

Making great work even better

Appraisal and Digital Curation of widely dispersed Electronic Textual Resources (c. 15th–19th cent.) in CLARIN-D

Full Paper for the International Conference "Historical Corpora 2012"
December 6–9, 2012; Goethe University, Frankfurt, Germany

Table of Contents

| | |
|--|----|
| Problem Statement..... | 1 |
| Towards a solution: | |
| Integrating distributed Text Resources into a Large Text Repository..... | 2 |
| Digital Curation: Select, enhance and prevail distributed resources..... | 4 |
| Conclusion..... | 7 |
| Illustrations..... | 8 |
| References/Affiliations..... | 10 |
| Further Reading..... | 10 |

Problem Statement

Numerous high-quality primary text sources—in the context of this paper, this means full-text transcriptions (and corresponding image scans) of German works originating from the 15th to the 19th centuries—are scattered among the web or stored remotely. E.g., transcriptions of historical sources are stored locally on degrading recording media and cannot be found, let alone accessed by third parties. Additionally, idiosyncratic, project-specific markup conventions and uncommon, out-of-date or inflexible storage formats often hinder further usage and analysis of the data. Often, textual resources are accompanied by scarce, insufficient or inaccurate bibliographic information, which is only one further reason why valuable resources, even if available on the web, remain undiscovered by and are of little use to the wider research community. The integration of these dispersed primary text sources into the sustainable, web and centres-based research infrastructure of CLARIN-D will be an important step to solve this problem.

Towards a solution:

Integrating distributed Text Resources into a Large Text Repository

The work described in this paper is a contribution to a wider, joint 'curation project' of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), the University of Gießen, the Herzog August Bibliothek Wolfenbüttel (HAB), and the Institut für Deutsche Sprache Mannheim as partner institutions in CLARIN-D.¹ The aim of the curation project is to process the equivalent of ~35 000 pages printed between the 15th to the 19th centuries from large text collections, digital libraries, ongoing and terminated research projects, scholarly editions, etc. The data will be integrated into the partner's respective repositories and, from there, be made available in the CLARIN-D framework under a Creative Commons license. CLARIN-D, funded by the Federal Ministry of Education and Research (BMBF), is the German contribution to the EU-wide project CLARIN. It develops a web and centres-based research infrastructure, primarily for language-centred research in the social sciences and humanities. CLARIN-D aims at providing linguistic data, tools and services, and will offer a federated search and sophisticated retrieval facilities. Its service centres will share their data and tools in an integrated, interoperable and scalable way, and will see to their long-term archiving to ensure persistent public access.²

This paper illustrates an exemplary approach taken by the »Deutsches Textarchiv«³ (henceforth: DTA) at the BBAW to integrate high-quality textual resources and corresponding image scans from various sources into a large historical text corpus of its own and to insert these into the infrastructure of CLARIN-D. As part of DTA's contribution to the curation project, large scale collections such as Wikisource.org and Gutenberg.org as well as smaller, more specific sources will be critically reviewed to identify items appropriate to serve as valuable extensions of DTA's growing reference corpus for the historical German language. The selected text resources will be aggregated and standardized with respect to their storage and annotation format; structural and bibliographic information will be enhanced and corrected, if necessary. The integrity and significance of the collections in general and of each single item in particular will be evaluated thoroughly with respect to the curation project's criteria described below. The integration of selected items into the DTA corpora will be carried out with the help of DTA's enhancement module DTAE, which will be described in the following passage.

Exemplary workflow: The DTA and its enhancement module DTAE

The DTA started in 2007 and is building a TEI⁴-XML annotated full-text corpus of German language works. About 1,300 volumes printed between the 17th and the 19th century will be processed and published in two major phases until 2013/14. Scientific texts, as well as fiction, poetry, drama, or essays and everyday literature combine to a comprehensive collection documenting the development of the modern German Language. TEI-XML-annotated full-text transcriptions of the primary source accompanied by detailed bibliographic metadata are made available for free download and are displayed on the internet alongside digital facsimiles. The transcriptions are true to the source,

¹ CLARIN-D: Common Language Resources and Technology Infrastructure, <http://clarin-d.net> [retrieved 2012-06-15, as for all URL cited in this paper]. The curation project »Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur« will be counselled by the CLARIN-D working group »Deutsche Philologie« (German Philology) at the University of Gießen and will be coordinated by the DTA. It will be carried out by the CLARIN-D service centers at the BBAW and the IDS, and by the HAB.

² Note that CLARIN-D is only one example of a wide-span research infrastructure. By offering an OAI-PMH protocol, the resources aggregated in the course of the project described here can be made available also within other national, European or international infrastructures such as DARIAH, Europeana, TextGrid, Project Bamboo, etc.

³ Deutsches Textarchiv (DTA), www.deutschestextarchiv.de. The DTA is funded by the German Research Foundation (DFG).

⁴ Cf. TEI: Guidelines for Electronic Text Encoding and Interchange <http://www.tei-c.org/Guidelines/>.

show a high level of accuracy and are annotated with structural information following the TEI P5-compliant ›base format‹ of the DTA.⁵ The electronic full-texts are enriched with linguistic information in stand-off markup gained through tokenization, lemmatization, and part-of-speech-analysis. Each text is analyzed with CAB, a set of rewrite rules for automated normalization of historical text material.⁶ Currently (September 2012), there are 531 texts dating from 1780–1900 online, and over 200 more are prepared to be published, comprising a total of more than 270,000 digitized pages with more than 440 million characters and roughly 63 million token. An equally large amount of text from the period between 1650 and 1780 is to be processed and will likewise be made freely available for non-commercial use under a Creative Commons license by 2013/14.⁷

The prospect of more than 1,300 original works from three centuries to be published until 2014 is promising for (computer-aided) research in linguistics, semantics, typology and other areas. But still, certain discourses and genres, subject fields or domains are less well documented than others, and the number of witnesses per decade (Ø 52 texts) may—for some purposes—seem relatively small. So, to enhance this ›core collection‹, i.e. to substantially broaden the text base and to improve the balance of the corpus, the software module DTAE (“E” for Extensions) was developed.⁸ DTAE provides routines and scripts for the conversion of metadata, text and images, as well as tools for the (semi-)automatic conversion from different source formats (HTML, doc, txt, PDF, ...) into the DTA ›base format‹. In the course of the curation project described here, DTAE will be used as the platform for conversion and publication of high-quality resources from various contexts. With the help of DTAE, external resources will be integrated into DTA’s extended corpus, and at the same time into the CLARIN-D research infrastructure, where tools for further analysis of the data are provided and their long-term preservation will be taken care of.⁹

DTAE was developed to facilitate the production of new high-quality transcriptions of primary sources in cooperation between the DTA and external researchers as well as for the enhancement and integration of existing resources. While the former, i.e. the co-operative text production is carried out successfully at the DTA—for example, with the Alexander-von-Humboldt-Forschungsstelle and the Marx-Engels-Gesamtausgabe (MEGA) project at the BBAW, as well as the Forschungsstelle für Personalschriften at the Philipps-Universität Marburg (Arbeitsstelle der Akademie der Wissenschaften und der Literatur, Mainz)—, the focus here will be on the latter, i.e. on the aspect of enhancement and integration of existing resources. Large amounts of text data are currently integrated into DTAE from born-digital scholarly editions like those of works of J. v. Sandrart, J. F. Blumenbach, and from the centenary-spanning 'Polytechnisches Journal' founded by J. G. Dingler.¹⁰ The integration of these text resources is relatively straightforward, thanks to the TEI-compliant encoding provided by the projects mentioned. Therefore, instead of going into detail any further on this aspect, the remainder of this paper will deal with the much higher obstacles on the way to identify, enhance, refine and integrate text converted from various storage formats.

⁵ Deutsches Textarchiv – Basisformat, <http://www.deutschestextarchiv.de/doku/basisformat>. The DTA ›base format‹ is a subset of TEI P5, containing about 100 elements and their possible attributes and values. It restricts the number of elements from the TEI Guidelines in order to reduce the application of inconsistent tagging for similar structural phenomena within the corpus. By this, it aims at gaining coherence at the annotation level, given the heterogeneity of the DTA texts regarding time of origin (1650–1900) and text type (e.g. fiction, functional texts, or scientific texts). Cf. Haaf/Wiegand/Geyken 2012b.

⁶ Cf. Jurish 2010 and Jurish 2011. CAB provides an automated normalization of the historical orthography in order to allow for lemma based, spelling-tolerant corpus searches.

⁷ All DTA texts are available for download in different formats: in TEI-XML, rendered HTML, and as plain text transcription. CMDI metadata comprising TEI header information may be harvested via OAI-PMH.

⁸ Deutsches Textarchiv – DTAE, <http://www.deutschestextarchiv.de/dtae>.

⁹ An infrastructure similar to DTAE is currently being developed at the HAB in the context of the project AEDit (<http://www.hab.de/forschung/projekte/aedit-e.htm>), while the project partners at the IDS will carry out respective steps to integrate resources on their behalf.

¹⁰ For further information on these editions, cf. the project’s respective web sites: www.sandrart.net, www.blumenbach-online.de, and www.polytechnischesjournal.de.

Digital Curation: Select, enhance and prevail distributed resources

Criteria: What to look for?

Digital curation—or, encompassing a wider scope of activities, digital stewardship—for the purpose of this paper entails the careful selection, refinement and analysis, archiving and maintenance of digital assets.¹¹ First of all, appropriate items for the curation project have to be identified with the help of a set of criteria. To accomplish the first step, i.e. to identify suitable items, a number of criteria will have to be considered. The curation project generally is putting an emphasis on quality over quantity:¹² It is rather 'hand-picking' than following a 'down-them-all' approach, where, as a prize for the greater amount of data to be gained in a single sweep, one has to put up with the downside, i.e. the minor quality a considerable number of single items in the collection—and, as a result, of the corpus as a whole—will display.

The criteria described in the following were defined in accordance with the general guidelines of the DTA.¹³ First of all, the digitized print sources should be first or early editions of the text represented. As a project with a strong orientation in Historical Text/Corpus Linguistics and Lexicography, the DTA offers text true to the primary source, without later 'normalizations' in spelling and other severe intrusions distorting the historical text. Furthermore, the works in question should be expressive witnesses of the development of the New High German Language, and/or relevant to a certain field of scientific or cultural history, and/or instances of a certain special discourse, documenting specific aspects of different kinds of language use, including everyday language. The transcribed text should contain or be accompanied by information about the method of data acquisition (uncorrected, 'dirty' OCR or OCR with proofing, Double Keying, ...), its creator and editing status (complete, draft, working transcription, ...). The image scans should show a high resolution, preferably be full-colour copies with ≥ 300 dpi in TIFF format. The metadata describing the source should be accurate and as detailed as possible, or has to be completed in the curation process. Certainly, legal aspects concerning text, metadata and images will have to be kept in mind: Every item, i.e. images, metadata and text should be available under a free license at least for reuse in a scientific context.

The text should be transcribed true to the source. Any alterations, e.g. the replacement of certain letters or ligatures by other characters, the correction of printing errors, etc., should be documented and be done consistently. Line breaks, or at least page breaks found in the source document should be marked in the transcription.¹⁴ The transcription should prove high accuracy on the level of

¹¹ According to the Digital Curation Centre (DCC) (2007), "Digital curation is maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management of data throughout the research lifecycle [...], including the provision of access to data and data reuse. Meeting this obligation will be enabled by good data stewardship." While 'digital curation' puts the emphasis on the cycle of creation, selection and preservation, 'digital stewardship' is used in a somewhat broader sense. It emphasises the activities of curation as crucial, but equally stresses the responsibility for ongoing, active work on preserved objects in the asset. Quite often however, and also in DCC's definition quoted above, the terms were (and still are) used interchangeably or in the sense that one concept entails the other, cf. for example Rusbridge et al. 2005: 2 or Lee/Tibbo 2007. For the purpose of this paper the definition given above will suffice. For an overview of recent publications on this topic cf. Bailey, Jr. 2012.

¹² Nevertheless, selected 'working transcriptions' can also be integrated to be revised step by step to finally meet the curation project's criteria. Especially in this respect, recommendations of CLARIN-D's discipline-specific working groups (Fachspezifische Arbeitsgruppen) will be taken into consideration, and members of the community will be encouraged to help improve the resources e.g. by proofreading and correcting.

¹³ Cf. DTA-Leitlinien, www.deutschestextarchiv.de/doku/leitlinien.

¹⁴ This is essential for the (automated) alignment of source images and transcribed text. It also allows for a rough, general comparison between source and derived text in order to evaluate the quality of the resource. In this sense and beyond that, it facilitates anticipative as well as retrospective quality assurance, e.g. proofreading. For a documentation of DTA's profound experience with quality assurance in large text corpora cf. Geyken et al. 2012 and Haaf et al. 2012a.

characters (preferably 99.5+ %) and, with respect to the annotation, should contain at least the most basic structural information (i.e. chapters, headers, paragraphs).

Although, at first glance, these criteria might seem to form quite a low threshold, they will help to guarantee a high quality and integrity of the acquired data, as they allow for a good orientation to separate the wheat from the chaff.¹⁵ For example, most of the works represented in the text collection of Gutenberg-DE¹⁶—and, although with some notable exceptions, also that of zeno.org¹⁷—do not meet the curation project's criteria in every respect. A great number of the transcriptions are based on philologically questionable editions, bristling with undocumented and, often enough, inconsistent alterations of the original text. In some cases, forewords, dedications, and other 'supplementary' parts printed in the primary source remained unconsidered altogether, and some transcriptions do not show the accuracy required. In this respect, they do not meet the curation project's criteria applied for the selection of resources.

Sources: Where to look?

Large Text (and Image) Collections: Wikisource, Gutenberg and the like

Most promising for a considerable amount of appropriate documents fulfilling the criteria described above are large collections such as the German partition of Wikisource and German-language texts from the American Project Gutenberg (PG). The quality of the resources assembled in those 'opportunistic' collections with its many individual contributors differs strongly, but there nonetheless are some high quality representations of historic documents to be discovered.¹⁸ They offer accurate transcriptions of historic primary sources, often along with corresponding image scans in good quality. Unfortunately, these fine examples are somewhat hidden among the vast total amount of objects there. To make sure its integration is worth an effort, each possible candidate has to be evaluated following the criteria described in the previous section—a non-trivial task itself, given for example the amount of >24 500 German-language texts in the German Wikisource.¹⁹

The metadata describing the collected objects displayed on the website often is insufficient, sometimes inadequate. The navigational structure of the respective sites is rather opaque, and the on-site retrieval facilities are often quite basic. The options to browse and search the collection are quite limited and it is hard to get an overview.²⁰ This holds for Wikisource and PG, but is also true for

¹⁵ As a welcome side effect, the criteria help to narrow the focus of the project described here to a manageable amount of text resources.

¹⁶ Projekt Gutenberg-DE, <http://gutenberg.spiegel.de/>.

¹⁷ Zeno.org, <http://www.zeno.org/>. In 2009, the whole collection was acquired by the research infrastructure project TextGrid, funded by means of the BMBF. The text files from zeno.org were converted into the Text Grid 'Baseline Encoding', a TEI-conformant basic encoding format used mainly to allow for project-specific as well as cross-text queries within the Textgrid Repository (Cf. Textgrid (2007–2009): p. 6.). In this process, basic structural information was gained by automated analysis of the source markup. XML-ids were added to each line of the transcription to allow for more exact referencing. Since July 2011, the data stock of the literature folder is available for download. The original transcriptions of historic works for zeno.org were almost exclusively derived from partly modernized editions from the 19th/20th century. During the transformation to TextGrid, they were not proofed against reliable scholarly editions or compared to the primary sources. Likewise, proofing and correction of the metadata is yet to be done (cf. <http://www.textgrid.de/en/digitale-bibliothek.html>).

¹⁸ Of course, but with the reservations mentioned above in mind, selected, high-quality items from zeno.org and Gutenberg-DE meeting the project's criteria will also be integrated.

¹⁹ Cf. <http://de.wikisource.org>, Hauptseite > Wikisource Aktuell > Statistik.

²⁰ Wikisource, for example, offers no query or download API for ingesting the full descriptive metadata of project's resources, although its development obviously has been discussed for some time, cf. http://de.wikisource.org/wiki/Wikisource:Metadaten#Weitergabe_der_Metadaten and http://de.wikisource.org/wiki/Wikisource:Skriptorium/Archiv/2006/3#Professionalisierung_von_Wikisource.

other, comparably smaller collections under consideration. Therefore, the sites in focus will have to be critically scoured manually pursuing different strategies.

Research Projects and Scholarly Editions

As a second domain for historical text resources, research projects and scholarly editions will have to be taken into account, as their data in general incorporate the expertise and scrutiny of acknowledged specialists. Without doubt, these would be of high interest for the purpose of this curation project, but first of all, the data has to be retrieved and often enough legal issues have to be solved: The access, even to the 'raw' data of the project might be problematic, e.g. because of restrictive contracts with publishing institutions.²¹ Both tasks, retrieving the data and securing access to it, are even harder to accomplish once the research is done: Staff members will be off to other places, while the work done—especially the fundamental steps *before* publication of the research outcomes—too often is not documented well enough and not stored on every level. Furthermore, the project-specific transcription and markup conventions applied might have become incomprehensible to others. They would have to be laboriously reconstructed in order to be able to evaluate the resource in the first place.

If the data are available at all, a further and no less severe problem to be expected concerns its storage format. Until recently, the majority of scholarly editions of historical text material were produced to the ends of a printed (or print-like) documentation of the work.²² Therefore, the text base was produced with the help of GUI based word processors and other office tools. It was published and/or stored in formats such as Adobe InDesign, LaTeX or PDF. The most severe problems are the evolving obsolescence of certain (esp. proprietary) data formats (older versions of MS Word, WordStar, WordPerfect etc.), and the fact that GUI based word processors and the output formats tend to indistinguishably mix layout information with structural information. The data needs reformatting and refreshing to different extends before the old format becomes obsolete and before the intellectual work explicit and implicit in the documents becomes incomprehensible.

Special Collections and Singular Resources

Finally, and in addition to large collection and research projects, smaller compilations of texts on a certain topic, representing a particular discourse or epoch will be considered. Often built and run by enthusiastic private scholars or layman investing a lot of energy and their spare time, these thematic collections sometimes reveal astonishing discoveries. Singular findings will be integrated, hopefully with the approval and also with the support of its producers.

For each single integrated item the appreciation of the work of others will be made visible in the source documentation. In order to establish a culture of shared access and usage, the importance of a reputation system must and will not be omitted. To overcome problems of format obsolescence and inflexibility, conversion of the data into a consistent, standardised and flexible format such as (TEI-)XML is necessary. However, the amount of time and manual work this process requires differs strongly depending on the data base. This in mind, it will be decisive for the success of the curation project to keep a sound balance between the effort it takes to integrate the 'chosen ones' and the (anticipated) value they represent to the research community addressed in CLARIN-D, and to carefully weigh the quality against the quantity of the aggregated resources.

²¹ 'Raw data' in the context of this paper could mean an uncommented, but exact transcription of the primary source, which forms the basis of almost every scholarly edition of a text. These transcriptions would be of great value to other projects (not only the curation project described in this paper, but also for corpus projects like the DTA in general), which seldom seems to be considered while negotiating the terms of publication. Often, this 'raw data' is less taken care of in the process of critical editing and commenting, and therefore it even more likely becomes outdated and inaccessible by (storage) format evolution over time.

²² Of course, this is still a wide-spread conduct, while it would be of great benefit for the research community to produce and preserve data in exchangeable, well documented formats like (TEI-)XML from the beginning.

Integration: How to proceed?

Once a relevant resource meeting the named criteria is identified, the full-text transcription, image scans and metadata are acquired and being integrated into DTA's enhancement module DTAE. To this purpose, the electronic documents are enriched with detailed bibliographic and structural information. In the next step, the bibliographic data and full-text transcription are converted into the DTA 'base format'. The acquired text and metadata are published alongside the corresponding image scans via the DTAE framework. Each text is analyzed with CAB for automated normalization of the historical text: the great variance in spelling of terms is being mapped onto its modern form, thereby allowing for spelling-tolerant and arbitrarily complex queries in the growing text corpus. The linguistic analysis encompasses tokenization, lemmatization, and PoS-tagging in stand-off markup. The text can be displayed page-wise in an HTML version automatically rendered from the underlying TEI-XML (*Illustration 1*), it can be searched and explored as a single resource, in the context of the different sub-corpora compiled at the DTA, or in the context of the DTA 'core corpus'. The resource descriptions and bibliographic information are standardized conformant to authority formats (e.g. CMDI²³ or DC²⁴) in order to be shared via OAI-PMH²⁵ and to be integrated into CLARIN-D's service architecture.

In parallel, all new items added to the DTA corpus can be accessed via DTA's quality assurance platform DTAQ.²⁶ In DTAQ, texts may be proofread page by page in comparison to their source images (*Illustration 2*). This way errors can be detected which may have occurred during the former transcription and annotation process, or that were overlooked or not taken care of during integration. While the transcription can best be inspected in the rendered HTML version, the underlying annotation can conveniently be checked in TEI-XML. The automated analysis of the full-text with CAB can be checked as well.

Conclusion

The CLARIN-D curation project described here will integrate the equivalent of some 35 000 pages into a large corpus for the written German language between the 15th and the 19th centuries. From the large text repository under construction, balanced reference corpora can be derived. This will help to improve the situation for corpus-based research, particularly in Historical Linguistics, but also in the Humanities in general. By curating existing textual primary resources, as yet dispersed items will be incorporated into the dynamic and growing corpora compiled by the DTA (mainly 17th–19th centuries), at the HAB (15th–18th centuries) and at the IDS in Mannheim. By applying a consistent, interoperable encoding based on the recommendations of the TEI and by integrating the resources into the CLARIN-D infrastructure, the data will be explorable in a broader context. Access to and sustainability of these resources will be improved substantially. Via CLARIN-D, users will have central access to a formerly dispersed, large variety of corpus text that can be processed by the elaborated tool chain CLARIN-D offers. By establishing methods of interoperation, a system of quality assurance and credit, and a set of technical practices that allow integrating resources of different origin, CLARIN-D will contribute significantly to the scholarly community. A culture of sharing corpus resources in a collaborative way will be encouraged.

²³ CMDI: Component Metadata Infrastructure, <http://www.clarin.eu/cmdi>.

²⁴ Cf. DCMI: The Dublin Core Metadata Initiative, <http://dublincore.org/>.

²⁵ Open Archives Initiative Protocol for Metadata Harvesting, www.openarchives.org/pmh.

²⁶ Deutsches Textarchiv – Qualitätssicherung (DTAQ), www.deutschestextarchiv.de/dtaq. [Users must register and have their accounts activated by a DTA staff member.]

Illustrations

Illustration 1: Wikisource-item in DTAE: image, transcription in rendered HTML, metadata and further information, i.e. on the transcription and annotation guidelines applied in the production of the resource.

Grimmelshausen, Hans Jakob Christoffel von: Deß Weltberuffenen SIMPLICISSIMI Pralerey und Gepräng mit seinem Teutschen Michel. [Nürnberg], 1673, image 11.

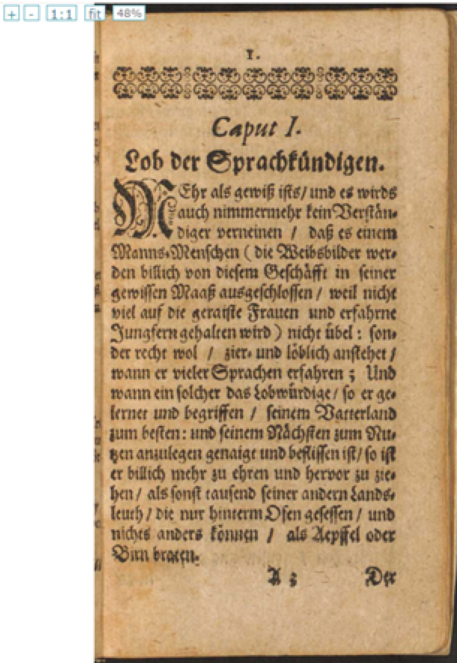
DTAE zuletzt gelesen · Hilfe · Zufallsseite ChristianThomas | Admin | Profil | ausloggen

grimmelshausen_michel_1673 (Wikisource) offene Tickets: 0 (0 ganzes Buch)
Stand: Mon May 14 10:36:36 2012

Text Text/Bild Darstellung TEI-XML 🌟

0 - 1 - 133 0 - 1 - 133 0 - 0 - 134 0 - 0 - 134

Bild: 0011 << vorherige Seite



vorherige Seite

nächste Seite >>

Caput I.
Lob der Sprachkündigen.

Mehr als gewiß ists / und es wirds auch nimmermehr kein Verständiger verneinen / daß es einem Manns-Menschen (die Weibsbilder werden billich von diesem Geschäft in seiner gewissen Maaß ausgeschlossen / weil nicht viel auf die geraifte Frauen und erfahrene Jungfern gehalten wird) nicht übel: sonder recht wol / zier- und löblich anstehet / wann er vieler Sprachen erfahren; Und wann ein solcher das Lobwürdige / so er gelernet und begriffen / seinem Vatterland zum besten: und seinem Nächsten zum Nutzen anzulegen genaigt und beflissen ist / so ist er billich mehr zu ehren und hervor zu ziehen / als sonst tausend seiner andern Landsleuth / die nur hinterm Ofen gesessen / und nichts anders können / als Aepffel oder Birn braten.

Metadaten

Titel: Deß Weltberuffenen SIMPLICISSIMI Pralerey und Gepräng mit seinem Teutschen Michel

Untertitel: Jedermänniglichen / wanns seyn kan / ohne Lachen zu lesen erlaubt

Aut.-Daten: **Aut. 1, Vorn.:** Hans Jakob Christoffel von
Aut. 1, Nachn.: Grimmelshausen
Aut. 1, PND: 118542273

Ersch.-Jahr: 1673
Ort: Nürnberg

- Anmerkung zu Metadaten verfassen
- Metadaten ändern

Informationen

Quelle: Wikisource

Umfang: **134** Scans
ca. 120711 Zeichen
ca. 17825 Token / [[:alnum:]]/
ca. 5150 Oberflächentypes

Schriftart: Fraktur

Genre:

Verfügbarkeit: zugänglich

im DTAE seit: 2012-05-14 10:34:21

Weitere Informationen:

Die Textgrundlage dieses Werkes stammt von Wikisource.
Quelle der Scans: MDZ München.
Anmerkungen zur Transkription:

- als Grundlage dienen die Editionsrichtlinien von Wikisource.
- Überschriebene „e“ über den Vokalen „a“, „o“ und „u“ werden als moderne Umlaute transkribiert.
- Der Seitenwechsel erfolgt bei Worttrennung nach dem gesamten Wort.
- Abkürzungen werden aufgelöst.
- æ und œ werden durch ae bzw. oe, ē als ae wiedergegeben.

[Anmerkung zu Informationen verfassen](#)

Illustration 2:
Quality Assurance
in DTAQ: image,
XML-view; 'ticket'
system to report
findings, e.g.
printing errors,
transcription
errors and
inconsistencies in
annotation

*Kerner, Justinus:
Geschichten
Besessener neuerer
Zeit. Karlsruhe,
1834, image 127.*

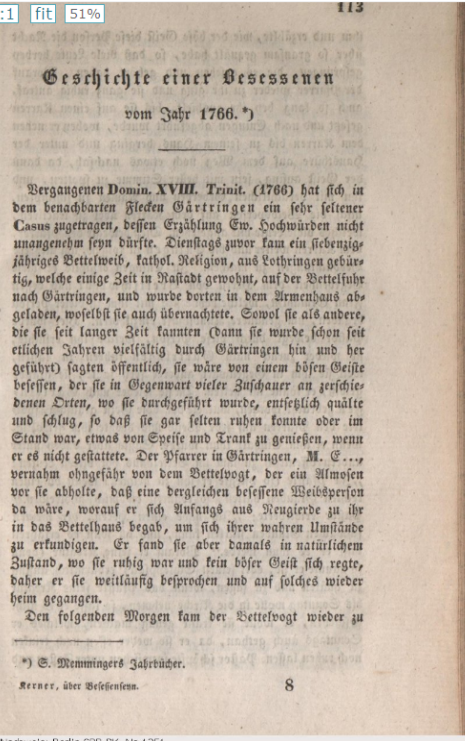
DTAQ
zuletzt gelesen · Hilfe · Zufallsseite ChristianThomas | [Admin](#) | [Profil](#) | [ausloggen](#)

kerner_besessene_1834 (CN)

offene Tickets: 2 (0 [ganzes Buch](#))
Stand: Mon Jun 11 12:29:43 2012

| Text | Text/Bild |
|--------------|--------------|
| 0 - 45 - 162 | 0 - 12 - 195 |

Bild: 0127 << vorherige Seite

+ - 1:1 fit 51%


Nachweis: Berlin S80-PK, Na 1351

nächste Seite >>

```

<text>
<body>
  <pb facs="#f0127" n="113"/>
  <div n="1">
    <head><hi rendition="#b"><hi rendition="#g">Geschichte
    einer Besessenen</hi></hi><lb/>
    vom Jahr 1766. <note place="foot" n="*">">S. Memmingers Jahrbücher.
    </note></head><lb/>

    <milestone rendition="#hr" unit="section"/>
    <p>Vergangenen <hi rendition="#aq">Domin. XVIII.
    Trinit.</hi> (1766) hat sich in<lb/>
    dem benachbarten Flecken <hi rendition="#g">Gärtringen</hi> ein
    fehr feltener<lb/><hi rendition="#aq">Casus</hi> zugetragen,
    deffen Erzählung Ew. Hochwürden nicht<lb/>
    unangenehm feyn dürfte. Dienftags zuvor kam ein fiebenzig-<lb/>
    jähriges Bettelweib, kathol. Religion, aus Lothringen gebür-<lb/>
    tig, welche einige Zeit in Raftadt gewohnt, auf der Bettelfuhr<lb/>
    nach Gärtringen, und wurde dorten in dem Armenhaus ab-<lb/>
    geladen, wofelbft fie auch übernachtete. Sowol fie als andere,<lb/>
    die fie feit langer Zeit kannten (dann fie wurde schon feit<lb/>
    etlichen Jahren vielfältig durch Gärtringen hin und her<lb/>
    geführt) sagten öffentlich, fie wäre von einem böfen Geifte<lb/>
    beffenen, der fie in Gegenwart vieler Zufchauer an zerfchie-<lb/>
    denen Orten, wo fie durchgeführt wurde, entfetzlich quälte<lb/>
    und fchlug, fo daß fie gar felten ruhen konnte oder im<lb/>
    Stand war, etwas von Speife und Trank zu genießen, wenn<lb/>
    er es nicht gefattete. Der Pfarrer in Gärtringen, <hi
    rendition="#aq">M.</hi> E ...,<lb/>
    vernahm ohngefähr von dem Bettelvoigt, der ein Almofen<lb/>
    vor fie abholte, daß eine dergleichen befeffene Weibsperson<lb/>
    da wäre, worauf er fich Anfangs aus Neugierde zu ihr<lb/>
    in das Bettelhaus begab, um fich ihrer wahren Umftände<lb/>
    zu erkundigen. Er fand fie aber damals in natürlichem<lb/>
    Zustand, wo fie ruhig war und kein böfer Geift fich regte,<lb/>
    daher er fie weitläufig befprochen und auf folches wieder<lb/>
    heim gegangen.</p><lb/>

    <p>Den folgenden Morgen kam der Bettelvoigt wieder zu<lb/>
    <fw place="bottom" type="sig"><hi rendition="#g">Kerner</hi>, über
    Befeffenfeyn. 8</fw><lb/></p>
</div>
</body>
</text>
</TEI>

```

Buchdaten

DTA-Informationen
Metadaten
Ansichten (Webversion)
nächstes Ticket

Korrekturstatus

✔ Text von mir kontrolliert

✔ Text/Bild von mir kontrolliert

✔ Darstellung von mir kontrolliert

✔ TEI-XML von mir kontrolliert

Tickets für diese Seite

neu: Ticket

#32247 [2012-07-09T17:12, ChristianThomas
Druckf.: Schreibfehler zweifelhaft
an zerfchie- denen Orten
an verfhie- denen Orten

Suche

DDC

grep

Anmerkung zu dieser Seite anlegen

References/Affiliations

CLARIN-D curation project »Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur«, cf. <http://www.clarin-d.de/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1.html>

The curation project is carried out by the following institutions:

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), www.bbaw.de
Deutsches Textarchiv (DTA), www.deutschestextarchiv.de

Justus-Liebig-Universität Gießen, www.uni-giessen.de
Institut für Germanistik, Prof. Dr. Thomas Gloning, www.uni-giessen.de/gloning/

Herzog August Bibliothek Wolfenbüttel (HAB), www.hab.de
Projekt AEDit, <http://www.hab.de/forschung/projekte/aedit.htm>

Institut für Deutsche Sprache (IDS), www.ids-mannheim.de
Programmbereich Forschungsinfrastrukturen, <http://www.ids-mannheim.de/fi/>
Programmbereich Korpuslinguistik, Korpusausbau, <http://www.ids-mannheim.de/kl/projekte/korpora/>

Initial samples of curated resources are available in DTAE, www.deutschestextarchiv.de/dtae.

In the broader context of a CLARIN-D panel on corpus development, the major outlines of the curation project were presented at the Digital Humanities 2012 at the University of Hamburg, 19 July 2012. Cf. Geyken, Gloning, Stäcker 2012.

Further Reading

Bauman, Syd (2011): "Interchange vs. Interoperability." Presented at Balisage: The Markup Conference 2011, Montréal, Canada, August 2–5, 2011. In: *Proceedings of Balisage: The Markup Conference 2011*. Balisage Series on Markup Technologies, vol. 7, doi:10.4242/BalisageVol7.Bauman01.

Bailey, Jr., Charles W. (2012): *Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works*. Available as open access PDF file via <http://digital-scholarship.org/dcpb/dcb.htm>.

Digital Curation Centre (DCC) (2007): "What is digital curation?", <http://www.dcc.ac.uk/digital-curation/what-digital-curation>; "DCC Curation Lifecycle Model", <http://www.dcc.ac.uk/resources/curation-lifecycle-model>.

Digital Preservation Coalition (DPC) (2008): *Preservation Management of Digital Materials: The Handbook*, online version <http://www.dpconline.org/advice/preservationhandbook>.

Geyken, Alexander et al. (2012): „TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv.“ In: *Jahrbuch für Computerphilologie*, online version <http://www.computerphilologie.de/jg09/geykenetal.html>.

Geyken, Alexander, Thomas Gloning and Thomas Stäcker (2012): "Compiling large historical reference corpora of German: Quality Assurance, Interoperability and Collaboration in the Process of

Publication of Digitized Historical Prints", <http://www.dh2012.uni-hamburg.de/conference/programme/abstracts/compiling-large-historical-reference-corpora-of-german-quality-assurance-interopability-and-collaboration-in-the-process-of-publication-of-digitized-historical-prints/>. Cf. the a/v documentation provided by 'lecture2go' of Hamburg University, <http://lecture2go.uni-hamburg.de/konferenzen/-/k/13952>.

Haaf, Susanne, Frank Wiegand and Alexander Geyken (2012a): "Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text. Accepted for publication in: *Journal of the Text Encoding Initiative* (jTEI) 4, 2012.

Haaf, Susanne, Frank Wiegand and Alexander Geyken (2012b): The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora. In: 11th Conference on Natural Language Processing (KONVENS) – Empirical Methods in Natural Language Processing, Proceedings of the Conference. Edited by Jeremy Jancsary. Wien, 2012 (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5), online version (Sept. 14th 2012) <http://www.oegai.at/konvens2012/proceedings.pdf#page=383>

Jurish, Bryan (2011): *Finite-state Canonicalization Techniques for Historical German*. PhD thesis, Universität Potsdam, January, 2011, URN urn:nbn:de:kobv:517-opus-55789, URL <http://opus.kobv.de/ubp/volltexte/2012/5578/>.

Jurish, Bryan (2010): "More than Words: Using Token Context to Improve Canonicalization of Historical German." In: *Journal for Language Technology and Computational Linguistics* (JLCL), vol. 25/1, 2010: 23–39, online version: http://media.dwds.de/jlcl/2010_Heft1/bryan_jurish.pdf.

Lee, Christopher A. and Helen R. Tibbo (2007): "Digital Curation and Trusted Repositories: Steps toward Success." In: *Journal of Digital Information* (JoDI), Vol. 8, No 2: Digital Curation & Trusted Repositories (2007), <http://journals.tdl.org/jodi/article/view/229/183>.

Rusbridge C. et al. (2005). "The Digital Curation Centre: A Vision for Digital Curation." In: Proceedings from the IEEE Conference *Local to Global: Data Interoperability – Challenges and Technologies*. Forte Village Resort, Sardinia, Italy, 2005: 1–11. Available from: <http://eprints.erpanet.org/82/>.

TextGrid (2007–2009): "TextGrid's Baseline Encoding for Text Data in TEI P5", <http://www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf>.

Unsworth, John (2011): "Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI". In: *Journal of the Text Encoding Initiative* (jTEI) 1, <http://jtei.revues.org/215>.

Whyte, Angus, and Andrew Wilson (2010): "How to Appraise & Select Research Data for Curation. A Digital Curation Centre and Australian National Data Service 'working level' guide." Digital Curation Centre, PDF download: http://www.dcc.ac.uk/webfm_send/828, online version: <http://www.dcc.ac.uk/resources/how-guides/appraise-select-data>.