Die Grenzboten on its Way to Virtual Research Environments and Infrastructures

Manfred Nölte and Martin Blenkle

Journal of European Periodical Studies, 4.1 (Summer 2019)

ISSN 2506-6587

The *Journal of European Periodical Studies* is hosted by Ghent University

Website: ojs.ugent.be/jeps

To cite this article: Manfred Nölte and Martin Blenkle, '*Die Grenzboten* on its Way to Virtual Research Environments and Infrastructures', *Journal of European Periodical Studies*, 4.1 (Summer 2019), 19–35

Die Grenzboten on its Way to Virtual Research Environments and Infrastructures

Manfred Nölte and Martin Blenkle
State and University Library Bremen
noelte@suub.uni-bremen.de and blenkle@suub.uni-bremen.de

ABSTRACT

The State and University Library Bremen (SuUB) is dedicated to the digitization of its historical collections. Digitization is an important instrument for improving the accessibility of valuable information contained in fragile historical documents. It facilitates academic research and teaching and is indispensable to the digital humanities. Especially the research of digital serial publications benefits from, as Michael Piotrowski puts it, 'recent systematic digitization efforts, often initiated by libraries […]. More and more historical periodicals and other serial publications are now digitally available in full, i.e. all of their issues'.[1] The historical journal presented in this article is one of these, and the final section will discuss why it can be considered a complete corpus.

Usually, digitization projects produce digital images, metadata for cataloguing and web-navigation purposes, and optical character recognition (OCR) full text for searching. This information is made available through the library's web portal for digital collections. However, digital humanists need high-quality full texts enriched with metadata in the appropriate format to analyze them with powerful software tools.

The historical journal *Die Grenzboten* serves as an exemplary model to bridge the gap between digitization projects in libraries and research infrastructures. *Die Grenzboten* is a long running serial publication (1841–1922). It can be classified as a literary journal that also covered politics and arts. We demonstrate that OCR post correction and a page-wise structuring are prerequisites for the creation of a high-quality Text Encoding Initiative (TEI) version of a full text. The TEI version was created in cooperation with the Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW). A fully automated OCR post correction developed at SuUB is freely available on GitHub.

To enable scientists to work with powerful software tools, the transfer of high-quality full texts to research infrastructures is a necessary step. We describe transfers of full text and the experience we have had, but still some general questions persist: What has to be done to prepare raw OCR output for this purpose in a reasonable and cost-effective manner? What quality is needed or expected? Which metadata and file formats are needed? Should there not be a closer cooperation between research infrastructures and libraries handling the digitization? OCR full texts, even post corrected, are not

1 Michael Piotrowski, 'Historical Models and Serial Sources', *Journal of European Periodical Studies*, 4.1 (2019), 8–18.

perfect but character recognition rates around 99% certainly provide more options than just being used as a search index. There is a vast amount of textual resources available, ready to be made fully accessible for scientific research! Finally, some suggestions for scholars and the researchers working on digital serial publications are given.

KEYWORDS

Digitization, full text, OCR post correction, research infrastructures

Introduction

Since 1999, the State and University Library Bremen (SuUB) has been dedicated to the digitization of its historical collections, such as historical maps, publications of Bremen's regional history, or material of interest to scientists, such as historical journals or German seventeenth-century newspapers. Digitization is an important instrument for improving the accessibility of valuable information contained in fragile historical documents. It facilitates academic research and teaching. Especially in combination with the generation of full texts, it is indispensable to the digital humanities and other fields such as linguistics. In the following, we present the origin of the full text, and describe the characteristics and the level of transcription errors. We show what the scholar or the researcher has to consider when dealing with full texts originating from digitization projects. Best results will be obtained if the intended use, the full text quality, and the approaches and methods fit together: 'for example, information retrieval is generally more tolerant towards transcription errors than linguistic analysis'.[2] The final section, 'Conclusions for the Digital Humanities', discusses this issue in more detail, and not only with respect to transcription errors.

Usually, digitization projects produce digital images, metadata for cataloguing and web-navigation purposes, and — if possible — optical character recognition (OCR) full text for searching. This information is made available through the library's web portals dedicated to digital collections. There is a wide range of different types of portal software systems for digitized material and hence different ways to access images, metadata, and full texts via the existing interfaces.

The scientists working with this material often do not have the knowledge or the time to access this material directly via these various interfaces. There is a need for easy, accessible, high-quality full texts enriched with metadata in the right format to be able to analyze them with powerful software tools.[3] This need is not restricted to the digital humanities. The historical journal *Die Grenzboten* serves as an exemplary model to bridge this gap between digitization projects in libraries and the requirements of the digital humanities. To achieve this, it seems most reasonable to transfer the digital material to research infrastructures such as CLARIN-D, DARIAH-DE/Textgrid or the DTA.[4] Hence the question is: What has to be done to prepare raw OCR output for this purpose in a reasonable and cost-effective manner?

To answer this question, we will describe and discuss two projects funded by the Deutsche Forschungsgemeinschaft (DFG). The first project was solely a digitization project subsequently followed by the second one, which was already targeted at the transfer to the mentioned research infrastructures. To deal with over 185,000 pages in a cost-effective manner, we developed a fully automatic post correction software at SuUB, that was able to eliminate 2.84 million character errors, achieving a character recognition rate of 98.83%. In the following sections we will describe and assess this reduction of about 32% of all character errors. As a second aspect of text quality, we enhanced the level of document structure according to an agreed standard format in

2 Michael Piotrowski, *Natural Language Processing for Historical Texts*, Synthesis Lectures on Human Language Technologies (Lexington, KY: Morgan & Claypool, 2012).

3 Thomas Stäcker, 'Konversion des kulturellen Erbes für die Forschung: Volltextbeschaffung und -bereitstellung als Aufgabe der Bibliotheken', *o-bib*, 1.1 (2014), 220–37 [accessed 14 June 2018].

4 The German part of the 'Common Language Resources and Technology Infrastructure' (CLARIN-D); the German part of the 'Digital Research Infrastructure for the Arts and Humanities' (DARIAH-DE/Textgrid); and the Deutsches Textarchiv (DTA) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) [all accessed 24 May 2018].

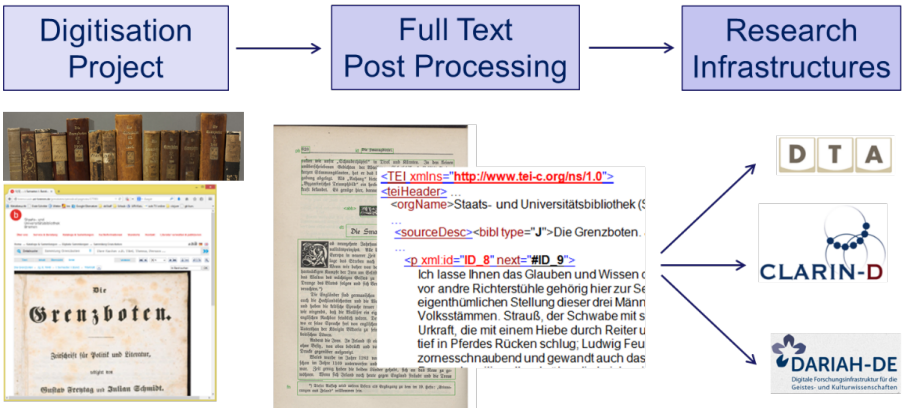


Fig. 1 Transfer of full text from a finished digitization project to research infrastructures

cooperation with our partners, the DTA. Using this structure information, we converted the OCR output format to an interoperable Text Encoding Initiative (TEI) format.

Actively supporting these processes as a digitizing library will result in considerably improved outcome in all fields of automated and computer-aided research across disciplines working with digitized material. It will enable the employment of quantitative methods and approaches such as authorship attribution studies, clustering techniques (i.e. for literary genre analysis), Topic Modeling, etc.[5]

Setting up the above-mentioned processes in cooperation with research infrastructures will have effects of standardization and centralization, enabling the matching of software tools requirements and digital full texts.

The SuUB is in contact with various research groups within the fields of German philology, linguistics, Topic Modeling, full text quality improvement, and research infrastructures. An example of the former is a cooperation with a research group at the University of Bremen conducting a project on the exploration of so-called 'Bildprosa'. This is a textual phenomenon within the nineteenth-century prose, a characteristic of which is the use of 'image terms' ('Bildbegriffe') as a reaction to the increasing differentiation of knowledge in that period. Examples of these terms are 'Zeit-, Welt-, Lebensbilder'; 'Reise-/Kulturbilder'; 'Charakterbilder'; 'Sittengemälde'; 'Umrisse'; and 'Skizzen, Studien'.

Undoubtedly, within this context there is potential for agile modelling.[6] So far, our first approach is the 'mere' identification of candidates of interesting sections. The methodological space between the 'mere identification of sections of interest' and agile modelling is an example of the future scale of possibilities in the digital humanities (DH). Fig. 2 illustrates some kind of 'methodological evolution' from the state of the art, via progress within the theoretical digital humanities to a future methodological change.

5 Shawn Graham, Scott Weingart, and Ian Milligan, 'Getting Started with Topic Modeling and MALLET', *UWSpace* (2012) and Andrew Kachites McCallum, *MALLET: A Machine Learning for Language Toolkit* (2002) [both accessed 14 June 2018].

6 In this volume, Piotrowski mentions the potential 'to create computational models […] in an agile manner'. See also Fig. 2.

Convenience, an aspect discussed by Michael Piotrowski as 'applying mechanical tools to automate tedious work'.[7]
→
DH as an 'enabling technology': 'enabling new approaches by automation'
'On the other hand, what is new is that […] 'complete corpora' of periodicals are now available, easily accessible, and can be analyzed automatically — thus enabling historians to quickly perform analyses that used to be practically impossible before because they would have required too much time and manual labour.'
→
Progress [within the *theoretical digital humanities*]: improving and applying single DH methods (or workflows)
→
Methodological progress: 'computational agile modelling'
A step towards a 'methodological change […] to create computational models that make assumptions explicit, can be rigorously tested, revised, and shared in an agile manner'.

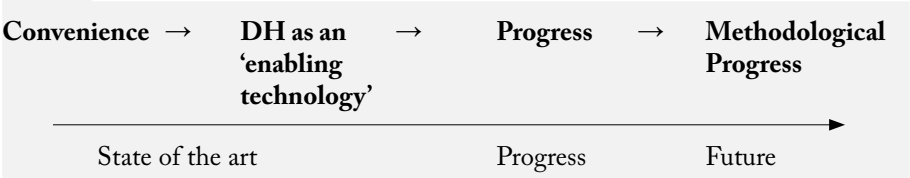


Fig. 2 Methodological evolution

The Digitization Project *Die Grenzboten*

The journal *Die Grenzboten* is a long-running serial publication covering eight decades of German history. It was founded in 1841 by Ignaz Kuranda in Brussels, and later published in Leipzig and Berlin. As of 1871, the journal was subtitled *Zeitschrift für Politik, Literatur und Kunst*. The publication of this periodical was discontinued in 1922. It can be classified as a literary journal that also covered politics and arts. The long publication period allows for the analysis of continuity and change of cultural values as well as media structures of German nationalism. As one of the most important journals of the nineteenth century, *Die Grenzboten* is an outstanding source for historical science, cultural studies (e.g. literature, arts, and musical history), press history, and more.

Funded by the DFG, the digitization project ran at the SuUB from November 2011 to April 2013. We digitized more than 185,000 single pages in 270 volumes. Almost 33,000 articles were digitized via OCR, and the titles of the articles were manually captured. The resulting OCR full text was processed by the OCR software ABBYY Finereader 9, and consists of approximately 500 million characters and 65 million tokens. The digitization was conducted according to the DFG practical guidelines on digitization following version of April 2009.[8] We integrated the results of this project into the Digital Collections of the State and University Library Bremen (the SuUB). The digitization management and portal software Visual Library is provided by the companies semantics and Walter Nagel. During the digitization project we were in

7 Michael Piotrowski, 'Historical Models and Serial Sources', *Journal of European Periodical Studies*, 4.1 (2019), 8–18. All quotes in Fig. 2 are from this source.

8 *DFG-Praxisregeln: Digitalisierung* (Bonn: Deutsche Forschungsgemeinschaft, 2016) [accessed 8 June 2018] and *DFG-Praxisregeln: Digitalisierung* (Bonn: Deutsche Forschungsgemeinschaft, 2009) [accessed 14 June 2018].

contact with scientists requesting full text, but at that time we could provide only plain text or ABBYY XML files.

The next section will discuss quality issues of the full text, which fundamentally depends on the typeset and of course on the quality of the scanned images. Fig. 3 shows that we were able to automatically identify problematic pages such as advertisements (having diverse layouts and typesets), tables and pages containing foreign languages, Antiqua type, or low image quality (low contrast, bad type face, small font size, translucent and damaged pages).

Fig. 3 Sample results of an automated quality assurance using OCR full text

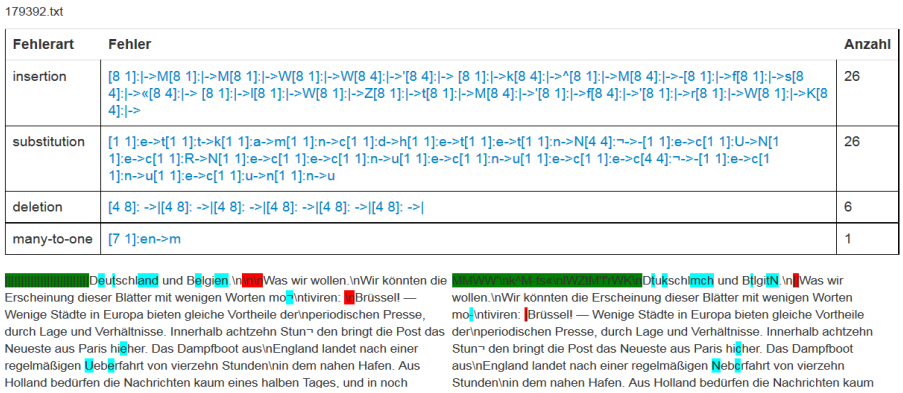
OCR post correction

OCR post correction and enhancing the level of document structure (see next section) were the main objectives of a subsequent — and also DFG-funded — project described here. The DTA at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) joined this project as a partner.

Fig. 4 Different types of gothic/black letter font

As a result of the above-mentioned digitization project we calculated a character recognition rate of 98.28%. This meant every fifty-eighth character was faulty. This is due to the gothic typesets, which are quite difficult to process by OCR (see Fig. 4). The characteristics of these typesets are predominantly using thick vertical bars and relatively thin horizontal hairlines, making it difficult for an OCR to distinguish between the letters 'u' and 'n', or 'c' and 'e', for example.

In order to be able to generate statistics for the OCR character errors we created 300 pages of ground truth text (i.e. error-free full text to compare with) and developed the software tool ocr-visualizer (see Fig. 5), which is able to align OCRed text and ground truth text (or post corrected text) and count character errors. Furthermore, ocr-visualizer is able to count different types of character errors (insertions, substitutions, and deletions) also classifying different types of glyphs (alphabet and special characters, numbers, diacritics, and punctuation marks). We generated statistics for the most



179392.txt

Fehlerart	Fehler	Anzahl
insertion	[8 1]:|->M[8 1]:|->M[8 1]:|->W[8 1]:|->W[8 4]:|->'[8 4]:|-> [8 1]:|->k[8 4]:|->^[8 1]:|->M[8 4]:|->-[8 1]:|->f[8 1]:|->s[8 4]:|->«[8 4]:|-> [8 1]:|->l[8 1]:|->W[8 1]:|->Z[8 1]:|->t[8 1]:|->M[8 4]:|->'[8 1]:|->f[8 4]:|->'[8 1]:|->r[8 1]:|->W[8 1]:|->K[8 4]:|->	26
substitution	[1 1]:e->t[1 1]:t->k[1 1]:a->m[1 1]:n->c[1 1]:d->h[1 1]:e->t[1 1]:e->t[1 1]:n->N[4 4]:¬->-[1 1]:e->c[1 1]:U->N[1 1]:e->c[1 1]:R->N[1 1]:e->c[1 1]:e->c[1 1]:n->u[1 1]:e->c[1 1]:n->u[1 1]:e->c[1 1]:e->c[4 4]:¬->-[1 1]:e->c[1 1]:n->u[1 1]:e->c[1 1]:u->n[1 1]:n->u	26
deletion	[4 8]: ->|[4 8]: ->|[4 8]: ->|[4 8]: ->|[4 8]: ->|[4 8]: ->|	6
many-to-one	[7 1]:en->m	1

Deutschland und Belgien.\n\nWas wir wollen.\nWir könnten die Erscheinung dieser Blätter mit wenigen Worten mo-\ntiviren: Brüssel! — Wenige Städte in Europa bieten gleiche Vortheile der\nperiodischen Presse, durch Lage und Verhältnisse. Innerhalb achtzehn Stun¬ den bringt die Post das Neueste aus Paris hi¬her. Das Dampfboot aus\nEngland landet nach einer regelmäßigen Ue¬berfahrt von vierzehn Stunden\nin dem nahen Hafen. Aus Holland bedürfen die Nachrichten kaum eines halben Tages, und in noch

Deutschland und Belgien.\n\nWas wir wollen.\nWir könnten die Erscheinung dieser Blätter mit wenigen Worten mo-\ntiviren: Brüssel! — Wenige Städte in Europa bieten gleiche Vortheile der\nperiodischen Presse, durch Lage und Verhältnisse. Innerhalb achtzehn Stun¬ den bringt die Post das Neueste aus Paris hi¬her. Das Dampfboot aus\nEngland landet nach einer regelmäßigen Ue¬berfahrt von vierzehn Stunden\nin dem nahen Hafen. Aus Holland bedürfen die Nachrichten kaum

Fig. 5 The software tool ocr-visualizer also generates statistics for the OCR character errors

frequent OCR character errors, which subsequently were used for the parametrization of the post correction algorithm.

We developed an almost fully automated OCR post correction tool at SuUB, keeping in mind that a post correction of millions of pages (coming up within the next digitization projects) should be cost-effective. Our approach is based on the assumption that predominantly typical OCR errors occur, such as the character errors 'e'/'c', 'u'/'n', 's'/'f', etc.[9] These character substitutions together with a factor based on the frequency operate as a simple error model. This might differ for varying OCR systems, typesets, and quality of scanned images. We also considered multi-character substitutions such as 'm'/'in' or 'en'/'m'. We collected this information in a list of typical OCR errors, and used it as a parameter for the algorithm described below.

Table 1 List of historical word forms

Frequency	Surface Form	Transliteration into the subset of ISO-8859-1 (Latin-1)	Modern Form
278187	und	und	und
239487	der	der	der
233389	die	die	die
...			
28268	iſt	ist	ist
14885	ſo	so	so
...			
600	Aehnlichkeit	Aehnlichkeit	Ähnlichkeit
322	Säugethiere	Säugethiere	Säugetiere
319	theilt	theilt	teilt
...			
6	ältlicher	ältlicher	ältlicher
1	̊ältlicher	ältlicher	ältlicher

A list of 1.7 million historical word forms from nineteenth-century corpora, provided by the DTA, was used to check against (see Tab. 1). The algorithm is fully automated, only depending on the above-mentioned parametrization and some heuristics, such as 'do not correct words that appear in the list of historical word forms'.

9 This assumption was confirmed by the above-mentioned digitization project.

With a runtime of four hours and fifteen minutes, the algorithm was efficiently matching the eighty million tokens from the journal *Die Grenzboten* against the list of 1.7 million historical word forms. In addition, it had to take into account eighty-two character substitutions and multi-character substitutions together with their frequencies. One can imagine this matching process such as the implementation of a word distance respecting OCR errors. For each of the roughly eighty million words the algorithm does the following:

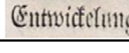
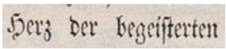
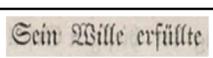
- Check the word against the list of 1.7 million historical word forms
- If there is a match: do nothing
- If there is no match:
  - Create modified words using the list of typical OCR errors (candidates)
  - Sort these candidates by the distance to the initial word
  - An evaluation function takes this sorting together with criteria such as frequency of the candidate word and the frequency of the OCR character error
  - Take best candidate with respect to the evaluation function

The erroneous word 'gelden', for example, leads to the following sorted list mentioned in the outline of the algorithm:

'gelden' → **gelten**, gelben, gelden, gelder, gelbem, gelteu, gelber, getten, geiten

With a runtime of four hours and fifteen minutes, we were able to eliminate 2.84 million character errors, reaching a character recognition rate of 98.83%, reducing all character errors by about 32%. Some correction examples are *Eutwicklimg/Entwickelung*[10] (using two-character substitutions: 'u'/'n', 'c'/'e', and a multi-character substitution: 'im'/'un'), and 'crsüiltc'/'erfüllte', having four character errors within eight characters. Due to the above-mentioned heuristic we were not able to correct words such as 'Aber'/'Ader' or 'dem'/'dein'. More details of these results were published in 2016.[11]

Table 2 OCR post correction examples

Print Image	Ground Truth	OCR-Text	Correction
	Aber	Ader	Ader
	dem	dein	dein
	Zeitschrift	Zeitschrist	Zeitschrift
	gewonnen	gcwouneu	gewonnen
	Entwickelung	Eutwickclimg	Entwickelung
	Herz der begeisterten	Hcrz dcr bcgeisterteu	Herz der begeisterten
	Sein Wille erfüllte	Sciu Willc crsüiltc	Sein Wille erfüllte

10 'Entwickelung' is a historical version of 'development' ('Entwicklung').

11 Manfred Nölte, Jan-Paul Bultmann, Maik Schünemann, and Martin Blenkle, 'Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift *Die Grenzboten*', *o-bib*, 3.1 (2016), 32–55 (p. 32) [accessed 14 June 2018].

Why couldn't we have more corrected errors? This is due to three factors. First, the list of historical word forms itself was erroneous and of course not complete (see estimation of the theoretical limitations below). These are examples of two major types of errors:

- '20pferdekräftiger', '1874–86', 'ἀλφη-', 'partic', and 'essayaient'
- 'uud', 'Dentschland'

The first type of error was not a problem. The algorithm tolerates these character strings to be part of the list of word forms. It is most unlikely for the above described algorithm to produce false-positive corrections because of these character strings. A bigger problem are words such as 'uud' and 'Dentschland'. This directly hinders the correction of typical and frequent OCR errors.

Secondly, we did not take into account the context of a potential error, for example using ngrams or grammatical information. Combined with context information there would have been a stronger capability to correct errors such as 'Aber'/'Ader' or 'dem'/'dein' shown in Tab. 2. Thirdly, altogether the error model, the heuristics, and the whole approach have their theoretical limitations. For example, the list of typical OCR errors was not complete, and will never be so. Extending it massively will increase the runtime significantly. Furthermore, the approach does not contain the concept to correct errors by splitting or merging words. 'erstauntenDeutschlaude' or 'Dentschlandssorgen' are examples or errors that need a splitting and the correction of split words. We roughly estimated the theoretical limitations of the algorithmic approach pretending to work with a complete and error-free list of historical word forms. This list was generated based on 370 pages of ground truth text. The result was that the algorithmic approach never would have been able to produce a character recognition rate above 99.22%. This means that the list of word forms has a great influence, but it is also neatly quantified at what level we need more powerful concepts within the algorithmic approach.

Fig. 6 shows an example of a post corrected page of the journal *Die Grenzboten*. On the left side there is uncorrected full text with character errors highlighted in light blue. The text on the right is showing much less remaining character errors. This graphic was generated automatically by the above-mentioned software ocr-visualizer.

During the second project phase we also looked for further OCR post correction tools and other approaches.[12] At the DATeCH conference (Madrid 2014) we became acquainted with the cloud service OverProof, which offers an online correction service for the English language with a sophisticated approach. In comparison with our straightforward approach, the OverProof approach uses a more powerful framework.[13] Within our project we took the opportunity to be directly in touch with the scanned and OCRed material and we had all the tools, requirements, and more supporting material from our project partners. An example is the list of historical word forms that OverProof didn't have in 2014. However, OverProof is capable of being retrained to different languages from different time periods. Its error model performs a deep

12 Lenz Furrer and Martin Volk, 'Reducing OCR Errors in Gothic-Script Documents', *ERCIM News*, no. 86 (2011), 29–30; Thorsten Vobl, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz, 'PoCoTo: An Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts', in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (New York, NY: Association for Computing Machinery, 2014), pp. 57–61; and Florian Fink, *Postcorrection Tool (PoCoTo) Manual* (Munich, Centrum für Informations- und Sprachverarbeitung, 2015) [all accessed 14 June 2018].

13 John Evershed and Kent Fitch, 'Correcting Noisy OCR: Context Beats Confusion', in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage*, pp. 45–51 [accessed 11 June 2018].

Fig. 6 Example of a post corrected page

probabilistic search for corrections involving multiple character edits per word, and it is also able to correct errors involving splitting or joining words.

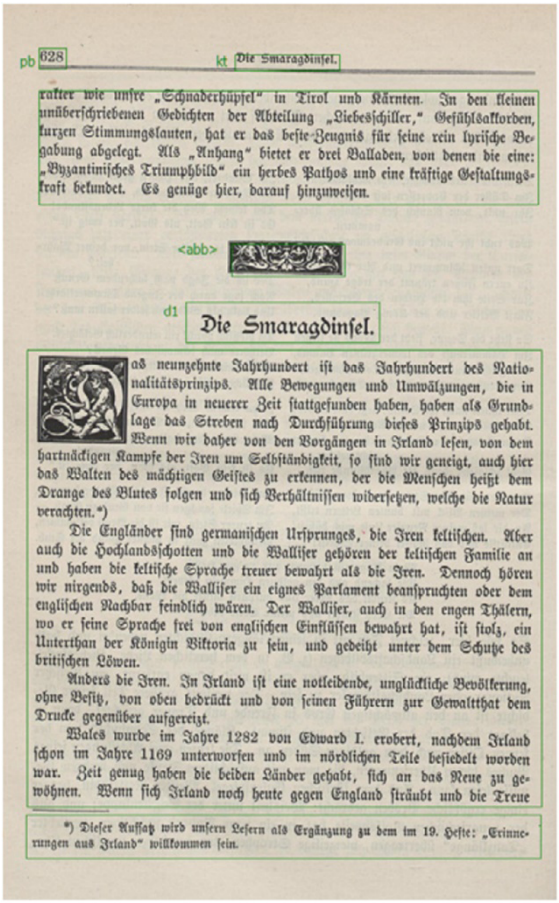
In a very fruitful cooperation, the cloud service OverProof was enhanced by the ability to correct historical German texts that have been OCRed out of gothic font. Now it is provided fee-based by the Australian company ProjectComputing as an easy-to-use web or cloud service which also processes ABBYY XML files. Called file by file for over 185,000 single pages representing files the correction process took three days and twenty-two hours. The correction performance was comparable to the above-mentioned results. ProjectComputing continues to develop the OverProof cloud service.

Creation of a TEI Version

At the DTA all 185,000 pages have been structured using page-wise image clipping, as during the project the automation of this process was not solved satisfactorily. The DTA aims at providing historical documents with very high quality, especially regarding character recognition quality and document structure. That should allow for interoperability: Structural elements such as headings, headers, notes, or footers have been tagged following the TEI-XML-based DTABf,[14] and are thus separated from the text body. This allows the correct treatment of hyphenations at the bottom of pages, for example, to separate the actual text from running headers, notes, marginalia, etc. Structuring the text transcription also makes computational processes such as sentence segmentation, part-of-speech tagging, or syntactic parsing much easier as no pre-processing has to be applied to separate the main text from elements interrupting the linear text flow (e.g. captions or footnotes). Another example to illustrate the importance

14 Susanne Haaf, Alexander Geyken, and Frank Wiegand, 'The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources', *Journal of the Text Encoding Initiative*, no. 8 (2014/15), no page.

of this step is the computation of term frequency: to avoid overestimations, terms in running headers, reoccurring on every page, should not be counted.


pb 628 kt Die Smaragdinsel.

rakter wie unsre „Schnaderhüpfel“ in Tirol und Kärnten. In den kleinen unüberschriebenen Gedichten der Abteilung „Liebesschiller,“ Gefühlsakkorden, kurzen Stimmungslauten, hat er das beste Zeugnis für seine rein lyrische Begabung abgelegt. Als „Anhang“ bietet er drei Balladen, von denen die eine: „Byzantinisches Triumphbild“ ein herbes Pathos und eine kräftige Gestaltungskraft bekundet. Es genüge hier, darauf hinzuweisen.

<abb>

d1 Die Smaragdinsel.

Das neunzehnte Jahrhundert ist das Jahrhundert des Nationalitätsprinzips. Alle Bewegungen und Umwälzungen, die in Europa in neuerer Zeit stattgefunden haben, haben als Grundlage das Streben nach Durchführung dieses Prinzips gehabt. Wenn wir daher von den Vorgängen in Irland lesen, von dem hartnäckigen Kampfe der Iren um Selbständigkeit, so sind wir geneigt, auch hier das Walten des mächtigen Geistes zu erkennen, der die Menschen heißt dem Drange des Blutes folgen und sich Verhältnissen widersetzen, welche die Natur verachten.*)

Die Engländer sind germanischen Ursprunges, die Iren keltischen. Aber auch die Hochlandsschotten und die Walliser gehören der keltischen Familie an und haben die keltische Sprache treuer bewahrt als die Iren. Dennoch hören wir nirgends, daß die Walliser ein eignes Parlament beanspruchten oder dem englischen Nachbar feindlich wären. Der Walliser, auch in den engen Thälern, wo er seine Sprache frei von englischen Einflüssen bewahrt hat, ist stolz, ein Unterthan der Königin Viktoria zu sein, und gedeiht unter dem Schutze des britischen Löwen.

Anders die Iren. In Irland ist eine notleidende, unglückliche Bevölkerung, ohne Besitz, von oben bedrückt und von seinen Führern zur Gewaltthat dem Drucke gegenüber aufgereizt.

Wales wurde im Jahre 1282 von Edward I. erobert, nachdem Irland schon im Jahre 1169 unterworfen und im nördlichen Teile besiedelt worden war. Zeit genug haben die beiden Länder gehabt, sich an das Neue zu gewöhnen. Wenn sich Irland noch heute gegen England sträubt und die Treue

*) Dieser Aufsatz wird unsern Lesern als Ergänzung zu dem im 19. Hefte: „Erinnerungen aus Irland“ willkommen sein.

Fig. 7 ‘Zoning’ — adding structural annotation

The structuring was done using the zoning tool ZOT, which was developed at the DTA. The last step for the creation of an interoperable TEI format was the merging of zoning information, OCR full text and metadata.

The TEI version, which is now available in the DTA, CLARIN-D, and DARIAH-DE, also includes a back link to the scanned image in the digital collections of SuUB. Unfortunately, this back link still is not a persistent link. At SuUB, we have persistent uniform resource name (URN) links to the digitized material but until now not at the URN granular level.

Working with Full Texts within Digitization Portals and Virtual Research Environments or Research Infrastructures

Digitization portals aim to make digitized material immediately accessible worldwide and without any time limits. Scanned images and metadata are offered. Not all material is obtainable along with OCR full text, due to different reasons such as costs or image quality. Gradually, there are more and more digitization projects producing OCR full text. For example, the DFG practical guidelines on digitization now require an OCR to be performed if the material was published after 1850. Hence there is more full text coming up to be expected from DFG-funded digitization projects.

Within digitization portals full text is used as the basis for a search index to provide fast search responses. Most portal software supports a search functionality for multiple words, and allows finding different forms of declension of nouns, but they are not meant to be a platform to work intensively with. For example, resulting big lists of hits are usually not manageable at a reasonable level. To overcome this problem and to be able to work intensively with full text material, scientists might want to be able to access and download the full text (and the metadata) primarily. On their local computational infrastructure or within research infrastructures they will have or find the tools to work intensively and properly with the material.

Within virtual research environments or research infrastructures there are facilities and many software tools to work with full texts. But there are some requirements to be met. The full text has to be uploaded. Some quality requirements have to be complied with. In order to obtain interoperable full texts and fit the requirements of the software tools (such as regular expression search, part-of-speech tagging, named-entity recognition, or Topic Modeling), format issues have to be considered. Some quantitative tools, such as mallet (Topic Modeling) only need plain text. But the pre-processing or the whole tool chain (e.g. including a graphic presentation for the analysis findings) nearly always requires the above-mentioned features: structured pages (i.e. semantically tagged full text) and metadata (year of publication, author names, etc.).

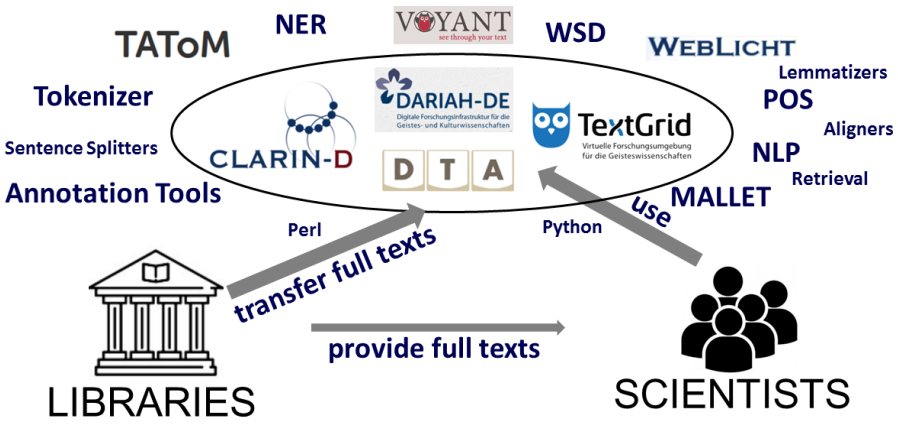


Fig. 8 Full text transfer between libraries, scientists and research infrastructures (and software tools)

Access and Transfer of Full Texts

Digitization portals usually offer access to PDFs or plain text and to technical interfaces, such as special URLs for ALTO (Analyzed Layout and Text Object) or ABBYY XML files, or to an OAI-PMH (Open Archives Initiative — Protocol for Metadata Harvesting) interface. Another possibility to access the full text is asking a contact person at the library. They also might help find and compile the metadata according to the requirements. A more reasonable and efficient approach might be that libraries directly transfer full text to research infrastructures (see Fig. 8). Having done this transfer a few times we list some criteria that might be specified together with potential requirements that should be discussed:

Quality

Transferring full texts to research infrastructures there might be some quality limits; e.g., a minimum character recognition rate, possibly defined as an interval 99± x%. Quality limits might also be given for the structuring of the text (see below under 'File formats and metadata'). These criteria vary for different centuries or decades or different software tools or scientific approaches.

File formats and metadata

Transferring plain text is not an option, nor is the output of OCR engines (such as ABBYY-XML) as well. There has to be a decision for ALTO or TEI, possibly together with 'annotation guide lines' that define subsets of the TEI standard (see DTABf above). From the point of view of digitization projects, the preliminary work (the above-mentioned zoning) for page-wise structuring and semantical tagging is quite cost intensive. Requiring information by 'annotation guidelines' directly corresponds to the amount of effort and costs during the zoning of single scanned pages.

Persistent back links

Within the full text there should be back links (page-wise or at least by sections) to the scanned images in the digital collections of the respective library or archive. If possible, these back links should be persistent at a URN granular level.[15] Researchers appreciate having the possibility to check the original image quality or to have access to supplemental material such as graphics, images, advertisements, or vignettes.

Line breaks

There should be a guideline for whether to transcribe line breaks as is. We have cooperated with partners with varying opinions on this question. The DTA wanted line breaks as is, whilst the transcription for Wikisource had to be without wrapped words.

Strictness of character transcription

Within historical full texts the spelling, of course, should be transcribed as is; for instance, 'Säugethiere' with 'th' and 'Entwickelung' instead of the modern form 'Entwicklung'. The same should apply to the transcription of single historical characters using UTF8 codes; see two examples in Fig. 9.

Documentation: 'full text metadata'

Ideally, there should be a documentation listing all the above-mentioned information: the level of the 'full text and metadata' quality, whether there are further file formats available, the availability of back links, and the status of line breaks and character transcriptions. If this 'full text metadata' would be realized with computer readable XML formats, pre-processing scripts might automatically decide what pre-processing remains to be done, and what analysis tools or scientific approaches might be applicable.

15 Dorothea Sommer, 'Persistent Identifiers: The "URN Granular" Project of the German National Library and the University and State Library Halle', *LIBER Quarterly*, 19.3–4 (2010), 259–74.

Certainly, there are more reasonable criteria and requirements to discuss. Licensing and intellectual property would be a further major issue to address. The above-mentioned issues emerged during the cooperation with various other digitizing libraries, three research infrastructures, the Wikisource group and diverse OCR post correction services and tools.

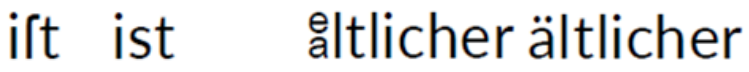


Fig. 9 Two examples of historical characters; left: the long-s; right: a historical version of an umlaut-a

We have shown some necessary steps for the transfer of high-quality full texts to research infrastructures to enable scientists to work with powerful software tools. We would like to continue optimizing these processes to accomplish them in an efficient way. Therefore, there are still a few remaining questions: Promoting full texts to research infrastructures, what quality is needed or expected? What digitization services or system interfaces should libraries offer? An 'attended direct transfer' (handled by two contact persons on either side) might be a good objective, but then a fully automatic transfer or a harvesting (download) option available via standardized interfaces should be the next desirable step. Which metadata and file formats are required, and in what order of priority? What criteria have to be fulfilled? For cost reasons it will not always be easy to fulfil all wishes, but we should streamline and prioritize these issues together. Lastly, what documentation should be submitted along with the full text? Standardizing this documentation or even creating computer readable full text metadata would have another huge potential, too.

It is appropriate, therefore, to have a closer cooperation between research infrastructures and digitizing libraries. In this context, there certainly is a potential to adapt the services of libraries and research infrastructures.

OCR full texts, even post corrected, are not perfect, but character recognition rates around 99% certainly provide more options than just being used as a search index in a web portal. Extending the focus of researchers and research infrastructures to digitized textual resources would add a vast amount of full texts. Making these fully accessible for scientific research would be a great benefit.

Conclusions for the Digital Humanities

This paper suggests the following three points to researchers or scholars working on digital serial publications. First, we would like to recommend digitized journals such as *Die Grenzboten*, and to clarify the entire origination process and its specific characteristics. As described above, it was a long running serial publication with a rich diversity of themes that was digitized in its entirety. The digitized full text was running through processes of OCR post correction, and the enhancing of the full text structure (resulting in a TEI version). Another advantage is the fact that this journal 'is digitally available in full, i.e. all of [its] issues. [It belongs to those collections that] can be considered "complete" in the sense that it contains, in fact, all items of a particular kind'. As Piotrowski states: 'in the sense of a model of some extralinguistic historical phenomenon, such a collection may then also be considered a corpus'. Also, the criterion for a corpus that differs from a mere collection of texts holds: having 'precise criteria of construction, which are motivated by its intended purpose'.[16] Those more than

16 Piotrowski, 'Historical Models and Serial Sources'.

33,000 articles of the journal have been written directly motivated by historic events making the historical context itself an important criterion in the construction process of the journal. Therefore, this journal, like all complete sources, is a valuable corpus for historical research.

The journal *Die Grenzboten* has been used for several publications,[17] within workshops and university teaching.[18] The project website at the SuUB lists five other research projects in the context of the humanities. It is clear that this usage is a result of actively providing the journal full text to the scholarly community and to research infrastructures as described in this paper. Up to now this journal, together with some other corpora (mentioned by Piotrowski[19]), are examples of a few serial sources available at this level of quality and accessibility. A lot more have been digitized or are in the process of being digitized.

Second, this library-driven contribution to this special issue also helps to remember the researcher of the original source of the material. With an OCRed full text being the 'model' of a paper original Piotrowski mentions within this volume the 'mapping property' and the 'reduction property' of models that researches should be aware of. For example, the analyses of serial sources should also consider basic textual properties. An example is the distribution of text quality over time (i.e. year of publication), which might have an impact on methods used in the digital humanities. It is therefore necessary to develop methods with a stability[20] towards varying OCR error rates. Methods and models requiring a constant error rate should be adapted or it should be possible to critically assess and interpret the results. Similarly, some of the mentioned criteria (such as 'line breaks' and 'strictness of character transcription') will also have an impact on the need for adaptation or interpretation of the used methods or an adapted pre-processing of the full texts.

What about the role and/or the conclusions for libraries? As a final point, we suggest that researchers should even more intensively see libraries (and research infrastructures) as partners to get access to historical full text materials. Everybody might agree that digitizing libraries contribute significantly to the amount of the available digitized material. Still a novel approach is to actively transfer full texts to the researchers and to research infrastructures, which is the central issue of this paper. The idea of a more extensive cooperation between libraries and research infrastructures is still a matter of the future.

And not least, libraries themselves also benefit from methods used in the digital humanities. Topic Modeling, for example, is an automated way to help with content

17 Martin Fechner and Andreas Weiß, 'Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts', *Zeitschrift für digitale Geisteswissenschaften*, no. 2 (2017); Fotis Jannidis, 'Quantitative Analyse literatischer Texte am Besipied des Topic Modeling', *Der Deutschunterricht*, 68.5 (2016), 24–35; Peer Trilcke, 'KONF: Digitale Methoden und die Literatur des späten 19. Jahrhunderts', *Potsdamer Arbeitstreffen zur digitalen Literaturwissenschaft* (2 December 2016), no page; Wikisource; and Thomas Wernecke and Maret Nieländer, 'Zur Anwendung forschungsunterstützender digitaler Methoden und Werkzeuge in Erziehungsund Geschichtswissenschaft: "DiaCollo" und die schulpolitischen Debatten in *Die Grenzboten*' (2017) [accessed 8 October 2018].

18 The journal *Die Grenzboten* as research data on GitHub; 'Expertenworkshop: Topic Modeling' at the University of Göttingen (May 2018); a workshop by Bryan Jurish, Thomas Wernecke, and Maret Nieländer on 'Diacollo and *Die Grenzboten*: Exploring Diachronic Collocations in a Historical German Newspaper Corpus' at the Genealogies of Knowledge I — Translating Political and Scientific Thought across Time and Space conference at the University of Manchester (December 2017); 'CIS OCR Workshop v1.0: OCR and Post Correction of Early Printings for Digital Humanities' at the Ludwig Maximilian University of Munich (LMU) (September 2015); and a module, also at LMU, by Florian Fink on *PoCoTo: Practice* (2015) [all accessed 8 October 2018].

19 Piotrowski, 'Historical Models and Serial Sources'.

20 The 'stability' of an algorithm or method refers to the quality of the results of a stable method that does not degrade (that much) whilst the given input has a reduced or degraded quality.

analysis or subject indexing, which is a prototypical librarian task. It is very helpful to find or preselect sections of interest within thousands of sections or millions of pages.

Manfred Nölte studied mathematics and bioinformatics. At the State and University Library Bremen, he is now engaged with the digitization of historical books and periodicals. He is interested in making high-quality textual resources fully accessible for scientific research as well as considering the role of libraries as full-text providers in the interplay of digital humanities and research infrastructures.

Martin Blenkle studied chemistry and library science. He is the Head of the Digital Services Department at the State and University Library Bremen. The IT group at Bremen has been developing the catalogue E-LIB Bremen since 1999, and is interested in user-orientated design and integration of online library services.

Bibliography

Deutsche Forschungsgemeinschaft, *DFG-Praxisregeln: Digitalisierung* (Bonn: Deutsche Forschungsgemeinschaft, 2009) [accessed 14 June 2018]

——, *DFG-Praxisregeln: Digitalisierung* (Bonn: Deutsche Forschungsgemeinschaft, 2016) [accessed 8 June 2018]

Evershed, John, and Kent Fitch, 'Correcting Noisy OCR: Context Beats Confusion', in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (New York, NY: Association for Computing Machinery, 2014), pp. 45–51 [accessed 11 June 2018]

Fechner, Martin, and Andreas Weiß, 'Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts', *Zeitschrift für digitale Geisteswissenschaften*, no. 2 (2017), no page

Fink, Florian, *Postcorrection Tool (PoCoTo) Manual* (Munich, Centrum für Informations- und Sprachverarbeitung, 2015) [accessed 14 June 2018]

Furrer, Lenz, and Martin Volk, 'Reducing OCR Errors in Gothic-Script Documents', *ERCIM News*, no. 86 (2011), 29–30 [accessed 14 June 2018]

Graham, Shawn, Scott Weingart, and Ian Milligan, 'Getting Started with Topic Modeling and MALLET', *UWSpace* (2012) [accessed 14 June 2018]

Haaf, Susanne, Alexander Geyken, and Frank Wiegand, 'The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources', *Journal of the Text Encoding Initiative*, no. 8 (2014/15), no page

Jannidis, Fotis, 'Quantitative Analyse literatischer Texte am Besipied des Topic Modeling', *Der Deutschunterricht*, 68.5 (2016), 24–35

McCallum, Andrew Kachites, *MALLET: A Machine Learning for Language Toolkit* (2002) [accessed 14 June 2018]

Nölte, Manfred, Jan-Paul Bultmann, Maik Schünemann, and Martin Blenkle, 'Automatische Qualitätsverbesserung von Fraktur-Volltexten aus der Retrodigitalisierung am Beispiel der Zeitschrift *Die Grenzboten*', *o-bib*, 3.1 (2016), 32–55 [accessed 14 June 2018]

Piotrowski, Michael, 'Historical Models and Serial Sources', *Journal of European Periodical Studies*, 4.1 (2019)

——, *Natural Language Processing for Historical Texts*, Synthesis Lectures on Human Language Technologies (Lexington, KY: Morgan & Claypool, 2012)

Sommer, Dorothea, 'Persistent Identifiers: The "URN Granular" Project of the German National Library and the University and State Library Halle', *LIBER Quarterly*, 19.3–4 (2010), 259–74

Stäcker, Thomas, 'Konversion des kulturellen Erbes für die Forschung: Volltextbeschaffung und -bereitstellung als Aufgabe der Bibliotheken', *o-bib*, 1.1 (2014), 220–37 [accessed 14 June 2018]

Trilcke, Peer, 'KONF: Digitale Methoden und die Literatur des späten 19. Jahrhunderts', *Potsdamer Arbeitstreffen zur digitalen Literaturwissenschaft* (2 December 2016), no page

Vobl, Thorsten, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter, and Klaus U. Schulz, 'PoCoTo: An Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts', in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage* (New York, NY: Association for Computing Machinery, 2014), pp. 57–61 [accessed 14 June 2018]

Wernecke, Thomas, and Maret Nieländer, 'Zur Anwendung forschungsunterstützender digitaler Methoden und Werkzeuge in Erziehungsund Geschichtswissenschaft: "DiaCollo" und die schulpolitischen Debatten in *Die Grenzboten*' (2017) [accessed 8 October 2018]