie M. Birkholz | Sally Chambers | *Ghent Centre for Digital Humanities and KBR – the Royal Library of Belgium*
Michal Hradiš | Pavel Smrz | *Brno University of Technology*

Historical newspapers are increasingly being analyzed in humanities research and thus digitized by cultural heritage institutions. In this process of digitization and OCRing, automatic classification, such as article segmentation, and full text indexing of the extracted text is completed most often with a language model to theoretically increase the quality of the results. In the case of multilingual sources, and further automating this process for large collections over different time periods, the efficiency of this approach is put to question. In this presentation, we will present results of the OCCAM (OCR, ClassificAtion & Machine Translation) project's digital humanities case which implements a pipeline utilizing PERO-OCR, resulting in high-quality OCR of multiple languages of historical newspapers.

PERO is a novel, learning-based, fully adaptable and customizable open-source recognition engine that aims to improve accessibility of digitized historic documents. It is based on state-of-the-art methods from computer vision, machine learning (esp. deep neural networks), and language modeling. PERO extends automation and capabilities of the digitization pipeline by providing tools for automated quality assessment and control, quality improvement, automated text transcription of historic printed documents, semi-automated handwritten text transcription, and automatic extraction of semantic information from semi-structured documents. Particular attention is paid to low-quality historic printed and handwritten documents that cannot be automatically processed by the currently available tools.

We explain how images of textual sources in multiple languages can efficiently be OCRed using a machine learning based model, enabling both the annotation of corrections and automatic text recognition of a set of multilingual historical newspapers from KBR - the Royal Library of Belgium's historical newspaper collections: BelgicaPress (Figure 1.). Through examples from a set of Dutch and French language newspapers from the early 1900s, we present the results obtained using PERO-OCR and compare this with the original OCR. This results in high-quality OCR, that is flexible to diverse layouts of historical newspapers of varying quality (Figure 2.), with adaptation needed for line detection / layout identification (Figure 3.); and various combinations of printed and handwritten text (e.g. signatures in newspapers). In generating a PAGE XML format, the PERO-OCR platform also affords an adaptable output for post-processing. This suggests that a general model in comparison to the novel learning-based recognition engine of PERO-OCR results in a flexible and adaptable tool for full text results for large multilingual collections.

Figure 1. Multilingual example from within one newspaper article of *De Standaard* from 1919

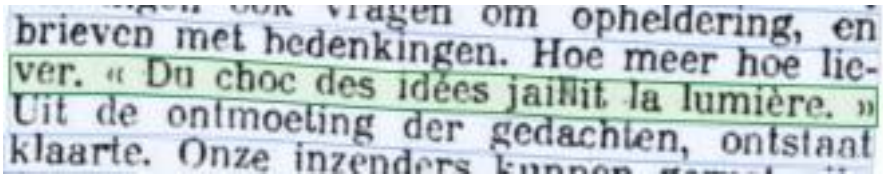

Figure 2. Signaling potential errors due to image quality issues

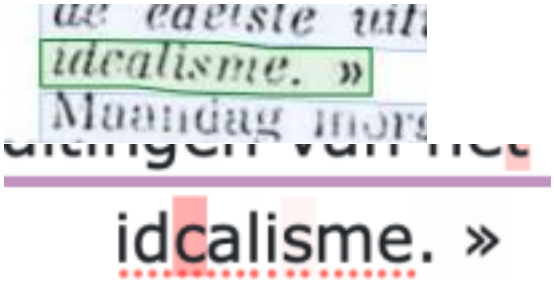

Figure 3. Layout identification

Two Examples of Analysis of Textual Documents in Oriental and Under-Resourced Languages

Chahan Vidal-Gorène | *École Nationale des Chartes*

Digital Humanities are more and more included in projects for the digitization and preservation of collections of heritage conservation institutions, notably through handwritten and printed character recognition, but also through text analysis technologies (e.g. named-entity recognition). The implementation of these technologies, now essentially based on artificial intelligence, pursues two main objectives: (1) increase a document's accessibility by integrating it into several searchable and on a like-for-like basis databases (e.g. in-text research for information in a manuscript page, etc.) and (2) accelerate the research work by providing tools for analysis and automatic comparison. These technologies are subject to the creation and existence of large interoperable databases for the exchange of information (e.g. IIIF protocol) and for the customized training of dedicated intelligent systems (e.g. the Europeana Newspapers project).

The issue of data access (for the public and for researchers) arises in the case of digitally under- resourced and/or non-Latin scripts languages, for which such databases (whose creation process is long and expensive) do not exist yet or for which existing tools do not fit well. These languages may also suffer from the lack of specialists who could carry out the preliminary annotation work. Our presentation will, therefore, focus on the automatic analysis of textual documents for oriental and under-resourced languages (Armenian, Georgian, Arabic and Syriac). Through the introduction to two ongoing digitization projects, namely the catalog of Armenian manuscripts in Venice (Calfa-Congrégation des Pères Mekhitaristes) and the digitization of manuscripts in Maghrebi script (Calfa-BULAC), we will evaluate the methodologies and tools implemented for the constitution of dedicated databases.

These two projects are modest in size (5,000 pages and 800 pages respectively) but offer many various difficulties (page layout, state of conservation, writing). For each project, we have favored the creation of customized training databases (both in the analysis of areas of interest in the documents and in the extraction of the text) in order to maximize the quality of the results produced and thus limit the post-processing work generally necessary in this kind of project. This relevant methodological choice implies nevertheless to be capable to quickly build the necessary data to be efficient, despite the small size of the corpuses considered. In view of the documents concerned, we have developed the Calfa Vision tool, which combines public involvement (crowdsourcing) and automatic annotation. The presentation is intended to be a feedback on the automatic processing of documents in under-resourced languages.

Session 4: Challenges and Perspectives in Digital Research: Workflows, Pipelines, Digital Libraries and Digital Literacy

Challenges for Digital Literacy in the Humanities: The Open, Community-Based and Multilinguistic Approach of The Programming Historian

Sofia Papastamkou | CNRS / IRHiS, University of Lille
Jessica Parr | Simmons University
Riva Quiroga | Pontifical Catholic University of Chile

Introduction

How do researchers in the humanities perceive the need to seize digital methods and code literacy, and how can they act towards its fulfillment? An open access, community-based and multilingual journal of tutorials on digital methods, The Programming Historian can be viewed as such a response. Its aim is to help humanists become familiar with a wide range of digital tools, techniques, and workflows to facilitate research and teaching.

It currently exists in three languages and proposes 84 lessons in English, 47 in Spanish, and 15 in French. From 1st January to 13 December 2020, some 1.5 million unique users around the globe visited the website with 2.4 millions of pages viewed.

Open Values

The Programming Historian was initially founded in 2007 in English, and became an open access peer reviewed journal in 2012. Four years later, linguistic sub-teams started to join the project to create Spanish (2016), French (2018) and Portuguese (2020) speaking versions. The Programming Historian proudly advocates open values; it is open access under the CC BY 4.0 licence and relies heavily on and promotes open software and programming languages. Due to these choices and, furthermore, its uses, The Programming Historian appears plainly as an open educational resource following UNESCO's definition (Rojas Castro et al. 2019). One of its core values is open peer review. This commitment is illustrated, and further enhanced, by the choice of its editorial and publishing infrastructure (GitHub platform) that allows it to be widely open to the community (Walsh 2021).

A Community (of Practice)-Based Approach

The Programming Historian is run by a team of editors who are mostly volunteers. Infrastructure building and maintenance, as well as administration, fund fundraising and global development are also assured by sub-teams that are part of the broader team. Beyond editing and related activities, authorship, peer review and translation of the lessons are assured by a community of collaborators who are all credited for their contribution. Correspondingly, The Programming Historian appears as what is defined as a community of practice (Wenger 1998) in the Digital Humanities trading zone and, proudly historian-oriented, can be viewed as one of Digital History's beating hearts (Clavert and Schafer 2019). The lessons are mostly case-studies that teach how to acquire, transform, analyze, present and sustain humanities research data using programming languages, mostly Python and R, or open source software and workflows. As such, they correspond to the needs of historians and humanists conducting in the field. They are used both for self-training and teaching in the classroom (Crymble 2018, Survey 2018).

Multilingualism

Publishing in more languages, in addition to English, since 2017, as well as ad hoc (not hosted on The Programming Historian website) translations in various languages thanks to open CC-BY licensing have actually been important steps towards internationalization in accordance with the journal's diversity policy (Sichani et al. 2019). Multilingualism seems to correspond to actual needs of global humanities

communities. For example, some 25% of the website's visits originate from Spanish speaking countries and five of them are among the top ten of users' countries (Isasi 2019). Furthermore, 2021 will see the launch of a lusophone version. The creation of different linguistic versions did not necessarily follow the same pattern. However, in all cases it highlights various and important challenges such as the development and maintenance of a flexible and evolving infrastructure (Lincoln 2020a, Lincoln 2020b); the need to develop original lessons in the other languages of The Programming Historian to fulfill different needs of the linguistic communities (Isasi and Rojas Castro 2019); or indeed the need to recruit editors with linguistic skills and multicultural acquaintance to assure infra-team and trans-subteams effective collaboration. However, one big challenge The Programming Historian faces is how to include under-represented communities in the global DH community and further contribute to digital and code literacy through enrichment with needs and approaches of culturally diverse environments.

Frédéric Clavert, Valérie Schafer (2019), 'Les humanités numériques, un enjeu historique', Quaderni, 98, DOI : 10.4000/quaderni.1417.

Adam Crymble (2018), "Uses of the Programming Historian", The Programming Historian Blog, https://programminghistorian.org/posts/Uses-Of-TheProgramming-Historian.

Jennifer Isasi (2019), "End of the Year 2019 Newsletter", The Programming Historian Blog, https://programminghistorian.org/posts/newsletter.

Jennifer Isasi and Antonio Rojas Castro (2019), 'Programming Historian en español: De la traducción a la creación de recursos educativos abiertos', Humanidades Digitales Hispánicas 2019, Toledo, Spain.

Matthew Lincoln (2020a), 'Multilingual Jekyll: How The Programming Historian Does That', Matthew Lincoln Website, https://matthewlincoln.net/2020/03/01/multilingual-jekyll.html.

Matthew Lincoln (2020b), "DOIs Added to All Lessons", The Programming Historian Blog, https://programminghistorian.org/posts/dois-for-phhttps://programminghistorian.org/posts/dois-for-ph.

Antonio Rojas Castro, Sofia Papastamkou, and Anna-Maria Sichani (2019), 'Three Challenges in Developing Open Multilingual DH Educational Resources: The Case of The Programming Historian', Accelerating DH Education Workshop, Digital Humanities 2019, Utrecht, Pays-Bas. http://doi.org/10.5281/zenodo.3384956.

Anna-Maria Sichani, James Baker, Maria José Afanador Llach, and Brandon Walsh (2019), 'Diversity and Inclusion in Digital Scholarship and Pedagogy: The Case of The Programming Historian', Insights 32 (1): 16 DOI: http://doi.org/10.1629/uksg.465.

Brandon Walsh (2021), 'The Programming Historian and Editorial Process in Digital Publishing', Walsh Brandon Website, http://walshbr.com/blog/the-programming-historian-and-editorial-process-in-digital-publishing.

Etienne Wenger (1998), *Communities of Practice. Learning, Meaning, and Identity*, Cambridge: Cambridge University Press.

The Programming Historian Community Survey, (unpublished) (2018).

Towards a Datafied Digital Library of Premodern Chinese

Donald Sturgeon | *Durham University*

As invaluable sources of information about the historical record, historical primary source materials have naturally been prime targets for digitization in many domains. While digital transcription of text directly enables many types of automated processing such as full-text search and natural language processing, concrete connections between the surface meaning of the text and the historical record often rely on a further provision of information about the meanings and referents of terms in the text, for example through the creation of annotated texts containing additional layers of data describing the referents of named entities such as persons, places, dates, institutions, and titles.

Accurate annotated editions of texts are typically expensive to create. The availability of appropriate knowledge bases such as historical gazetteers and prosopographical and bibliographic databases greatly facilitates their efficient creation while also enhancing their practical utility by providing connections to existing sources of structured data about the referenced entities. Conceptually, in many cases such domain-specific knowledge bases ultimately rely in part on many of the same primary source materials as the ultimate historical source for substantial portions of their contents. Both tasks also typically rely in practice upon the availability of accurate transcriptions of primary source materials, which – like annotations and structured knowledge – are costly to produce in the first instance.

This paper introduces a joint crowdsourced approach to the three tasks of transcription, annotation, and knowledge base construction. Crowdsourced transcription of texts provides a basis for the creation of semantically annotated editions referencing entities in an associated knowledge base, and the transcribed texts are simultaneously used to dynamically augment the knowledge base through user activity. As additional information is added to the knowledge base, this data is immediately leveraged to provide additional annotation assistance, including candidate ranking of likely referents for terms requiring disambiguation. Finally, the knowledge base and annotated texts are both exposed through combined user interfaces and editing interfaces, as well as via API for text and data mining, and integration with other projects.

The knowledge base and annotation interface are intentionally designed to incorporate very few domain-specific assumptions and data-specific processing rules, the one exception being for historical date references, typically made using the Chinese lunisolar calendar by reference (often implicitly relying in part on context) to particular rulers and/or era names. These are encoded as unambiguous machine-readable annotations allowing – where sufficiently precise primary source records exist – automated date translation to Gregorian and Julian calendars at the granularity of a single day, spanning a period of over 2,000 years of Chinese history. This in turn facilitates more precise datafication of historical facts about annotated entities than has been widely applied previously to these materials at scale. While an initial focus of annotation and knowledge base construction is on the 25 official dynastic histories (with composition dates ranging from the1st century BC through to the early 20th century), the digital library platform within which the project is based contains over 30,000 premodern Chinese works, thus enabling access to a 'long tail' of less mainstream materials as potential sources for machine-readable historical data.

From Chronicling America to Newspaper Navigator: Improving Access to Historic Newspaper Photos at the Library of Congress through Machine Learning

Ben Lee | *University of Washington*
Nathan Yarasavage | *Library of Congress*

The millions of digitized historic newspaper pages within Chronicling America, a joint initiative between the Library of Congress and the National Endowment for the Humanities, represent an incredibly rich resource for historians, journalists, genealogists, students, and members of the American public. Users regularly interact with the free and open-access collection via keyword search of the text -- but how do we navigate the abundant visual content?

In this talk, we will present Newspaper Navigator, a project created by Benjamin Lee through the Innovator in Residence Program at the Library of Congress. In collaboration with LC Labs, the National Digital Newspaper Program, and IT Design & Development at the Library of Congress, as well as Professor Daniel Weld at the University of Washington, Lee developed Newspaper Navigator to extract visual content (e.g. photographs, illustrations, maps, and comics) from 16+ million pages in Chronicling America (resulting in the Newspaper Navigator dataset) and to re-imagine how we search over such visual content using machine learning techniques through the Newspaper Navigator search application. We will also include a brief overview of the features of Chronicling America designed to support projects like Newspaper Navigator as well as research and scholarship in a variety of disciplines.

DAHN: An Accessible and Transparent Pipeline for Publishing Historical Egodocuments

Alix Chagué | Floriane Chiffoleau | *The French Institute for Research in Computer Science and Automation* (*Inria*)

The automation of the processing of documents oriented towards online publication and exploration by the humanities increases the rapidity of treatments like transcription, but they should also be an opportunity to make the experimentation and the resulting corpora sustainable and reusable. The DAHN Project (*Dispositif de soutien à l'Archivistique et aux Humanités Numériques)* relies on a joint interdisciplinary collaboration between Inria, the EHESS (School for Advanced Studies in the Social Sciences) and the University of Le Mans. By taking the example of egodocuments, the project aims to create a ready-to-use digital and scientific publishing pipeline going from the material archive to an online publication.

The workflow involves key steps such as detecting the text on an image resulting from a digitization, transcribing it, post-processing the transcription and encoding the result. Each of these steps is taken on by using open-source software, such as e-Scriptorium for the recognition of handwritten texts, and by implementing widely adopted standards for the description of texts, like TEI XML, and its associated methods and tools such as TEI Publisher. While the pipeline offers a full transformation from digitized sources to the publication of an annotated corpus, its modularity ensures its adaptability to projects with different objectives or to additional software. Although the workflow was elaborated on the specific case of egodocuments - typewritten correspondence in particular - it mixes generic tasks, which are available for any type of corpus, with closely constructed modelizations for which we maintain documentation. Furthermore, we explore the necessity to durably store and distribute the data resulting from each step, therefore ensuring the reproducibility of the whole process.

We would like to present our method and guidelines for the processing of non-digital-native textual documents using open-source and easily hackable tools that guarantee visibility across an accessible pipeline, thus challenging the notions of a black box or scattered tools which tend to be hard to maintain in the long run. The presentation will use examples taken from DAHN's edition of Paul d'Estournelles de Constant's war correspondence.

The Case for Magazines: Citizen Historian and Citizen Scientist Perspectives

Timlynn Babitsky | Jim Salmons | *Citizen Scientists*
James Hyman | *HYMAG*
Steven Lomazow | *The Great American Magazine Collection*

The goal of this panel is to turn a proportion of Digital Humanities and Cultural Heritage research attention from digitization of newspapers to focus on magazines. While both these serial publications share the challenges of complex document structure and deep content depiction, there are qualitative differences that warrant this shift in research attention. Newspapers tend to be geographically and governance-bounded publications that shine an editorial 'floodlight' on their territorial domain. Magazines, by contrast, tend to be 'spotlights' on non-geographic, topical domains of human knowledge and interest.

Over the last decade, intense interest and funding has poured into massive projects to digitize tens of millions of pages of historic newspapers. This effort has produced a wealth of publicly and scholarly available digital resources. This activity has also produced new and refined digitization technologies and workflows to produce and mine these newspaper resources. By tapping these matured technologies and workflow processes in support of the digitization of magazines, we will plumb the deeper and focused cultural heritage resource captured in these equally historic and valued serial publications.

To make our case, panelists Steven Lomazow and James Hyman will present the Citizen Historian perspective of passion-driven cultural heritage collectors. Their magazine collections are among the world's largest and most diverse of this historic cultural heritage resource. Timlynn Babitsky and Jim Salmons present the Citizen Scientist perspective as unaffiliated independent researchers. Steven Lomazow and James Hyman will profile their extensive magazine collections and describe the origin and methods of their passion-driven collecting efforts. They will also describe past and prospective future interest in collaborating with Digital Humanities researchers and history scholars.

Jim Salmons and Timlynn Babitsky will profile their research developing the MAGAZINEgts ground truth storage format and reference implementation for the Softalk magazine collection at the Internet Archive. This publication chronicles the early history of the microcomputer and digital age. They will highlight their Digital Humanities collaboration with the FAU Germany Pattern Recognition Lab and other EU researchers as part of their community involvement in the EU's Time Machine Project.

With special thanks to:

Conference Scientific Committee

Marion Ansel (National Library of France)
Sally Chambers (Ghent Centre for Digital Humanities and KBR - the Royal Library of Belgium)
Antoine Doucet (University of La Rochelle)
Tonica Hunter (Austrian National Library)
Max Kaiser (Austrian National Library)
Minna Kaukonen (National Library of Finland)
Roger Labahn (University of Rostock)
Amanda Maunoury (National Library of France)
Jean-Philippe Moreux (National Library of France)
Eva Pfanzelter (University of Innsbruck)
Cyrille Suire (University of La Rochelle)
Hannu Toivonen (University of Helsinki)

Conference Session Moderators

Eva Pfanzelter (University of Innsbruck)
Sally Chambers (Ghent Centre for Digital Humanities and KBR - the Royal Library of Belgium)
Sarah Oberbichler (University of Innsbruck)
Jean-Philippe Moreux (National Library of France)
Juha Rautiainen (National Library of Finland)

Keynote Speakers

Ann Dooms (Vrije Universiteit Brussel)
Ian Milligan (University of Waterloo)
Clemens Neudecker (Berlin State Library)
Gerben Zaagsma (University of Luxembourg)

Technical Assistance

Benjamin Eichhorn, Sophie Hammer and Christoph Steindl (Austrian National Library)

The NewsEye project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 770299.