Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 969–976
Marseille, 11–16 May 2020

Allgemeine Musikalische Zeitung as a Searchable Online Corpus

Bernd Kampe[1] **Tinghui Duan**[1,2] **Udo Hahn**[1]
[1] Jena University Language & Information Engineering Lab (JULIE) Lab
[2] Graduate School "Romanticism as a Model"
Friedrich-Schiller-Universität Jena, Jena, Germany
{bernd.kampe | tinghui.duan | udo.hahn}@uni-jena.de

Abstract

The massive digitization efforts related to historical newspapers over the past decades have focused on mass media sources and ordinary people as their primary recipients. Much less attention has been paid to newspapers published for a more specialized audience, e.g., those aiming at scholarly or cultural exchange within intellectual communities much narrower in scope, such as newspapers devoted to music criticism, arts or philosophy. Only some few of these specialized newspapers have been digitized up until now, but they are usually not well curated in terms of digitization quality, data formatting, completeness, redundancy (de-duplication), supply of metadata, and, hence, searchability. This paper describes our approach to eliminate these drawbacks for a major German-language newspaper resource of the Romantic Age, the *Allgemeine Musikalische Zeitung* (*General Music Gazette*). We here focus on a workflow that copes with *a posteriori* digitization problems, inconsistent OCRing and index building for searchability. In addition, we provide a user-friendly graphic interface to empower content-centric access to this (and other) digital resource(s) adopting open-source software for the purpose of Web presentation.

Keywords: Digital Humanities, historical newspapers, corpus compilation, romanticism research, data integration

1. Introduction

The on-going large-scale digitization of historical newspapers lays the infrastructure foundations for profoundly changing information management workflows in the humanities (Trace and Karadkar, 2017; Blanke and Hedges, 2013) and, in the end, allows humanities scholars to explore entirely new research avenues. For example, the *Oceanic Exchanges* project[1] analyzes historical newspapers from 1840 through 1914 to examine patterns of information flow across national and language borders (Oiva et al., 2019). The *Living with Machines* project[2] focuses on newspapers published in the United Kingdom during the Long Nineteenth Century (c.1780–1918) and examines the ways in which technology altered people's work and life. In several case studies of the *NewsEye* project,[3] topics such as migration, gender and nationalism are investigated based on digitized historical newspapers from 1850 through 1950 provided by the national libraries of Austria, Finland and France. A case study of the *impresso* project[4] on the anti-European movements explores how to make newspapers a valuable source for examining public opinion in the 19th and 20th century.

Whilst all of the above-mentioned projects concentrate on general newspapers aimed at ordinary people as primary recipients, newspapers targeting narrower defined intellectual circles still suffer from a pronounced lack of attention in digitization campaigns. Hence, researchers in such fields miss the advantages resulting from digital access, search and automated analysis tools (cf., e.g., the frameworks provided by Niekler et al. (2018), Frank and Ivanovic (2018), Pustejovsky et al. (2017), Brooke et al. (2015), or Hinrichs and Krauwer (2014) and Hinrichs et al. (2018)). Moreover, they face the well-known obstacles of intellectual paper work — typically, lacking completeness of the locally available newspaper collection, or, if complete volumes are available, the dilemma to be unable to sift through entire volumes by year or even decade in sufficient depth, given a specific content focus. Still, some of these newspapers have been (often only partially) digitized, possibly even with options for searchable OCRed text. But as long as they are not curated carefully (e.g., with checks for digitization quality, correctness of OCRing, completeness of coverage, duplicated portions, etc.), they cannot reasonably be handed over to computational tooling. Meanwhile, humanities scholars who just want to search for certain keywords in these newspapers will quickly get lost in too many hits, because the filtering functionality provided by most digital libraries is often very limited, sometimes even erroneous.

Indeed, all of the problems addressed above apply to the *Allgemeine Musikalische Zeitung* (*General Music Gazette*, *AMZ*) which was published between 1798 and 1865 in Leipzig, Germany. This gazette, alongside of the *Allgemeine Literatur Zeitung, ALZ* (Hahn and Duan, 2019), is regarded as one of the most important text sources for research on the German Romanticism. Not only musicologists (Ringer, 1990; Milsom, 2011; Neubauer, 2017), but also literary scholars (Donovan and Elliott, 2004), among others, attribute a supreme potential to this resource. Some important literary works of German romanticism, such as E.T.A. Hoffmann's *Ritter Gluck* (Kremer, 2010), were even published in the *AMZ* for the first time.

In order to make advanced computational content analytics accessible for the *AMZ*, on the one hand, and to provide a user-friendly search interface for humanities scholars who typically lack in-depth technical skills, on the other hand, we created a complete digital full-text corpus of the *AMZ*.

We assembled available portions of that resource from different digital libraries scattered across various physical

[1] https://oceanicexchanges.org
[2] https://livingwithmachines.ac.uk
[3] https://newseye.eu
[4] https://impresso-project.ch

sites, generated a homogeneous format for the textual data and made them browsable and searchable by adopting modern Web technologies. In this paper we describe the workflow underlying these efforts and present a unique and comprehensive platform of intellectual newspapers of German Romanticism for the Digital Humanities community, especially for researchers working on German Romanticism.

2. Related work

There are two streams of activities related to our work — researchers initiating thematically narrow digitization projects, typically as a side issue of their main area of work, on the one hand, and professional digital archivists running large-scale digitization campaigns, often serving a national archiving agenda, on the other hand. The first one is due to humanities scholars who commmonly make use of some specific historical newspapers because they find important evidence in them for their own research work. In order to make these valuable sources accessible to other interested researchers independent of time and place, they initiated digitization projects to transform paper-based newspaper collections into a digital, more easily sharable form.

One of these projects is the *Literarische Zeitschriften um 1800* (Literary Periodicals around 1800),[5] funded by the German Research Foundation (DFG) and carried out between 2007 and 2017. In this context, four well-known periodicals were digitized: the *Allgemeine Literatur-Zeitung* (*General Literature Gazette*, 1785–1849), *Jenaische Allgemeine Literatur-Zeitung* (*Jena General Literature Gazette*, 1804–1841), *Journal des Luxus und der Moden* (*Journal of Luxury and Fashions*, 1787–1812) and *Leipziger Literaturzeitung* (*Leipzig Literature Gazette*, 1802–1834). They were thoroughly scanned and enriched with metadata at the article level, such as date, author and category. However, for three of them (with the exception of the *Leipziger Literaturzeitung*), the scanned pages have not been OCRed, so automated search is precluded. The *Leipziger Literaturzeitung*[6] comes with OCRed full text, but neither full texts nor scanned pages can be bulk-downloaded.

Similar digitization projects like *"Der Blick auf den Krieg"* (The View on the War)[7] or *"Historische Presse der deutschen Sozialdemokratie online"* (Historical Press of the German Social Democracy Online)[8] were initiated by special interest groups who were interested in thematically even more constrained newspapers. For these efforts, they apply up-to-date digitization techniques, make their digitized newspapers searchable and present them in a user-friendly way. However, these projects are usually not based on open source software and do not comply with a standardized workflow. Consequently, their frameworks are neither reusable, nor are their data interoperable.

In summary, such digitization projects *selectively* transform historical newspapers in a digital format, yet advanced computational analytics, such as topic modeling (Glenny et al., 2019), document clustering (Hoenen, 2018), social network analysis (Jayannavar et al., 2015; Agarwal et al., 2013; Elson et al., 2010), semantic technologies (Meroño Peñuela et al., 2015; Wang et al., 2012), information extraction or text mining (Higuchi et al., 2018; Widlöcher et al., 2015), and visualization of the analytical results they yield (Scrivner and Davis, 2017; Bradley et al., 2016; El-Assady et al., 2016) are completely out of reach. Such sophisticated techniques require cleaned textual data, common data formats and search indexes to properly address content items.

The second stream of related work originates from data curators who see their main task in digitizing printed newspapers at a much larger scale. The American *National Digital Newspaper Program* (NDNP),[9] launched in 2004, is a partnership between the National Endowment for the Humanities (NEH) and the Library of Congress (LC). It aims at the creation of a national digital resource of historically significant newspapers published between 1690 and 1963, from all the U.S. states and territories. Since 2005, 15,855,607 pages from 155,857 newspaper titles were digitized, including 278,486 pages in German language.[10] A similar initiative, called TROVE, is due to the Australian Newspaper Digitisation Program (ANDP)[11] where 23,407,352 newspaper pages and 2,026,782 gazette pages from 1806 to around 2007 have been digitized, amounting overall to 143 million articles (Cassidy, 2016). The British Newspaper Archive, launched in 2011 as a partnership between the British Library and Findmypast (a private company), has digitized more than 35 million pages of newspapers from 1700s until today. Upon payment, full text searching is available. However, full texts cannot be downloaded and scanned pages can only be downloaded page by page as PDF.

The National Library of Finland has digitized a large proportion of the historical newspapers published in Finland between from 1771 to 1929, yet suffers from lexical quality issues of OCRed documents (Kettunen and Pääkkönen, 2016; Kettunen et al., 2020). This collection currently contains approximately 18,6 million pages in Finnish and Swedish. The National Library's Digital Collections are offered by a Web service, also known as DIGI.[12] Part of this material is also freely downloadable from *The Language Bank of Finland* provided by the FIN-CLARIN consortium. Further national digital newspaper programs include DDB NEWSPAPER PORTAL of the German Digital Library,[13] ANNO (AustriaN Newspapers Online)[14] and many more. What these projects have in common, is that they *massively* digitize primarily historical newspapers, *as complete as possible*. However, their clear focus is on printed mass media targeted at a general audience.

[5] http://projekte.thulb.uni-jena.de/literaturportal/

[6] https://zs.thulb.uni-jena.de/receive/jportal_jpvolume_00220472

[7] https://hwk1.hebis.de

[8] https://www.fes.de/bibliothek/vorwaerts-blog/

[9] https://www.loc.gov/ndnp/

[10] https://chroniclingamerica.loc.gov

[11] https://trove.nla.gov.au/newspaper/about

[12] https://digi.kansalliskirjasto.fi/etusivu

[13] https://www.dnb.de/EN/Professionell/ProjekteKooperationen/Projekte/DDB-Zeitungsportal/DDB-Zeitungsportal_node.html

[14] http://anno.onb.ac.at

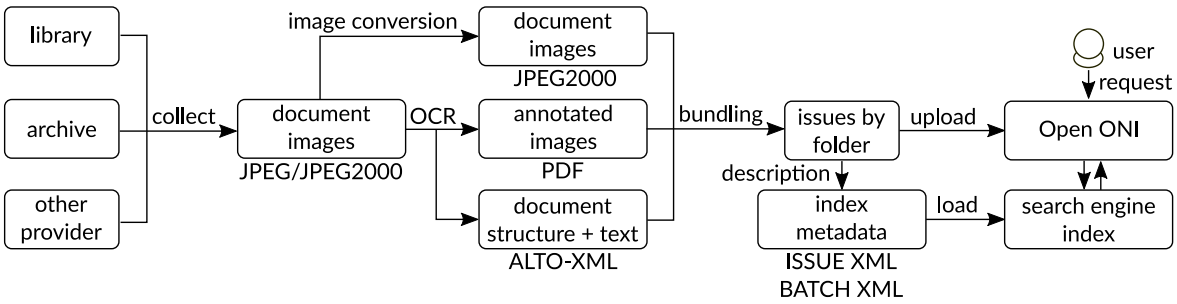

Figure 1: Schema of the workflow to set up OPEN ONI with the *AMZ* corpus.

In order to make more advanced computational content analytics accessible for specialized newspapers as well, we developed a general workflow for assembling scattered digitized fragments of historical newspapers from different digital libraries into a single full-text corpus (Hahn and Duan, 2019). On the example of the *Allgemeine Literatur-Zeitung* (General Literature Gazette, *ALZ*), a major text source for research on the German Romanticism, a full-text corpus of 126,612 pages containing 120,369,005 tokens was created and released in XML-format. This corpus already covers about 82% of the entire volumes of the *ALZ*. Although this approach provided a text base for computational analytics, yet for humanities scholars who lack in-depth technical skills, a searchable and browsable interface would be a much desired add-on. Hence, in this paper, we go one step further. We not only compile a fully digitized, interoperable text corpus, but also provide a user-friendly search interface by adopting open source Web applications.

3. Technicalities of Corpus Compilation

The technical foundation of our project is given by the Open Online Newspaper Initiative (OPEN ONI), which is both the name of a piece of software, as well as an institutionalized collaboration with the goal to make it easier to set up software to display digital newspapers. OPEN ONI[15] is a fork of the Library of Congress' Chronicling America (Yarasavage et al., 2012), yet follows the standards established by its institutional predecessor.

The starting point of the workflow depicted in Figure 1 are already digitized versions of the *AMZ*, as available in digital libraries, archives or other sources worldwide.

At that point, only document images are available in different formats. To comply with the Library of Congress' image format specifications (Buckley and Sam, 2006), one of the major steps is the harmonization of all heterogeneous file formats. We converted all document images that were not already in the target format to JPEG2000 using the freely available IMAGEMAGICK.[16] A second major step addresses the analysis of the page structure and the annotation of the pictures with the results of the OCR step. This makes each page accessible as an image itself, a PDF file containing the page and the corresponding text layer, as well as the text, both as plain text and ALTO-XML. All of these different formats are made available for download through OPEN ONI. Layout information and the results of OCR are to be kept in ALTO-XML files, a standard which was created during the METAe project (2000–2003) and is nowadays hosted by the Library of Congress.[17] This information enables highlighting of search results directly in the document images (as illustrated in Figure 2, mentions of "Schumann" are marked across the 850 results). The files of all generated resources listed above are bundled together,

[15] https://open-oni.github.io

[16] https://imagemagick.org

[17] https://www.loc.gov/standards/alto/

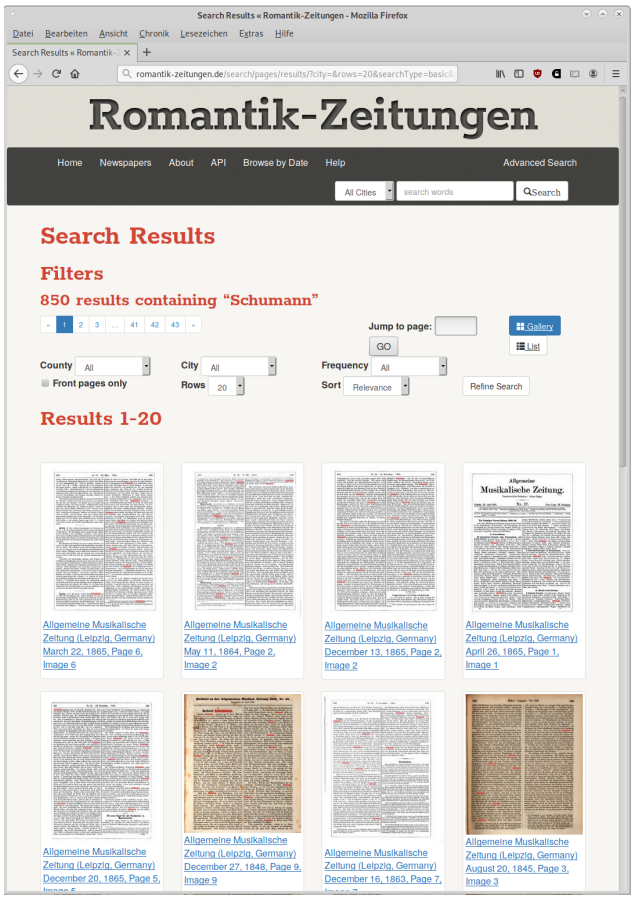

Figure 2: Screenshot of a search result for the query "Schumann", including page-wise highlighted mentions (in red) of the composer and conductor Robert Schumann and his wife Clara, also a composer and pianist. Both are highly appraised representatives of the Romantic era.

	Bavarian State Library			Internet Archive			*AMZ* Full-Text Corpus		
Library	**volumes**	**pages**	**tokens**	**volumes**	**pages**	**tokens**	**volumes**	**pages**	**tokens**
Bavarian State Library	50	24,882	21,051,116	1	476	458,881	51	25,358	21,509,997
University of Oxford	-	-	-	2	1,070	672,226	2	1,070	672,226
New York Public Library	-	-	-	-	2	1,404	-	2	1,404
Total	50	24,882	21,051,116	3	1,548	1,132,511	53	26,390	22,183,627

Table 1: Quantitative breakdown of the components of the *AMZ* collection taken from the Bavarian State Library, University of Oxford, and New York Public Library constituting our newly assembled full-text *AMZ* corpus

issue by issue, in separate directories as described in Subsection 4.2.

Finally, the folder contents is indexed into ISSUE XML and BATCH XML files as described by the guidelines of the National Digital Newspaper Program, which in turn are used to generate the search index. To help with missing metadata that were not available to us, we made use of existing project data of the *Salt Lake Tribune*[18] and used it as a template, automatically substituting the content of XML fields with information from our folders. The actual query processing is powered by the APACHE SOLR search engine[19] via an interface created with DJANGO.[20]

4. Workflow for Corpus Construction

A workflow scheme similar to the one depicted in Figure 1, yet without the OPEN ONI framework, has already been shown to yield promising results for the literary-focused *Allgemeine Literatur-Zeitung* (Hahn and Duan, 2019). It consists of the following logical steps that were adapted to the *Allgemeine Musikalische Zeitung*:

1. Collecting all available digitized versions of *AMZ* from various digital libraries (the set of documents);
2. Image processing of those documents;
3. Assessing the quality of the OCR processing for the documents;
4. Assembling the complete *AMZ* corpus.

4.1. Data Collection

Most of the volumes were gathered from Bayerische Staatsbibliothek (BSB),[21] except for volumes 2 (1799/1800), 5 (1802/1803), and 39 (1837). Volumes 2 and 5 were not available from BSB, while pages in volumes 39 contained partial content from adjacent pages, which was detrimental to the quality requirements of our OCR process. Those volumes were retrieved from alternative Internet Archive sites: 2 & 5 from the University of Oxford, 39 from the New York Public Library (for details and quantitative data for *AMZ*, cf. Table 1).

While the document archives of BSB also contained the results of their native OCR for each text file, but lacked the specific coordinates of the text boxes within the pages, they only served as a benchmark for us. So we could still decide which OCR engine to use for our workflow.

4.2. Document image processing

Another necessity for the operation of the search portal is the separation of each (yearly) volume into individual (weekly) issues. OPEN ONI expects this content to be available in folders marked with timestamps (employing a 'YYYYMMDDII' format, where 'YYYY' denotes the year, 'MM' the month with leading zero, and the same holds for the placeholder for days, 'DD'). As several issues could have been published on the same day, 'II' specifies the issue number, which in the case of the *AMZ* is always '01' right now. We do not distinguish between the actual issue itself and possibly attached *Intelligenzblättern* (News Sheets) which either contain announcements and advertisements for the readers or addenda.

In order to partially automate the arrangement of the issues, we employed a script that examined the upper part of all document pages and matched text in that section with the string 'MUSIKALISCHE ZEITUNG' given a maximum Levenshtein distance of 3. Such a conditional match gives strong evidence that an issue header is really identified. Since each volume usually starts on a Wednesday (with the exception of starting at the 1st of January or the 1st of October in a few rare cases), this default day allowed us to calculate the date of the first Wednesday per volume and then incrementing this date by 7 every time a header was found.

This process still required manual curation to check whether all 52 issues (or 65 for volume 9) had been found and to look for potential gaps in the sequence of title pages. During this step, we realized that the last page of issue 34 and the first page of issue 35 in volume 25 were missing from the BSB archive and had to be obtained from an alternative source. Another strong outlier was volume 39 where only 26 out of 52 title pages were found automatically by the script. A closer look at the entire volume then led us to replace it entirely, as already mentioned in Section 4.1.

4.3. Assessment of OCR

Reul et al. (2019) have recently made available OCR4ALL,[22] a tool suite geared towards the processing of historical printings. Another larger project that deals with mass digitization of historical newspapers is OCR-D, which has a special focus on German OCR (Neudecker et al., 2019). As there is currently no mapping between PageXML and the ALTO standard, we were not able to benefit from this specialization. The same holds true for OCR-D.

[18] https://chroniclingamerica.loc.gov/lccn/sn83045396/

[19] https://lucene.apache.org/solr/

[20] https://www.djangoproject.com

[21] https://app.digitale-sammlungen.de/bookshelf/

[22] https://github.com/OCR4all/OCR4all

In order to assess the quality of OCR for *AMZ*,[23] we transcribed a small random sample of pages of issue 25. This volume was chosen since the quality of the scans is in general dependent on the age of the issue, with scans of older issues being of lower quality than more recent ones.
Tables 2 and 3 depict the Character Error Rate (CER) and Word Error Rate (WER) of different approaches (ignoring and including punctuation errors, respectively) in relation to the transcribed ground truth (for an extrinsic evaluation of OCR errors, cf. Tanner et al. (2009)). The CER is defined as the edit distance of two strings divided by their maximum length. *German* and *German_best* refer to the output of Tesseract's standard recognition model for German and the slower but (according to the developers) best trained model, respectively. For comparison, in an evaluation of OCR4ALL with book-specific models these produced a CER between less than 1% and 5.3% (Reul et al., 2019).

page	BSB	German	German_best
183	1.49 / 9.78	1.29 / 3.51	1.49 / 5.17
231	1.35 / 5.38	4.69 / 11.26	4.70 / 12.77
297	1.49 / 4.02	2.62 / 4.62	2.83 / 6.41
377	1.63 / 3.54	3.80 / 6.34	3.93 / 7.82

Table 2: Evaluation of CER/WER on a random sample of volume 25 (excluding punctuation errors).

page	BSB	German	German_best
183	1.83 / 9.78	2.28 / 3.51	3.43 / 5.17
231	1.58 / 5.38	7.17 / 11.26	8.81 / 12.77
297	1.90 / 4.02	3.60 / 4.62	4.67 / 6.41
377	2.63 / 3.54	4.62 / 6.34	5.26 / 7.82

Table 3: Evaluation of CER/WER on a random sample of volume 25 (including punctuation errors).

A large amount of those character errors can be attributed to the class of punctuation marks. A closer look at the alignment of words and characters revealed that most of them could be attributed to stains on the paper which were mistaken by the OCR engine as full stops, commas, colons or semicolons. The proportion of whitespace errors (e.g., *"performedby"*, a much more common source of digitization errors in OCRed historical newspaper corpora (Soni et al., 2019), however, is comparatively low in our corpus.
Tesseract's[24] designated best OCR model turned out to perform slightly worse than the standard model on all inputs, which shows the dependency of good results on suitable training datasets. Its most recent 4.1 version added the ALTO-XML standard to its list of supported output formats.
From a computational infrastructure and resource consumption perspective, the generation of the PDF and ALTO-XML files for all 53 issues took 307 hours of elapsed sequential wall time on an Intel(R) Xeon(R) workstation with an E5-1620 v4 @ 3.50GHz. Table 1 highlights the magnitude of this task.

[23]The impact of OCR errors for digital libraries is thoroughly discussed by Chiron et al. (2017)

[24]`https://github.com/tesseract-ocr/tesseract`

5. Conclusions

We described the compilation of a complete and fully searchable digitized corpus for a specialized historical newspaper, the *Allgemeine Musikalische Zeitung*. Together with the *Allgemeine Literatur-Zeitung*, it will be available online as a hub for German-language Romanticism research.[25] Furthermore, the code used to transform the corpus and generate all the resources we presented in this paper is available at `https://github.com/JULIELab/romantik-zeitungen`. The corpus has also been archived at Zenodo and can be downloaded from `https://zenodo.org/record/3708427`.
The current search functionality is limited to a free-text search mode, with well-known drawbacks, e.g., lack of coverage for synonyms, short forms, etc. Following previous work, e.g. by Neudecker (2016) on the Europeana newspaper corpora, one of the major next steps will be to enhance the corpus substantially by semantic metadata in terms of named entities keeping an eye on effects of corrupted OCR input (Grover et al., 2008; Alex and Burns, 2014; Kim and Cassidy, 2015; Kettunen and Ruokolainen, 2017; Bircher, 2019).

6. Acknowledgements.

This work was conducted within the Graduate School "Romanticism as a Model" (GRK 2041) and the Collaborative Research Center AquaDiva (CRC 1076). Both are supported by the German Research Foundation (*Deutsche Forschungsgemeinschaft*, DFG). We especially want to thank Christiane Wiesenfeldt for pointing out to us the relevance of *AMZ* for the entire Romantic period.

7. Bibliographical References

Agarwal, A., Kotalwar, A., and Rambow, O. C. (2013). Automatic extraction of social networks from literary text: a case study on *Alice in Wonderland*. In Hsin-Hsi Chen, et al., editors, *IJCNLP 2013 — Proceedings of the 6th International Joint Conference on Natural Language Processing. Nagoya, Japan, 14-18 October 2013*, pages 1202–1208. Asian Federation of Natural Language Processing (AFNLP).

Alex, B. and Burns, J. (2014). Estimating and rating the quality of optically character recognised text. In Apostolos Antonacopoulos et al., editors, *DATeCH 2014 — Proceedings of the 1st International Conference on Digital Access to Textual Cultural Heritage. Madrid, Spain, May 19-20, 2014*, pages 97–102. Association for Computing Machinery (ACM).

Bircher, S. (2019). Toulouse and Cahors are French cities, but T«i*louse and Caa.Qrs as well: a neural approach for detecting named entities in digitized historical newspapers. Master's thesis, Institute of Computational Linguistics, University of Zurich.

[25]`http://www.romantik-zeitungen.de`

Blanke, T. and Hedges, M. (2013). Scholarly primitives: building institutional infrastructure for humanities e-science. *Future Generation Computer Systems*, 29(2):654–661.

Bradley, A. J., Mehta, H., Hancock, M., and Collins, C. (2016). Visualization, Digital Humanities, and the problem of instrumentalism. In *#VIS4DH 2016 — Proceedings of the 1st Workshop on Visualization for the Digital Humanities @ VIS 2016. Baltimore, Maryland, USA, 24 October 2016*. Institute of Electrical and Electronics Engineers (IEEE).

Brooke, J., Hammond, A., and Hirst, G. (2015). GUTENTAG: an NLP-driven tool for Digital Humanities research in the Project Gutenberg corpus. In Anna Feldman, et al., editors, *CLfL 2015 — Proceedings of the 4th Workshop on Computational Linguistics for Literature @ NAACL-HLT 2015. Denver, Colorado, USA, June 4, 2015*, pages 42–47. Association for Computational Linguistics (ACL).

Buckley, R. and Sam, R., (2006). *JPEG 2000 Profile for the National Digital Newspaper Program*. Library of Congress, Office of Strategic Initiatives & Xerox Global Services.

Cassidy, S. (2016). Publishing the TROVE newspaper corpus. In Nicoletta Calzolari, et al., editors, *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 4520–4525, Paris. European Language Resources Association (ELRA-ELDA).

Chiron, G., Doucet, A., Coustaty, M., Visani, M., and Moreux, J.-P. (2017). Impact of OCR errors on the use of digital libraries: towards a better access to information. In Adam Jatowt, et al., editors, *JCDL '17 — Proceedings of the 17th ACM/IEEE-CS Joint Conference on Digital Libraries. Toronto, Ontario, Canada, 19-23 June 2017*, pages 249–252. Institute of Electrical and Electronics Engineers (IEEE).

Siobhán Donovan et al., editors. (2004). *Music and Literature in German Romanticism*. Camden House.

El-Assady, M., Gold, V., John, M., Ertl, T., and Keim, D. A. (2016). Visual text analytics in context of Digital Humanities. In *#VIS4DH 2016 — Proceedings of the 1st Workshop on Visualization for the Digital Humanities @ VIS 2016. Baltimore, Maryland, USA, 24 October 2016*. Institute of Electrical and Electronics Engineers (IEEE).

Elson, D. K., Dames, N., and McKeown, K. R. (2010). Extracting social networks from literary fiction. In Jan Hajič, et al., editors, *ACL 2010 — Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden, 11-16 July 2010*, pages 138–147, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Frank, A. U. and Ivanovic, C. (2018). LITTEXT: building literary corpora for computational literary analysis. A prototype to bridge the gap between CL and DH. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 798–804, Paris. European Language Resources Association (ELRA).

Glenny, V., Tuke, J., Bean, N., and Mitchell, L. (2019). A framework for streamlined statistical prediction using topic models. In Beatrice Alex, et al., editors, *LaTeCH-CLfL 2019 — Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature @ NAACL-HLT 2019. Minneapolis, Minnesota, USA, June 7, 2019*, pages 61–70, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Grover, C., Givon, S., Tobin, R., and Ball, J. (2008). Named entity recognition for digitised historical texts. In Nicoletta Calzolari, et al., editors, *LREC 2008 — Proceedings of the 6th International Conference on Language Resources and Evaluation. Marrakech, Morocco, 26 May - June 1, 2008*, pages 1343–1346, Paris. European Language Resources Association (ELRA).

Hahn, U. and Duan, T. (2019). Corpus assembly as text data integration from digital libraries and the Web. In Maria Bonn, et al., editors, *JCDL '19 — Proceedings of the 19th ACM/IEEE-CS Joint Conference on Digital Libraries. Urbana-Champaign, Illinois, USA, June 2-6, 2019*, pages 25–28. IEEE Computer Society.

Higuchi, S., Freitas, C., Cuconato, B., and Rademaker, A. (2018). Text mining for history: first steps on building a large dataset. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 3754–3760, Paris. European Language Resources Association (ELRA).

Hinrichs, E. W. and Krauwer, S. (2014). The CLARIN research infrastructure: resources and tools for eHumanities scholars. In Nicoletta Calzolari, et al., editors, *LREC 2014 — Proceedings of the 9th International Conference on Language Resources and Evaluation. Reykjavik, Iceland, May 26-31, 2014*, pages 1525–1531, Paris. European Language Resources Association (ELRA).

Hinrichs, E. W., Ide, N. C., Pustejovsky, J. D., Hajič, J., Hinrichs, M., Elahi, M. F., Suderman, K., Verhagen, M., Rim, K., Straňák, P., and Mišutka, J. (2018). Bridging the LAPPS GRID and CLARIN. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 1294–1302, Paris. European Language Resources Association (ELRA).

Hoenen, A. (2018). From manuscripts to archetypes through iterative clustering. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 712–718, Paris. European Language Resources Association (ELRA).

Jayannavar, P. A., Agarwal, A., Ju, M., and Rambow, O. C. (2015). Validating literary theories using automatic social network extraction. In Anna Feldman, et al., editors, *CLfL 2015 — Proceedings of the 4th Workshop on Computational Linguistics for Literature @ NAACL-HLT 2015. Denver, Colorado, USA, June 4, 2015*, pages 32–41, Stroudsburg/PA. Association for Computational Linguistics (ACL).

Kettunen, K. and Pääkkönen, T. (2016). Measuring lexical quality of a historical Finnish newspaper collection: analysis of garbled OCR data with basic language technology tools and means. In Nicoletta Calzolari, et al., editors, *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 956–961, Paris. European Language Resources Association (ELRA).

Kettunen, K. and Ruokolainen, T. (2017). Names, right or wrong: named entities in an OCRed historical Finnish newspaper collection. In Apostolos Antonacopoulos et al., editors, *DATeCH 2017 — Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage. Göttingen, Germany, June 1-2, 2017*, pages 181–186. Association for Computing Machinery.

Kettunen, K., Koistinen, M., and Kervinen, J. (2020). Ground truth OCR sample data of Finnish historical newspapers and journals in data improvement validation of a re-OCRing process. *LIBER Quarterly*, 30(1):1–20.

Kim, S. M. and Cassidy, S. (2015). Finding names in TROVE: named entity recognition for Australian historical newspapers. In Ben Hachey et al., editors, *ALTA 2015 — Proceedings of the Australasian Language Technology Association Workshop 2015. Parramatta, New South Wales, Australia, 8-9 December 2015*, pages 57–65. Australasian Language Technology Association.

Kremer, D. (2010). Ritter Gluck. Eine Erinnerung aus dem Jahre 1809 / Fantasiestücke in Callot's Manier (1814/15) / Nachricht von den neuesten Schicksalen des Hundes Berganza / Der Magnetiseur / Der goldene Topf. In Detlef Kremer, editor, *E.T.A. Hoffmann. Leben – Werk – Wirkung*, pages 81–130. Walter de Gruyter, 2nd, enlarged edition.

Meroño Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., and van Harmelen, F. (2015). Semantic technologies for historical research: a survey. *Semantic Web — Interoperability, Usability, Applicability*, 6(6):539–564.

David Milsom, editor. (2011). *Classical and Romantic Music*. Routledge.

Neubauer, J. (2017). *The Persistence of Voice. Instrumental Music and Romantic Orality*. Brill.

Neudecker, C., Baierer, K., Federbusch, M., Boenig, M., Würzner, K.-M., Hartmann, V., and Herrmann, E. (2019). OCR-D : an end-to-end open source OCR framework for historical printed documents. In Apostolos Antonacopoulos, et al., editors, *DATeCH 2019 — Proceedings of the 3rd International Conference on Digital Access to Textual Cultural Heritage. Brussels, Belgium, May 8-10, 2019*, pages 53–58. Association for Computing Machinery (ACM).

Neudecker, C. (2016). An open corpus for named entity recognition in historic newspapers. In Nicoletta Calzolari, et al., editors, *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, pages 4348–4352, Paris. European Language Resources Association (ELRA-ELDA).

Niekler, A., Bleier, A., Kahmann, C., Posch, L., Wiedemann, G., Erdogan, K., Heyer, G., and Strohmaier, M. (2018). ILCM: a virtual research infrastructure for large-scale qualitative data. In Nicoletta Calzolari, et al., editors, *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, pages 1313–1319. European Language Resources Association.

Oiva, M., Nivala, A., Salmi, H., Latva, O., Jalava, M., Keck, J., Martínez Domínguez, L., and Parker, J. (2019). Spreading news in 1904. The media coverage of Nikolay Bobrikov's shooting. *Media History*, [Epub ahead of print]. https://doi.org/10.1080/13688804.2019.1652090.

Pustejovsky, J. D., Ide, N. C., Verhagen, M., and Suderman, K. (2017). Enhancing access to media collections and archives using computational linguistic tools. In Thierry Declerck et al., editors, *CDH 2017 — Proceedings of the Workshop on Corpora in the Digital Humanities. Bloomington, Indiana, USA, January 19, 2017*, number 1786 in CEUR Workshop Proceedings, pages 19–28.

Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., and Puppe, F. (2019). OCR4ALL : an open-source tool providing a (semi-)automatic OCR workflow for historical printings. *Applied Sciences*, 9(22):#4853 (30pp.).

Alexander L. Ringer, editor. (1990). *Early Romantic Era. Between Revolutions, 1789 and 1848*. Palgrave Macmillan UK.

Scrivner, O. and Davis, J. (2017). Interactive text mining suite: data visualization for literary studies. In Thierry Declerck et al., editors, *CDH 2017 — Proceedings of the Workshop on Corpora in the Digital Humanities. Bloomington, Indiana, USA, January 19, 2017*, number 1786 in CEUR Workshop Proceedings, pages 29–38.

Soni, S., Klein, L. F., and Eisenstein, J. (2019). Correcting whitespace errors in digitized historical texts. In Beatrice Alex, et al., editors, *LaTeCH-CLfL 2019 — Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature @ NAACL-HLT 2019. Minneapolis, Minnesota, USA, June 7, 2019*, pages 98–103. Association for Computational Linguistics (ACL).

Tanner, S., Muñoz, T., and Ros, P. H. (2009). Measuring mass text digitization quality and usefulness. Lessons learned from assessing the OCR accuracy of the British Library's 19th century online newspaper archive. *D-Lib Magazine*, 15(7-8), July.

Trace, C. B. and Karadkar, U. P. (2017). Information management in the humanities: scholarly processes, tools, and the construction of personal collections. *Journal of the Association for Information Science and Technology*, 68(2):491–507, February.

Wang, S., Isaac, A., Schlobach, S., van der Meij, L., and Schopman, B. A. C. (2012). Instance-based semantic interoperability in the cultural heritage. *Semantic Web — Interoperability, Usability, Applicability*, 3(1):45–64.

Widlöcher, A., Béchet, N., Lecarpentier, J.-M., Mathet, Y., and Roger, J. (2015). Combining advanced information retrieval and text-mining for Digital Humanities.

In Christine Vanoirbeek et al., editors, *DocEng '15 — Proceedings of the 15th ACM Symposium on Document Engineering 2015. Lausanne, Switzerland, September 8-11, 2015*, pages 157–166, New York/NY. Association for Computing Machinery (ACM).

Yarasavage, N., Butterhof, R., and Ehrman, C. (2012). National Digital Newspaper Program: a case study in sharing, linking, and using data. In Karim B. Boughida, et al., editors, *JCDL '12 — Proceedings of the 12th ACM/IEEE-CS Joint Conference on Digital Libraries. Washington, D.C., USA, June 10-14, 2012*, pages 399–400, New York/NY. Association for Computing Machinery (ACM).