The many uses of digitized newspapers

2nd Baltic Summer School of Digital Humanities

Essentials of Coding and Encoding

Clemens Neudecker

National Library of Latvia

collections

CLARIN
about:me

- Studied Philosophy, Computer Science and Political Science at LMU University of Munich
- 2003 - 2009: Researcher at the Bavarian State Library
- 2009 - 2014: Research Coordinator at the National Library of the Netherlands
- 2014 - now: Research Manager at the Berlin State Library

- Main areas of interest: Optical Character/Layout Recognition, Natural Language Processing, Machine/Deep Learning, Digital Humanities

- Find me at https://cneud.net or on Twitter @cneudecker
Introduction to newspapers

What is a newspaper? → Too diverse, difficult to define
Appears in a serial fashion, with regular frequency
Shorter number of pages but larger page size
National/regional/local scope, specific communities (e.g. expats, minorities)
Why Newspapers are a great source for DH

Multimodal content - text, images, statistics

Broad wealth of topics: news, novels, humour, weather, births & deaths, etc.

Captures details of the daily life in the past - events and (details of) discussions that did not make it to the history textbooks

https://minorecs.hypotheses.org/495
Why Newspapers are a terrible source for DH

OCR quality

Article segmentation and reading order challenges

Lack of coverage/digitization bias
Representation and Absence

Quantities of newspapers – a) in project b) digitised in total c) in physical libraries

Source: European Newspaper Survey Report
Newspaper digitization
Figure 1. (a) original image (b) pixel level logical labeling result (c) labelling smoothing (d) grid of separators (e) result of the segmentation into articles (f) example of reading order
- On Tuesday last, Mr. Charles. IWilson, this 8 ; had vf these, a week ago, which terminateTih his death . ; ' ' - On Sunday, dJst nt. at. AsbtCnvHall, mar Lancaster, Mr.,Geo. Worn ick, many years house'steward hit late Once The Hamilton and Brandon. He locked himself h-o-fnTw-twtes; soon, twelve o'clock that day, and fi-l-a a loaded pistol 'through Ins head, which instantaneously killed him. Coronet's Verdict, shot himself in a temporary fit of Friday week.

Excerpt from The British Newspaper Archive, Chester Courant, Tuesday 6-Apr-1819, page 3.
Named Entity Recognition

In 1917, Einstein applied the general theory of relativity to model the large-scale structure of the universe. He was visiting the United States when Adolf Hitler came to power in 1933 and did not go back to Germany, where he had been a professor at the Berlin Academy of Sciences. He settled in the U.S., becoming an American citizen in 1940. On the eve of World War II, he endorsed a letter to President Franklin D. Roosevelt, alerting him to the potential development of "extremely powerful bombs of a new type" and recommending that the U.S. begin similar research. This eventually led to what would become the Manhattan Project. Einstein supported defending the Allied forces, but largely denounced using the new discovery of nuclear fission as a weapon. Later, with the British philosopher Bertrand Russell, Einstein signed the Russell–Einstein Manifesto, which highlighted the danger of nuclear weapons. Einstein was affiliated with the Institute for Advanced Study in Princeton, New Jersey, until his death in 1955.

Tag colours:
LOCATION TIME PERSON ORGANIZATION MONEY PERCENT DATE
Europeana Newspapers (history)
Final Report

Europeana Newspapers

Special Issue - Final Report

The Europeana Newspapers project (www.europena-newspapers.eu) is an EU-funded Best Practice Network to make European digital historical newspapers available via Europeana (www.europena.eu) & The European Library (www.theeuropeanlibrary.org).

20 August 2015

Free and Open Source Software Tools for Newspapers

The Europeana Newspapers project has developed a number of free and open source software tools. These tools can help libraries and other institutions to refine their digitised historical newspapers and to evaluate the quality of the refinement.

20+ million newspaper pages available online

The 18 partners in the Europeana Newspapers project aimed to make Europe's digital historical newspapers more accessible. The overall amount of digitised newspapers that can now be accessed online has been increased dramatically.

In terms of volume and functionality, the Europeana Newspapers portal is on par with other major international newspapers digitisation programmes such as...
Newspapers in Europe and the Digital Agenda for Europe

Our Wishlist

What to address first?

Collaboration & building consensus

Users' engagement

Improve OCR + textual corpus

Seamless open access

Copyright

We need

Copyright

Tech skills

Research

Newspapers are the second hand of history. This hand, however, is often damaged, sometimes improperly.

We need to handle all of Europe's newspapers.

What is the value of past printing?

Capital mining

Digitalisation is democratisation

ICT PSP

Europeana

Open end...
DDB Zeitungsportal
Interviews with Researchers

Q&A with newspaper researchers: Sophie Kurkdjian
Posted on 30 March, 2015 by euronews — 2 Comments

Within the Europeana Newspapers project, we often speak of the value of historic newspapers for the academic community but how
Continue reading →

Q&A with newspaper researchers: Timo Honkela
Posted on 30 October, 2014 by euronews — No Comments

Within the Europeana Newspapers Project, we often speak of the value of historic newspapers for the academic community but how
Continue reading →

Q&A with newspapers researchers: Toine Pieters
Posted on 31 March, 2014 by euronews — 1 Comment

Within the Europeana Newspapers project, we often speak of the value of historic newspapers for the academic community but how
Continue reading →

Q&A with newspaper researchers: Kārlis Vērdiņš
Posted on 20 April, 2014 by euronews — No Comments

Within the Europeana Newspapers project, we often speak of the value of historic newspapers for the academic community but how
Continue reading →
General Orders.

REPEATED Complaints having been made of the great losses sustained by the Settlers at Hawkesbury, from the vexatious conduct of the Boatsmen by whom they send their Grain to Sydney, the following Regulations are to be observed.

Every person sending grain from the Hawkesbury to Sydney in an open boat, or a boat that is not trust-worthy, the Magistrates are directed to take no notice thereof.

Upon proof it appears that the Master of a Boat receives more grain than the vessel ought to take with safety, the Master shall make good any quantity he may throw overboard, or otherwise damage, lose the weight of that part, and, on conviction before two Magistrates, forfeit 5l. to the Orphan Fund.

If it shall appear to the Magistrates that grain coming round to Sydney has been read to the Guard Hulk.

By Command of His Excellency W. N. Chapman

Government House, Feb. 21, 1803.

The receiving Granaries at Parramatta and Hawkesbury, being filled with Grain which is spoiling, no more can be taken at those places until further Orders, in payment for Government Debts, an Indemnity for Whalers Investments lodged in the Store.

Wheat will continue to be received at the Stores at Sydney, until further Orders.

Wheat will be issued to the Civil, Military, &c., until further Orders; except detachments and labouring people at Curl Curl, Seven Hills, and other Out Posts, will receive Flour, as they have not the convenience of Mills.
Digitised Newspapers and the Varieties of Value
Data visualization
Data Mining Digitized Newspapers
Retronews

– Les unes à la une, plus de 400 titres pour découvrir l'histoire de la presse

Parcourez plus de 400 titres de presse d'information générale, régionale, coloniale et spécialisée publiés entre 1631 et 1950. Chaque mois Retronews s'enrichit de nouveaux journaux pour une couverture approfondie de toutes les époques.

Voir tous les journaux
Viral Texts
Oceanic Exchanges (OcEx) is funded through the Transatlantic Partnership for Social Sciences and Humanities 2016 Digging Into Data Challenge.

The dramatic expansion of newspapers over the 19th century created a global culture of abundant, rapidly circulating information. The significance of the newspaper has largely been defined in metropolitan and national terms in scholarship, while digitization by local institutions further situates newspapers in national contexts. OcEx brings together leading efforts in computational periodicals research from six countries—Finland, Germany, Mexico, the Netherlands, the United Kingdom, and the United States—to examine patterns of information flow across national and linguistic boundaries.
Media Monitoring of the Past

Mining 200 years of historical newspapers

How can newspapers help understand the past? How to explore them?

The objective of the project "Media monitoring of the past, Mining 200 years of historical newspapers" is to enable critical text mining of newspaper archives with the implementation of a technological framework to extract, process, link, and explore data from print media archives. Supported by an INTERDISCIPLINARY CONSORTIUM composed of computational linguists, digital humanists, designers, historians, librarians and archivists, impresso (i.e. “what has been printed”) will tackle the challenges of content enrichment and data representation, visualization and analysis, completed by methodological and epistemological reflections. Expected outcomes include, among others, a set of natural language processing (NLP) tools dedicated to historical print media, visualization interfaces for active exploration and critical analysis of newspaper corpora in a transparent manner, as well as a digital history research project on resistance to European unification in the late 19th and early 20th centuries.

Time frame: September 2017 - August 2020

Funding: Swiss National Science Foundation
NewsEYE

Upcoming events

26 Aug 2019 IFLA WLIC CONFERENCE, NEWS+EYE PRESENTATION BY NATIONAL LIBRARY OF FINLAND

NewsEYE collects information about cultural, political and social events in a more detailed way than any other public record. Since their beginnings in the 17th century they are recording billions of events, stories and names, in almost every language, every country and every day. Newspapers were always an important medium for the dissemination of public and political opinions, literary works, essays and art. This thematic wealth sets them at the center stage for anyone interested in European cultural heritage.
More than a Feeling

Media Sentiment as a Mirror of Investors’ Expectations at the Berlin Stock Exchange, 1872–1930

Investors’ behaviour is not only influenced by information on fundamentals, such as future cash flows, but also by soft parameters, such as moods, emotions, or feelings. Accordingly, stock returns and trading volume are only partially determined by fundamentals. Recently, the notion of “investor sentiment” has attracted the attention of scholars in finance and financial economics. Although the understanding of the nature of sentiment is still far from being complete, there is considerable empirical evidence that sentiment plays a crucial role for financial markets. In this project, we follow the idea that sentiment is a belief about future developments that is not built upon facts and rational reasoning. Accordingly, sentiment can be regarded as being equivalent to investors’ (non-)fundamental expectations. In this project, we aim to investigate the influence of sentiment for the case of the Berlin stock exchange between 1872 and 1930. We will use different computational sentiment analysis techniques (dictionaries, machine learning) to automatically extract sentiment from a daily column in the Berliner Abend-Zeitung, which was the most important financial newspaper at the time. We will create a daily sentiment index which can be used to measure investors’ expectations as defined above. In combination with content analysis methods, such as topic modelling and stock market data, we will investigate the impact of expectations and the accompanying narratives on the Berlin stock market, both quantitatively and qualitatively. As we will apply different text mining methods on a large scale, we build this project on an interdisciplinary cooperation between economic history and digital humanities. Addressing the following questions: How did historical experiences, such as wars and political or economic shocks, change investor sentiment i.e. investors’ expectations, and how did sentiment influence financial markets? Has this influence changed over time? Which role did sentiment play in the development of financial and economic crises, and which were the narratives that accompanied high stock market fluctuations? Hence, our project aims to contribute directly to the priority programme’s three central questions.

Jun.-Prof. Dr. Manuel Burghardt, Leipzig (Principal Investigator)
Prof. Dr. Mark Spenner, Regensburg (Principal Investigator)
Lino Walchshausen, M.Sc., Regensburg
Siamese

Similar Advertisement Search

SIAMESE allows users to search for similar images in three ways:

- Paste a Delpher URL for an advertisement from Algemeen Handelsblad (1945-1969) or NRC Handelsblad (1970-1995) into the search box
- Click on one of the images to use it as a source image
- Click on 'random' for a random source image

Click on year to navigate to Delpher.

Source image

Top ten nearest neighbors
CHRONReader: Query Classified Historical Newspaper Images

Select date-range:
1860-01-01 to 1920-01-31

Select image type:
- Photo
- Drawing
- Photo or drawing

Select image category:
Building

Image with face:
- No
- Yes
- Yes or no

Text in article:
Train

http://www.kbresearch.nl/xportal?image_type="photo" and image_type="building" and "Train" and date within "1860-01-01T00:00:00Z" to "1920-01-31T00:00:00Z"
Newspapers and capitalism

The Newspaper group studies how newspapers transformed from the late eighteenth century to the early twentieth century as a channel for marketing goods and services. It further studies the relationship between advertising and journalistic ideals. The datasets used consist of digitized newspapers from the British Library's Nineteenth Century Newspapers. Studying the relation of advertisements and journalistic ideals in newspapers has two potentially important outcomes. First, it may contribute to a scholarly discussion on the development of print capitalism by using the advertisements as a proxy for larger political and social developments. Second, by better understanding the logic of advertisements in the press, it may help in developing article extraction (at least with regard to advertisements) and thus pave the way for future study that needs to be able to discern between text that emerges from advertisements and from journalistic text.
DH2019

- Oceanic Exchanges: Transnational Textual Migration And Viral Culture
- The Past, Present and Future of Digital Scholarship with Newspaper Collections
- Complexities in the Use, Analysis, and Representation of Historical Digital Periodicals
Coding da Vinci: Berliner Schlagzeilen

Die Berliner Volks-Zeitung war vor 100 Jahren die Tageszeitung für die Metropole Berlin. Mit Anspruch, Sensationen und Unterhaltung wurde sie zur erfolgreichsten Boulevardzeitung der Zeit. Der Twitter-Bot „Berliner Schlagzeilen“ tweetet die Schlagzeilen und ein Bild der Titelseite von vor 100 Jahren jeden Tag, auf den Tag.

Wir verbinden die Vergangenheit mit der Gegenwart und machen Geschichte einem breiten Publikum zugänglich. Schon jetzt haben wir mit diesem Konzept mehrere 10.000 Menschen (und Bots 😊) erreicht. Doch dies ist erst der Anfang.

Unser Projekt startete am Abend des 2. Tages des Hackatons mit den ersten Tweets und tweetet seitdem konstant jeden Tag, meist sogar 2 mal, neue Schlagzeilen. Über die Zeit haben wir das tweeten immer weiter automatisiert und werden dies soweit vorantreiben, dass wir voraussichtlich in den nächsten Monaten alle Tweets vorbereitet haben, die wir bis zum Auslaufen der Zeitung im Jahr 1930, also 2030 in unserer Zeit, tweeten können.

Unser Twitterfeed bekommt jeden Tag durchschnittlich einen neuen Follower und wird somit an unserem Todestag 4745 Follower haben. Mit 21.5K Ansichten unserer Tweets allein im November Können wir schon jetzt erste Highlights feiern.
Conclusion and Outlook

Large quantities of digitized newspapers are ready available, new digitization projects and portals are following.

OCR & OLR have had recent breakthroughs thanks to Machine Learning, so we can expect better full text and article segmentation to become standard soon.

Many diverse research activities and research communities around newspapers currently ongoing with future perspectives.

→ Now is the time for historical newspapers!
Thank you for your attention! Questions please?

2nd Baltic Summer School of Digital Humanities

Essentials of Coding and Encoding

Clemens Neudecker
<table>
<thead>
<tr>
<th>Reference</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td><a href="https://twitter.com/jpmoreux/status/1149052250905600000">https://twitter.com/jpmoreux/status/1149052250905600000</a></td>
</tr>
<tr>
<td>8</td>
<td><a href="https://arxiv.org/abs/1210.0999">https://arxiv.org/abs/1210.0999</a></td>
</tr>
<tr>
<td>9</td>
<td><a href="http://www.dlib.org/dlib/july09/munoz/07munoz.html">http://www.dlib.org/dlib/july09/munoz/07munoz.html</a></td>
</tr>
<tr>
<td>10</td>
<td><a href="http://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/">http://www.europeana-newspapers.eu/named-entity-recognition-for-digitised-newspapers/</a></td>
</tr>
<tr>
<td>11</td>
<td><a href="http://www.theeuropeanlibrary.org/tel4/newspapers">http://www.theeuropeanlibrary.org/tel4/newspapers</a></td>
</tr>
<tr>
<td>12</td>
<td><a href="http://europeananewspapers.github.io/">http://europeananewspapers.github.io/</a></td>
</tr>
<tr>
<td>14</td>
<td><a href="http://newspapers.europeana.eu">http://newspapers.europeana.eu</a></td>
</tr>
</tbody>
</table>
## References

<table>
<thead>
<tr>
<th></th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>15</td>
<td><a href="https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/DDB-Zeitungsportal/DDB-Zeitungsportal_node.html">https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/DDB-Zeitungsportal/DDB-Zeitungsportal_node.html</a></td>
</tr>
<tr>
<td>16</td>
<td><a href="https://zdb-katalog.de">https://zdb-katalog.de</a></td>
</tr>
</tbody>
</table>
## References

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td><a href="https://www.retronews.fr">https://www.retronews.fr</a></td>
</tr>
<tr>
<td>23</td>
<td><a href="https://viraltexts.org/">https://viraltexts.org/</a></td>
</tr>
<tr>
<td>24</td>
<td><a href="https://oceanicexchanges.org/">https://oceanicexchanges.org/</a></td>
</tr>
<tr>
<td>25</td>
<td><a href="https://impresso-project.ch/">https://impresso-project.ch/</a></td>
</tr>
<tr>
<td>26</td>
<td><a href="https://www.newseye.eu/">https://www.newseye.eu/</a></td>
</tr>
<tr>
<td>28</td>
<td><a href="https://lab.kb.nl/tool/siamese">https://lab.kb.nl/tool/siamese</a></td>
</tr>
<tr>
<td>Reference Number</td>
<td>Reference URL</td>
</tr>
<tr>
<td>------------------</td>
<td>---------------------------------------------------------</td>
</tr>
<tr>
<td>29</td>
<td><a href="https://lab.kb.nl/tool/chronreader">https://lab.kb.nl/tool/chronreader</a></td>
</tr>
<tr>
<td>30</td>
<td><a href="https://twitter.com/NewsDHH19">https://twitter.com/NewsDHH19</a></td>
</tr>
<tr>
<td>31</td>
<td></td>
</tr>
<tr>
<td>32</td>
<td><a href="https://codingdavinci.de/projects/2017/altpapier.html">https://codingdavinci.de/projects/2017/altpapier.html</a></td>
</tr>
<tr>
<td>33</td>
<td><a href="https://codingdavinci.de/projects/2017/ber_schlagzeilen.html">https://codingdavinci.de/projects/2017/ber_schlagzeilen.html</a></td>
</tr>
</tbody>
</table>