Corpus Assembly as Text Data Integration from Digital Libraries and the Web

Udo Hahn
Jena University Language & Information Engineering
(JULIE) Lab
Friedrich-Schiller-Universität Jena
Jena, Germany
udo.hahn@uni-jena.de

Tinghui Duan
Graduate School
"Romanticism as a Model"
Friedrich-Schiller-Universität Jena
Jena, Germany
tinghui.duan@uni-jena.de

ABSTRACT

We here explore a new corpus construction workflow which exploits the inherent potential of the growing number of Digital Libraries worldwide and the ever-expanding Internet Archive. Rather than building corpora from scratch (which typically consumes a huge amount of resources), we search the Web for fragments of relevant digitized contents scattered across the world, check their digitization quality, select those digital versions with highest quality, and finally assemble from those pieces an integrated corpus with a maximum coverage of the targeted resource. As a use case within the framework of Digital Humanities, we illustrate this approach for the *Allgemeine Literatur-Zeitung* (*General Literature Gazette*, ALZ) published from 1785 to 1849, which is considered as one of the most important text collections from the Romantic Age in Germany. With lots of incomplete and overlapping fragments physically scattered over many Web sites, we started to assemble these fragments, to bind these pieces together using a homogeneous format, and thus constructed the first (almost) complete corpus of ALZ, now accessible (in XML format obeying to TEI standards) as a whole for in-depth scientific investigations.

CCS CONCEPTS

• **Information systems** → **Digital libraries and archives**; **Digital libraries and archives**; **Extraction, transformation and loading**; *Mediators and data integration*; Data cleaning; • **Applied computing** → Document management and text processing.

KEYWORDS

Digital Humanities, Digital Libraries, Internet Archive, Document Management, German Romanticism, Allgemeine Literatur-Zeitung

ACM Reference Format:
Udo Hahn and Tinghui Duan. 2019. Corpus Assembly as Text Data Integration from Digital Libraries and the Web. In *JCDL '19: ACM/IEEE Joint Conference on Digital Libraries, June 02–06, 2019, Urbana-Champaign, IL, USA.* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

Digital Humanities face a tremendous resource acquisition problem. Typically, data resources (corpora) are very small and originate from single, physically dispersed collections so that often only fragments of complete works, editions, etc. of an author, a journal/newspaper, respectively can be digitized. If available at all, they furthermore suffer from different levels of quality concerning the digital reproduction of the raw data they contain, thus, discouraging subsequent reuse, e.g., by natural language processing analytics. According to a usability survey [2], users of digital libraries are often disappointed by the search results and the Web interface. In particular for literary researchers it is even more difficult to make use of resources provided this way, since they rely on the completeness of the underlying resources and the ease of use of interfaces (due to a lack of technical background knowledge).

We here explore a new approach to alleviate this dilemma, which builds on the inherent potential of the growing number of Digital Libraries worldwide and the ever-expanding Internet Archive. Rather than building corpora from scratch (which typically consumes a huge amount of resources), we take advantage of Linked (Open) Data and the Semantic Web [3, 4, 10]. In particular, we search the Web for fragments of relevant digitized contents scattered across the world, check their digitization quality, select those digital versions with highest quality, and finally assemble from those pieces an integrated corpus with a maximum coverage of the targeted resource.

As a use case, we illustrate this approach for the *Allgemeine Literatur-Zeitung* (*General Literature Gazette*, ALZ) published from 1785 to 1849 in Jena and Halle (Germany), which is considered as one of the most important text collections from the Romantic Age in Germany [11]. It consists of anonymous reviews of contemporary publications of all sorts, partially written by the most eminent representatives of that literary period and thus constitutes the most comprehensive review organ around 1800 in Germany. Therefore, a digital (currently non-existent) corpus covering all of the issues of ALZ would be a true desideratum for both computational and non-computational literary studies dealing with that cultural period.

2 COMPILATION WORKFLOW FOR THE ALZ CORPUS

In order to assemble a composite digital corpus of ALZ, we define the following workflow which will be discussed in more detail in the following subsections:

JCDL '19, June 02–06, 2019, Urbana-Champaign, IL, USA

Table 1: Components of the ALZ collection in Google Books, Internet Archive and the newly assembled full-text corpus

	Google Books			Internet Archive			ALZ Full-Text Corpus		
Library	**volumes**	**pages**	**tokens**	**volumes**	**pages**	**tokens**	**volumes**	**pages**	**tokens**
Bavarian State Library	1	285	289,269	-	-	-	152	78,675	76,568,508
University of Lausanne	4	2,008	1,969,395	-	-	-	-	-	-
Harvard University	4	3,056	2,645,386	4	3,054	2,520,832	2	1,528	1,322,693
Indiana University	67	33,721	32,255,992	49	24,923	25,146,053	3	2,641	2,547,872
New York Public Library	103	48,949	47,909,730	94	42,636	42,970,913	4	2,080	2,043,980
Princeton University	38	36,854	35,245,420	29	29,096	30,036,319	15	8,038	7,267,803
Stanford University	2	887	869,272	3	1,102	1,216,273	-	-	-
University of California	45	28,690	28,032,026	54	33,248	33,179,736	9	3,354	3,162,025
University of Illinois	-	-	-	4	3,192	2,820,205	2	1,512	1,299,100
University of Michigan	144	71,681	68,554,109	308	147,559	146,025,563	64	25,538	23,276,002
University of Oxford	46	20,995	20,506,977	56	26,181	25,185,788	8	3,246	2,881,022
Total	454	247,126	238,277,576	601	310,991	309,101,682	261	126,612	120,369,005

Note: The first column block from the right shows the data of our compiled corpus (ALZ Full-Text Corpus). While all 152 volumes provided by BSB were included, the other volumes were selected from Google Books and Internet Archive. At this point, it should be noted that in case one volume provided by Google Books, Internet Archive or BSB contains more than one volume, it was split into individual ones so that the volume number of the ALZ Full-Text Corpus has a slightly different counting logic.

(1) Collecting all available digitized resources of ALZ from various digital libraries,
(2) Evaluating the collected resources by their OCR quality,
(3) Selecting the best-quality OCR-ed full-texts,
(4) Assembling a(n almost) complete ALZ corpus.

2.1 Collecting all Available Resources from Digital Libraries

In December 2018, we started our compilation project by searching the WWW using the keyphrase "Allgemeine Literatur Zeitung" in a Google search. By following the hints of the search results based on the snowball principle, we were able to find more and more digital versions of ALZ in various digital libraries, including the *HathiTrust Digital Library*,[1] *Internet Archive*,[2] *Google Books*,[3] and several digital platforms of physical libraries such as the *Bavarian State Library* (BSB),[4] the *Austrian National Library* (ÖNB),[5] and the *Thuringian University and State Library* (ThULB).[6] As already mentioned, the digital libraries provide their resources in different technical formats. ThULB, e.g., yields scanned images and manually annotated metadata, but no full-texts, whereas HathiTrust and ÖNB excel with full-texts as well as scanned images, yet, bulk downloading is not possible in a reasonable way.

In the end, we gathered more than 1,000 volumes of full-texts of ALZ and their metadata (if available), in total. Unluckily, we found that many metadata provided by the digital libraries were incorrect. For example, according to the Internet Archive, the publication date of many volumes is 1785, but actually they were issued in later publication years. One might speculate whether a reason for this error could be the fact that 1785 was the starting year of the ALZ collection and this date was inherited by individual volumes as their publication date.[7] For this and other possible reasons, we thoroughly inspected all downloaded volumes and manually completed or corrected the metadata (publication year, volume, source physical library etc.).[8] Table 1 gives a comprehensive overview of the resources we took into account and their quantitative share on several dimensions (volumes, pages and tokens for each source physical library).

As Table 1 shows, the digital resources hosted by Google Books and Internet Archive were originally taken from other physical libraries. In most cases, the same resource is hosted by both Google Books and Internet Archive, yet with different full-text qualities (see Section 2.2). At the same time, we identified volumes in Google Books that were not provided by Internet Archive (e.g., the four volumes from the Cantonal and University Library of Lausanne) and vice versa (e.g., the four volumes from the library of the University of Illinois). Although the Bavarian State Library has a lower amount of digital resources, in total, it provides some unique volumes which are neither available from Google Books nor from the Internet Archive (e.g. issue 1785 vol. 1-3). So it makes sense to put all the sources together and select the best full-text for each unique volume.

2.2 Evaluating the Collected Digital Resources by their OCR Quality

After having collected the available digital resources of ALZ from different digital libraries, we evaluated the quality of the full-texts made available by the institutional providers by calculating the minimum edit distance [8] between them and a manually transcribed ground truth (the larger the number, the worse the resource in terms of text quality).

[1] https://www.hathitrust.org/
[2] https://archive.org/
[3] https://books.google.de/
[4] https://www.bsb-muenchen.de/
[5] https://www.onb.ac.at/
[6] https://zs.thulb.uni-jena.de/
[7] See, e.g., https://archive.org/details/allgemeineliter118unkngoog
[8] This work turned out to become the most time-consuming effort of the whole assembly process.

Table 2: Evaluation of full-text quality

Levenshtein Distance to ground truth	ÖNB	BSB	Google Books	Internet Archive	HathiTrust
Austrian National Library (ÖNB)	758	-	-	-	-
Bavarian State Library (BSB)	-	811	-	-	-
New York Public Library	-	-	847	2726	773
Princeton University	-	-	1176	-	1130
University of California	-	-	1060	2223/1637[a]	1012
University of Michigan	-	-	3466	4608	3423

Note: Pages 297-300, issue 1800, volume 2. Cell entries contain the respective Levenshtein distance value of each OCR-ed version relative to the ground truth.

[a] Two versions of the same source (University of California) are included in Internet Archive, yet with different OCR qualities: https://archive.org/details/allgemeineliter142unkngoog/page/n156; https://archive.org/details/bub_gb_uaU9AAAAIAAJ/page/n154

We will illustrate this process (see Table 2) by the following example from a review of Friedrich Schlegel’s novel “Lucinde”, a famous piece of writing from the Romantic Age. First, we manually transcribed two selected pages (issue 1800, volume 2, pages 297/298[9] and 299/300[10]) as the *ground truth* for our experiment. Then we compared the downloaded *OCR-ed version* available from each digital library with that manually established ground truth. In Table 2, the cell entries depict the minimum edit distance (Levenshtein Distance) [8] of each *OCR-ed version* relative to the manually defined *ground truth.* In future work, we might envisage to consider automatic means for OCR error correction [1, 12].

2.3 Selecting the Best-Quality OCR-ed Full-Texts

Although (based on our assessment, see Table 2) ÖNB seems to offer the highest quality for full-texts, their sources could not be included, since they do not allow bulk downloading (no special agreement could be negotiated). A similar argument applies to HathiTrust.

Among all those sources which are in line with our procedure and based on the exemplary evaluation in Subsection 2.2, we decided on the inclusion of documents into our corpus applying the following rule: Full-texts from the Bavarian State Library (BSB) are preferably included. These amount to 152 volumes (78,675 pages or 76,568,508 tokens). If not available via BSB, we preferably included the full-texts from Google Books, which amount to 106 volumes (45,830 pages or 42,011,823 tokens). The remaining volumes are taken from the Internet Archive. Since the Internet Archive provides only full-texts of entire volumes but we wanted to get full-texts of each scanned page, we post-processed the scanned pages downloaded from the Internet Archive with Tesseract 4.0;[11] and then got full-texts of 3 additional volumes (2,107 pages or 1,788,674 tokens).

2.4 Final Assembly of the ALZ Corpus

We compiled all full-texts of unique volumes of ALZ that we either picked up directly from BSB and Google Books as is, or that we OCR-ed (from scanned images downloaded from Internet Archive) using Tesseract 4.0. The ALZ corpus currently contains 126,612 pages of full-texts from 261 volumes which is equivalent to 120,369,005 tokens, including review volumes (the main part of ALZ), supplementary and intelligence notes. About 54 volumes were not included because they are neither provided by BSB, nor by Google Books, nor by the Internet Archive. In the future, we will struggle to complete the anthology corpus by cooperating with libraries that provide digital resources, scanned images or even printed versions of the missing volumes. Currently, based on our best estimate, our corpus covers about 82% of the entire ALZ.[12]

For downloading and post-processing, the corpus is available in XML format and accessible via https://github.com/JULIELab//alz. Relevant metadata are included as well, such as publication year, volume, source, reference, document provider, OCR method, licence and size (number of pages and tokens for each volume).

3 RELATED WORK

Corpora dealing with the Romantic Age are rare and typically medium- or small-sized. Hellrich & Hahn [5] list three corpora for the English language (the fiction part of Google Books Ngrams Corpus (GBF), Corpus of Historical American English (COHA), Royal Society Corpus (RSC)) and two for German (Deutsches Textarchiv (DTA) and the German part of Google Books Ngrams Corpus (GBG)). Their coverage of the Romantic period (roughly ranging from 1780 to 1850) is in the range of several millions of tokens, yet their coverage of literary texts varies a lot (e.g., GBF contains hundreds of millions of n-grams (up to pentagrams) from fiction writing, but RSC incorporates only scientific writing since the documents are taken from the Philosophical Transactions of the Royal Society of London).

As far as German language is concerned, the *Deutsches Textarchiv* (DTA)[13] constitutes the largest compilation of historical text documents, with approximately 55 million tokens from the Romantic

[9] https://api.digitale-sammlungen.de/iiif/presentation/v2/bsb10502018/canvas/233/view

[10] https://api.digitale-sammlungen.de/iiif/presentation/v2/bsb10502018/canvas/234/view

[11] https://github.com/tesseract-ocr/tesseract/wiki/4.0-with-LSTM we used the best (most accurate) trained LSTM model for German (https://github.com/tesseract-ocr/tessdata_best)

[12] It was not easy to measure the accurate coverage, since the scopes of individual volumes vary a lot. One volume can contain the reviews of three, four or six calendar months. Furthermore, in the first publishing years, no volumes of intelligence notes were issued.

[13] http://www.deutschestextarchiv.de/

period.[14] Project Gutenberg[15] hosts several publications from the Romantic Age (e.g., books of Charlotte von Ahlefeld (1781-1849), Ernst Moritz Arndt (1769-1860) or Bettina von Arnim (1781-1831) and Ludwig Achim Freiherr von Arnim (1785-1859), to mention a few). They are quite heterogeneous in terms of text genres, but constitute another reasonable resource for research on the Romantic Age. Even much lower numbers in size are reported by Lücking et al. [9] whose TGermaCorp corpus comprises only 122,902 tokens. Hence, the ALZ Corpus we constructed (with over 100 millions of tokens) constitutes, in terms of the number of tokens, the largest resource and contributes a substantial increase in the number of German-language documents available for that literary period.

It should be noted that our contribution does not belong to the field of reconstruction of ancient texts out of several (noisy) historical exemplars (cf., e.g., [6, 7]) which focuses on multiple corrupted versions of the same text and attempts to reconstruct the original text from which the extant corrupted versions were copied (typically via latent intermediary versions).

4 CONCLUSIONS AND FUTURE WORK

The standard procedure for creating a corpus from historical resources is to scan the original paper documents and produce a digital copy (typically in PDF format). Subsequently the e-copy is transformed – using some form of OCR processing – into a format that can be used for analytic purposes (full-text search or more advanced natural language processing), i.e., plain or XML-encoded text files. This process is typically expensive (high-quality scanners have to be purchased, qualified staff has to be hired and paid) and technically troublesome (the quality of OCR software and historical resources differs from institution to institution; the older the original resources are – often with ancient lettering, e.g., German Fraktur –, the worse the output of OCR) [13].

As a resource-savvy alternative, we here propose to reuse already existing fragments of digitized historical contents and compile an integrated corpus. Such an approach seems to be a viable alternative, since it builds on the precious work of others who have run through the painful processes outlined above. Thanks to the often hidden potential of the World Wide Web, such kind of data integration from physically scattered digital libraries has become possible now but, to the best of our knowledge, not been tried before.

Based on the current state of the integrated ALZ corpus, we will continue to complete the whole ALZ corpus (see Section 2.4), select all reviews with direct reference and immediate relevance to the Romantic Age as the size-wise largest resource for (non-) computational literary studies of this literary period. A link to additional metadata (as available from the ThULB) could further enrich the ALZ corpus.

Once this work will be finished, literary scholars are able to conduct quantitative studies on the semantics of documents from the Romantic Age using original, in-time resources.

ACKNOWLEDGMENTS

This work was conducted within the Graduate School "Romanticism as a Model" (http://modellromantik.uni-jena.de) supported by the German Research Foundation (DFG) under Grant No.: GRK 2041/1.

The authors would like to thank the anonymous referees for their valuable comments and helpful suggestions.

REFERENCES

[1] Haithem Afli, Zhengwei Qiu, Andy Way, and Páraic Sheridan. 2016. Using SMT for OCR error correction of historical texts. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan E. J. M. Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA-ELDA), Paris, 962–966.

[2] Maristella Agosti, Franco Crivellari, Giorgio Maria Di Nunzio, and Silvia Gabrielli. 2010. Understanding user requirements and preferences for a digital library Web portal. *International Journal on Digital Libraries* 11, 4 (2010), 225–238.

[3] Karen Coyle. 2013. Linked data as a new paradigm of data interconnection linked data: an evolution. *Italian Journal of Library, Archives, and Information Science* 4 (Special Issue on Global Interoperability and Linked Data in Libraries, ed. by Mauro Guerrin), 1 (January 2013), 53–61.

[4] Brighid M. Gonzales. 2014. Linking libraries to the Web: linked data and the future of the bibliographic record. *Information Technology and Libraries* 33, 4 (December 2014), 10–22.

[5] Johannes Hellrich and Udo Hahn. 2017. Exploring diachronic lexical semantics with JeSemE. In *ACL 2017 — Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. Vancouver, British Columbia, Canada, August 1, 2017*, Heng Ji and Mohit Bansal (Eds.). Association for Computational Linguistics (ACL), Stroudsburg/PA, 31–36.

[6] Armin Hoenen. 2015. Lachmannian archetype reconstruction for ancient manuscript corpora. In *NAACL-HLT 2015 — Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Denver, Colorado, USA, May 31 - June 5, 2015.* Association for Computational Linguistics (ACL), Stroudsburg/PA, 1209–1214.

[7] Moshe Koppel, Moty Michaely, and Alex Tal. 2016. Reconstructing ancient literary texts from noisy manuscripts. In *CLfL 2016 — Proceedings of the 5th Workshop on Computational Linguistics for Literature @ NAACL-HLT 2016. San Diego, California, USA, June 16, 2016*, Anna Feldman, Anna Kazantseva, and Stan Szpakowicz (Eds.). Association for Computational Linguistics (ACL), Stroudsburg/PA, 40–46.

[8] Vladimir I. Levenshtein. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10, 8 (1966), 707–710.

[9] Andy Lücking, Armin Hoenen, and Alexander Mehler. 2016. TGermaCorp: a (Digital) Humanities resource for (computational) linguistics. In *LREC 2016 — Proceedings of the 10th International Conference on Language Resources and Evaluation. Portorož, Slovenia, 23-28 May 2016*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan E. J. M. Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA-ELDA), Paris, 4271–4277.

[10] Vincenzo Maltese and Fausto Giunchiglia. 2016. Search and analytics challenges in digital libraries and archives. *Journal of Data and Information Quality* 7, 3 (September 2016), #10.

[11] Stefan Matuschek. 2004. Epochenschwelle und prozessuale Verknüpfung. Zur Position der Allgemeinen Literatur-Zeitung zwischen Aufklärung und Frühromantik. In *Organisation der Kritik Die ‚Allgemeine Literatur-Zeitung' in Jena 1785-1803*, Stefan Matuschek (Ed.). Number 5 in Ereignis Weimar-Jena. Kultur um 1800. Ästhetische Forschungen. Universitätsverlag Winter, Heidelberg, 7–17.

[12] Caitlin Richter, Matthew Wickes, Deniz Beser, and Mitchell P. Marcus. 2018. Low-resource post processing of noisy OCR output for historical corpus digitisation. In *LREC 2018 — Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki, Japan, May 7-12, 2018*, Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan E. J. M. Odijk, Stelios Piperidis, and Takenobu Tokunaga (Eds.). European Language Resources Association (ELRA), Paris, 2331–2339.

[13] Christoph Stollwerk. 2016. *Machbarkeitsstudie zu Einsatzmöglichkeiten von OCR-Software im Bereich "Alter Drucke" zur Vorbereitung einer vollständigen Digitalisierung deutscher Druckerzeugnisse zwischen 1500 und 1930.* Niedersächsische Staats- und Universitätsbibliothek Göttingen, Göttingen.

[14] This number is based on the tcf-version (2018-10-18) which can be downloaded from http://www.deutschestextarchiv.de/download#tcf

[15] https://www.gutenberg.org/