IDS

Korpuslinguistik und interdisziplinäre
Perspektiven auf Sprache

Band 5

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr
VERLAG

Jost Gippert / Ralf Gehrke (eds.)

Historical Corpora

Challenges and Perspectives

narr VERLAG

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

Dischingerweg 5 · D-72070 Tübingen

Gedruckt auf chlorfrei gebleichtem und säurefreiem Werkdruckpapier.

Internet: www.narr.de
E-Mail: info@narr.de

Redaktion: Melanie Steinle, Mannheim
Layout: Andy Scholz, Essen (www.andyscholz.com)
Printed in Germany

ISSN 2191-9577
ISBN 978-3-8233-6922-6

Contents

CHRISTIAN THOMAS / FRANK WIEGAND

Making great work even better[1]

Appraisal and digital curation of widely dispersed electronic textual resources (c. 15th-19th centuries) in CLARIN-D

Abstract

Numerous high-quality primary textual resources – in the context of this paper, this means full-text transcriptions (and corresponding image scans) of German texts originating from the 15th to the 19th century – are scattered among the web or stored remotely on institutional or private servers. They are often filed on degrading recording media and are encoded in out-of-date or inflexible storage formats. 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. Additionally, idiosyncratic, project-specific markup conventions often hinder further usage and analysis of the data. Because of these and other problems, a great amount of the abovementioned transcriptions of historical sources can hardly be found, let alone accessed by third parties, and are of little use to the wider research community. This situation is unsatisfying from the perspective of a (corpus-)linguistic project like the one described here, but also from the perspective of any text-based research in the humanities and social sciences. The integration of as many of these 'dispersed' high-quality primary textual resources as possible into an encompassing repository like the sustainable, web and centres-based research infrastructure of CLARIN-D[2] is an important step and at least a necessary prerequisite to solve this problem. This paper summarizes the work of an 18-month project funded by the German Federal Ministry of Education and Research (BMBF) which dealt with the curation and integration of historical text resources of the 15th-19th century into the CLARIN-D infrastructure.

[1] This paper is a thoroughly revised version of the original full paper by the same title, handed in for the International Conference "Historical Corpora 2012", December 6-9, 2012; Goethe University, Frankfurt am Main, Germany, and published in October 2012 on the edoc-server of the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), URN: urn:nbn:de:kobv:b4-opus-23081, URL: http://edoc.bbaw.de/volltexte/2012/2308/ [last retrieved April 30, 2014, as for all URLs cited in this paper].

[2] CLARIN-D: Common Language Resources and Technology Infrastructure, http://clarin-d.de/. Funded by the Federal Ministry of Education and Research (BMBF), CLARIN-D 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 offers a federated content search and sophisticated retrieval facilities. Its service centres share their data and tools in an integrated, interoperable and scalable way, and will see to their long-term availability and archiving to ensure persistent public access.

1. The Mission: curating and integrating distributed text resources into a large text repository

The work described in this paper was carried out in the context of a joint 'curation project' (duration: September 2012 until February 2014) of the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), the Justus-Liebig-Universität (JLU) Gießen, the Herzog August Bibliothek (HAB) Wolfenbüttel, and the Institut für Deutsche Sprache (IDS) in Mannheim as partner institutions in CLARIN-D.[3] Digital curation in the context of the project described here entails the careful selection, refinement and analysis, archiving and ongoing maintenance of digital assets.[4] The stated objective of this project was to process the equivalent of approx. 35,000 pages printed between the 15th and the 19th century from large text collections, digital libraries, ongoing and terminated research projects, scholarly editions, etc. When the project terminated in February 2014, more than 79,000 pages (encompassing approx. 21 million tokens) were integrated, thereby even doubling the targeted number of pages. The three major reasons for this over-achievement are worth mentioning:

[3] Cf. the project's web page "Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts in einer nachhaltigen CLARIN-D-Infrastruktur", Kurationsprojekt 1 der Facharbeitsgruppe 1 Deutsche Philologie, www.deutschestextarchiv.de/clarin_kupro. The project was counseled by the discipline-specific working group German Philology in CLARIN-D and coordinated at the BBAW. It was carried out by the CLARIN-D service centres at the BBAW and the IDS, and by the HAB. The basic ideas behind a cooperation like this and the more general aims and methods of corpus compilation are described in this volume in the contribution of Alexander Geyken and Thomas Gloning: "A living text archive of 15th-19th-century German. Corpus strategies, technology, organization."

[4] 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 are used interchangeably or in the sense that one concept entails the other, cf. for example Whyte/Wilson (2010), Lee/Tibbo (2007) or Rusbridge et al. (2005: 2). 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).

1) The project could rely on the elaborated corpus building infrastructure and the well-documented workflow set up at the cooperating project Deutsches Textarchiv (DTA)[5] at the BBAW.

2) A great amount of text (the equivalent of more than 36,000 pages) was integrated from projects associated with the HAB. Since they were already TEI[6]-encoded, these documents could easily be converted into the specific TEI-format of the DTA; after proofing some representative sample documents, the process of integration could be entirely automated for the rest of the HAB-corpus. For another large amount of text (approx. 20,000 pages) integrated from Wikisource, the manual effort could be reduced significantly with the help of a specialised web form.[7] This script-based integration form parses the cumbersome 'MediaWiki'-syntax and transforms as many elements as possible into TEI-XML.

3) It has to be kept in mind that the curation of digital assets is an ongoing process that does not end with the integration. For some of the HAB texts, but also for texts from the Max Planck Institute for the History of Science (MPI-WG) and other collections, further work has to be done to improve the quality of text and metadata. In the course of the project, we decided to first of all convert and integrate these texts into the corpus infrastructure at the DTA, in order to then be able to use the quality assurance mechanisms provided by the DTA and thereby support the ongoing process of data curation.

The integrity and significance of the collections in general and of each single item in particular was evaluated thoroughly with respect to the project's qualitative criteria described below. The selected items were integrated into the partner's respective repositories and, from there, made available in the CLARIN-D framework under a Creative Commons license.[8] In preparation for

[5] Deutsches Textarchiv (DTA), www.deutschestextarchiv.de. The DTA is funded by the German Research Foundation (DFG). All DTA texts are available for download in different formats: in TEI-XML, HTML, in the Text Corpus Format (TCF) used by WebLicht services in CLARIN-D, and as plain text transcriptions. Metadata, available as TEI-Headers, formatted in Dublin Core (DC, cf. http://dublincore.org/) and according CLARIN's Component MetaData Infrastructure (CMDI, cf. www.clarin.eu/cmdi), can be harvested via an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), available under www.deutschestextarchiv.de/api.

[6] TEI: Text Encoding Initiative, www.tei-c.org. Cf. Guidelines for Electronic Text Encoding and Interchange www.tei-c.org/Guidelines/.

[7] URL: www.deutschestextarchiv.de/dtae/submit/wikisource.

[8] Note that CLARIN-D is only one example of a wide-span research infrastructure. By offering the

and all through the course of the project, large scale collections such as Wikisource and Gutenberg.org as well as smaller, more specific sources were critically reviewed to identify text resources appropriate to serve as valuable extensions of a growing reference corpus for the historical German language. The selected items were aggregated and standardized with respect to their storage and annotation format; structural and bibliographic information was enhanced and corrected, if necessary.

To reach its aims, the curation project fortunately could make extensive use of the elaborated technical infrastructure and, not less important, the encompassing documentation of transcription- and format-specific guidelines developed by the DTA.

2. Exemplary workflow: the DTA and its enhancement module DTAE

The DTA project started in 2007 and is building a TEI-XML-annotated full-text corpus of German-language texts. More than 1,300 volumes printed between the 17th and the 19th century will be processed and published online until 2014/15. Scientific texts, as well as fiction, poetry, drama, 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 sources 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, show a high level of accuracy and are annotated with structural information following the TEI P5-compliant DTA ‘base format’ (DTABf).[9] The electronic full-texts are enriched with linguistic information in stand-off markup gained

data via OAI-PMH, 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.

[9] DTABf: Deutsches Textarchiv – Basisformat, www.deutschestextarchiv.de/doku/basisformat. The DTABf 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 means, the DTABf aims at gaining coherence at the annotation level, given the heterogeneity of the DTA texts regarding time of origin (~1600-1900) and text type (e.g. fiction, functional texts, or scientific texts). Cf. Geyken/Haaf/Wiegand (2012). The DTABf is recommended as best practice for the structural encoding of historical printed texts within CLARIN-D. Cf. CLARIN-D AP 5 (2012): Part II, ch. 6, subsection “Text Corpora”.

through tokenization, lemmatization, and part-of-speech-analysis. Each text is analyzed with CAB, a set of rewrite rules for automated orthographic normalization of historical text material.[10]

The prospect of substantially more than 1,300 original texts from three centuries to be published until 2014/15 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 represented in the corpus than others, and the number of witnesses per decade may – for some purposes – seem relatively small. So, to enhance DTA's '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. By the time this article was written (April 2014), there are 1,312 texts (core corpus plus curated extensions) dating from between the 16th and early 20th century online, comprising a total of more than 423,000 digitized pages with more than 691 million characters and roughly 100 million tokens. More than 500 additional volumes, mainly from the period between 1600 and 1780, are prepared to be published and will likewise be made freely available under a Creative Commons license. In the course of the curation project described here, DTAE was used as the platform for conversion and publication of high-quality resources from various contexts. The resources were 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 is taken care of.

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, docx, plain text, PDF, ...) into the DTABf. Thereby, the DTAE infrastructure and tools facilitate the production of new high-quality transcriptions of primary sources in cooperation between the DTA and external researchers, as well as allowing for the integration and enhancement of existing resources. Both ways of corpus building have been followed successfully at the DTA: co-operative text production – for example, together 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) – and corpus build-

[10] Cf. Jurish (2010; 2011). CAB provides an automated normalization of the historical orthography in order to allow for lemma-based, spelling-tolerant corpus searches.

ing via integration and enhancement of existing resources – for example, from born-digital scholarly editions like those of the works of J. v. Sandrart, J. F. Blumenbach, and from the centenary-spanning 'Polytechnisches Journal' founded by J. G. Dingler.[11] 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 in the context of the curation project.

3. Digital curation: select, enhance and preserve distributed resources

3.1 Criteria: what to look for?

Among a great number of possible sources, appropriate items for the curation project were identified with the help of a set of criteria. Generally speaking (and deliberately in contrast to other corpus building approaches), the curation project put an emphasis on quality over quantity:[12] This meant rather 'hand-picking' than following a 'DownThemAll!' approach, where, as a prize for the greater amount of data to be gained in one 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. To overcome problems of format obsolescence and inflexibility, conversion of the data into a consistent, standardised and flexible format such as (TEI-) XML was central. However, the amount of time and manual work this process requires differs strongly depending on the data base. With this in mind, it was decisive for the success of the curation project to have kept a sound balance between the effort it took to integrate the 'chosen ones' and the (anticipated) value they represent to the research community addressed in CLARIN-D, and to carefully have weighed the quality against the quantity of the aggregated resources.

[11] For further information on these editions, cf. the projects' respective web sites: www.sandrart.net, www.blumenbach-online.de, and www.polytechnischesjournal.de.

[12] Nevertheless, selected 'working transcriptions' were also integrated and 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 were taken into consideration, and members of the community were encouraged to help improve the resources, e.g., by proofreading and correcting.

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 towards 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. Any alterations, e.g. the replacement of certain letters like the long *s* (*ſ*) by the 'modern' round *s*, the dissolution of ligatures, 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 characters (preferably 99.5+ %) and, with respect to the annotation, should contain at least the most basic structural information (i.e. divisions/chapters, headers, paragraphs). Furthermore, the texts 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, single-handed transcription, double keying, ...), its creator and its editing status (completed, draft, working transcription, ...). The image scans should show a high resolution, preferably be full-colour master copies with ≥ 300 DPI in TIFF or JPEG2000 format. The metadata describing the source should be accurate and as detailed as possible (while it certainly still has to be complemented in the curation process, if only for the purpose of marking versions and stating editing responsibilities in the life cycle of the document).[15] Certainly, legal aspects concerning text, metadata and images will have to be sorted out: each item, i.e. images, metadata and text should be available under a free license at least for reuse in a scientific context.

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

[14] The marking of page breaks 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/Wiegand/Geyken (2013).

[15] See DCC (2007) for an illustration of a Curation Lifecycle Model.

Although, at first glance, these criteria might seem to form quite a low threshold, in the course of the curation project they still helped to guarantee a high quality and the integrity of the acquired data and provided a good orientation to separate the wheat from the chaff.[16] For example, most of the texts represented in the text collection of Gutenberg-DE[17] – and, although with some notable exceptions, also those of zeno.org[18] – did not meet the curation project's criteria in every respect. A great number of the transcriptions available there 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 simply did not show the accuracy required. So, instead of the two large collections Gutenberg-DE and zeno.org, quantitatively smaller, but – considering the curation project's criteria – qualitatively better sources became the major points of interest for the project.

3.2 Sources: where to look?

3.2.1 Large text (and image) collections: Wikisource and Project Gutenberg

The German partition of Wikisource and German-language texts from the American Project Gutenberg (PG) proved to be the most fruitful sources for a considerable amount of documents fulfilling the criteria described above.[19]

[16] As a welcome side effect, the criteria helped to narrow the focus of the project described here to a manageable amount of text resources.

[17] Projekt Gutenberg-DE, http://gutenberg.spiegel.de/.

[18] Zeno.org, www.zeno.org. In 2009, the whole collection was acquired by the research infrastructure project TextGrid, funded by the BMBF. The text files from zeno.org were converted into the TextGrid '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: 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. www.textgrid.de/en/digitale-bibliothek/.

[19] 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 were also integrated. For example, the accurate transcription of Hans Stadens "Warhaftige Historia und beschreibung eyner Landtschafft

The focus in the following passage is on Wikisource, which proved to be the richest source for appropriate texts. The quality of the single resources assembled in 'opportunistic' collections like Wikisource with its many individual contributors differs strongly, but nonetheless several high quality representations of historic documents could be discovered. The site offers accurate transcriptions of historic primary sources, often along with corresponding image scans in good quality. Unfortunately, the best items were somewhat hidden among the vast total number of objects. To make sure that its integration would be worth an effort, each possible candidate was evaluated following the criteria described in the previous section – a non-trivial task itself, given the amount of approx. 30,000 German-language texts (as of April 2014) in the German Wikisource.[20]

The metadata describing the collected objects displayed on the website often proved to be not sufficient to serve as a basis for a systematic selection of single items. The navigational structure of the site is rather opaque, and the on-site retrieval facilities are quite basic. The options to browse and search the collection are rather limited and it is hard to get an overview.[21] This holds true for Wikisource, but also for Project Gutenberg and other large scale collections under consideration. Therefore, the sites in focus had to be critically scoured manually pursuing different strategies. From Wikisource, the most prolific source among the large scale collections, 1,891 high-quality texts containing almost 20,000 pages were identified and integrated.[22]

der Wilden / Nacketen / Grimmigen Menschfresser Leuthen [...]." (Marpurg [Marburg], 1557), www.deutschestextarchiv.de/staden_landschafft_1557, was integrated from Gutenberg-DE. A number of works of female writers, a group notoriously under-represented in corpora of historical printed works, was drawn from zeno.org, cf. www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=zeno. Further additions where derived from Sophie – A Digital Library of Works by German-Speaking Women (http://sophie.byu.edu/), for example Louise Aston's "Aus dem Leben einer Frau" (Hamburg, 1847), www.deutschestextarchiv.de/aston_leben_1847.

[20] Cf. http://de.wikisource.org, Hauptseite > Wikisource Aktuell > Statistik.

[21] Unfortunately, Wikisource offers no query or download API for ingesting the full descriptive metadata of the 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.

[22] Cf. www.deutschestextarchiv.de/doku/clarin_kupro_liste?g=wikisource.

3.2.2 Research projects and scholarly editions

As a second domain for historical text resources, research projects and scholarly editions were taken into account, as their data in general incorporate the expertise and scrutiny of acknowledged specialists. Without doubt, the fruits of their labour were of high interest for the purpose of this curation project,[23] but first of all, the data had to be retrieved and often enough legal issues had to be solved: sometimes, access was impossible even to the 'raw' data of the project, e.g. because of restrictive contracts with publishing houses.[24] Both tasks, retrieving the data and securing access to it, were even harder to accomplish in cases where the research project in question had already ended: staff members were off to other places, while the work done – especially the fundamental steps *before* the publication of the research outcomes – often was hardly documented. As one result of this, the project-specific transcription and the markup conventions applied had become hard to comprehend by others. They had to be reconstructed a) in order to be able to evaluate the resource in the first place and b), if the item was to be integrated, in order to perform a lossless conversion of the data into the DTABf.

If the data was available for integration, a further and no less severe problem concerned its storage format. Until recently, the majority of scholarly editions of historical text material were produced with the goal of a printed (or print-

[23] A very successful cooperation was established with the editors of the historical-critical edition of Karl Gutzkow's works and letters, http://projects.exeter.ac.uk/gutzkow/Gutzneu/. The HTML-representation of Gutzkow's primary works was converted into the DTABf, encompassing linguistic analysis and indexing for full-text queries (which the Gutzkow edition itself does not offer) and allowing for collaborative quality assurance in DTAQ (cf. below, ch. 3.3). By preserving all editorial and other comments originally present in HTML and by discussing and carefully documenting the steps of conversion, the integration of all texts from the Gutzkow edition into the DTA corpus remains reversible. This was a crucial point for the fruitful cooperation between the projects. For example, if transcription errors are corrected or the text base is changed for other reasons in the process of quality assurance in DTAQ, or if the transcription is annotated more deeply (for example by annotating named entities), the results of this work can be re-transformed from DTA's site into the Gutzkow project.

[24] '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 taken less 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.

like) documentation of the work in mind.[25] Therefore, the text base was produced with the help of GUI-based text processors and other office tools. It was published and/or stored in formats such as MS doc or docx, Adobe InDesign, PDF or LaTeX. 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 text processors and their output formats tend to indistinguishably mix layout information with structural information. Therefore, the data demanded a notable amount of manual labour in reformatting to preserve the intellectual work explicitly and implicitly contained in the documents.[26]

3.2.3 Special collections and single resources

Finally, and in addition to large scale collections and scholarly projects, smaller compilations of texts on a certain topic, representing a particular discourse or epoch were considered. Often built and run by enthusiastic private scholars or layman investing a lot of energy and spare time, these thematic collections may reveal astonishing discoveries. Single findings were integrated, fortunately always with the approval and sometimes also with the support of their producers.[27]

[25] 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.

[26] Two examples to illustrate the outcome of this laborious, but worthwhile effort of data conversion are Theresia Lindnerin's "Koch Buch zum Gebrauch der Wohlgebohrenen Frau" (around 1780), which was transcribed, annotated and published as a PDF file under http://geb.uni-giessen.de/geb/volltexte/2009/7361/ by Thomas Gloning from the University of Gießen, and is now available in TEI-XML under www.deutschestextarchiv.de/lindnerin_kochbuch_1780; and a transcription of "Petrus de Crescentiis zu teutsch mit Figuren. Speyer, ca. 1493", www.deutschestextarchiv.de/crescentiis_figuren_1493, which Jakub Šimek had published as part of his University of Heidelberg's Magister thesis (cf. http://crescenzi.dyskanti.com/) and stored in LaTeX. We are grateful to both editors for offering their documents and their instructive comments on how to convert the valuable information.

[27] See, for example, Joseph Schauberg's three-volume "Vergleichendes Handbuch der Symbolik der Freimaurerei" (1861-63), www.deutschestextarchiv.de/schauberg_freimaurerei01_1861, .../schauberg_freimaurerei02_1861 and .../schauberg_freimaurerei03_1863, derived from the "Portal to the World of Freemasonry", www.internetloge.de. Altstuhlmeister Franz-L. Bruhns, webmaster and editor of this non-commercial web page, happily agreed to the re-use of the HTML-representation of the "Handbuch" (www.internetloge.de/symhandb/symb.htm) on the one condition that www.internetloge.de is appropriately credited as creator of the original transcription.

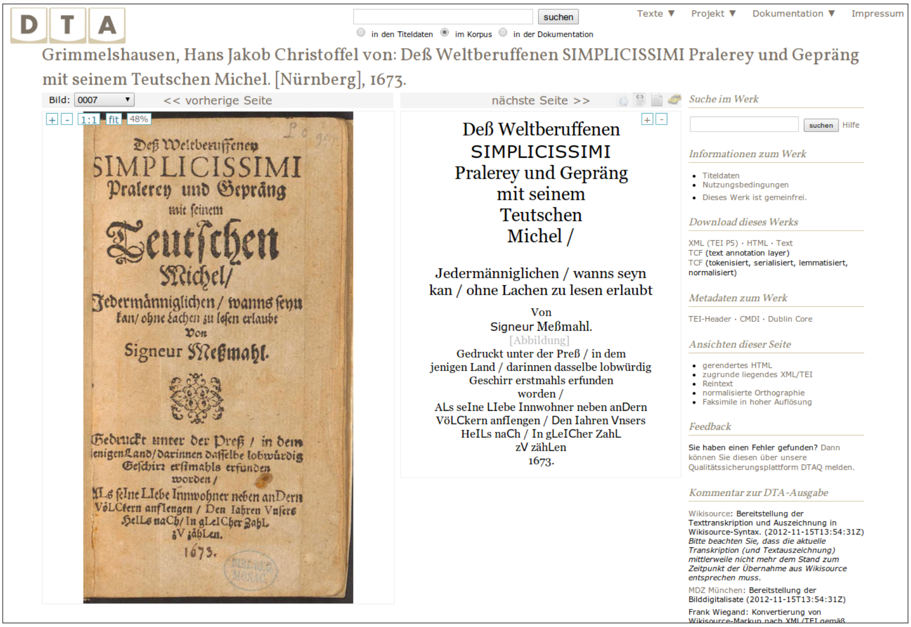
Now that the major sources of high-quality resources have been described and before the process of their integration in the course of the curation project is outlined, a word on appreciation and responsibilities is due. In order to establish a culture of shared access and usage, the importance of a reputation system must not be omitted. Therefore, and for each single item integrated, the appreciation of the work of others was made visible in the source documentation. Also, responsibilities in every stage of the text refinement were made transparent.

3.3 Integration: how to proceed?

Once a relevant resource meeting the named criteria was identified, the full-text transcription, image scans and metadata were acquired and integrated into DTA's enhancement module DTAE. In the course of this, the electronic documents were enriched with the acquired and enhanced bibliographic and structural information. In the next step, the bibliographic data and full-text transcription were converted into the DTABf. The text and metadata were then published alongside the corresponding image scans via the DTAE framework; it was also made available via the BBAW's CLARIN-D repository.[28] Each text was analyzed with CAB for automated orthographic 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 complex queries in the growing text corpus. The linguistic analysis furthermore encompasses tokenization, lemmatization, and PoS-tagging in stand-off markup.

Each integrated text can now be displayed page-wise in an HTML representation automatically rendered from the underlying TEI-XML (Fig. 1), it can be searched and explored as a single resource, in the context of the different subcorpora compiled by the DTA, in the context of the DTA 'core corpus' and in the greater context of all corpora available in CLARIN-D. The resource descriptions and bibliographic information are standardized conformant to authority formats (e.g. CMDI or DC) in order to be shared via OAI-PMH and to be integrated into CLARIN-D's service architecture.

[28] Cf. the repository at the CLARIN Service Center of Zentrum Sprache at the BBAW, http://clarin.bbaw.de/.



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

Deß Weltberuffenen
SIMPLICISSIMI
Pralerey und Gepräng
mit seinem
Teutschen
Michel /

Jedermänniglichen / wanns seyn
kan / ohne Lachen zu lesen erlaubt

Von
Signeur Meßmahl.
[Abbildung]
Gedruckt unter der Preß / in dem
jenigen Land / darinnen dasselbe lobwürdig
Geschirr erstmahls erfunden
worden /
ALs seIne LIebe Innwohner neben anDern
VöLCkern anfIengen / Den Iahren Vnsers
HeILs naCh / In gLeICher ZahL
zV zähLen
1673.

Suche im Werk

suchen Hilfe

Informationen zum Werk

- Titeldaten
- Nutzungsbedingungen
- Dieses Werk ist gemeinfrei.

Download dieses Werks

XML (TEI P5) · HTML · Text
TCF (text annotation layer)
TCF (tokenisiert, serialisiert, lemmatisiert, normalisiert)

Metadaten zum Werk

TEI-Header · CMDI · Dublin Core

Ansichten dieser Seite

- gerendertes HTML
- zugrunde liegendes XML/TEI
- Reintext
- normalisierte Orthographie
- Faksimile in hoher Auflösung

Feedback

Sie haben einen Fehler gefunden? Dann können Sie diesen über unsere Qualitätssicherungsplattform DTAQ melden.

Kommentar zur DTA-Ausgabe

Wikisource: Bereitstellung der Texttranskription und Auszeichnung in Wikisource-Syntax. (2012-11-15T13:54:31Z) *Bitte beachten Sie, dass die aktuelle Transkription (und Textauszeichnung) mittlerweile nicht mehr dem Stand zum Zeitpunkt der Übernahme aus Wikisource entsprechen muss.*

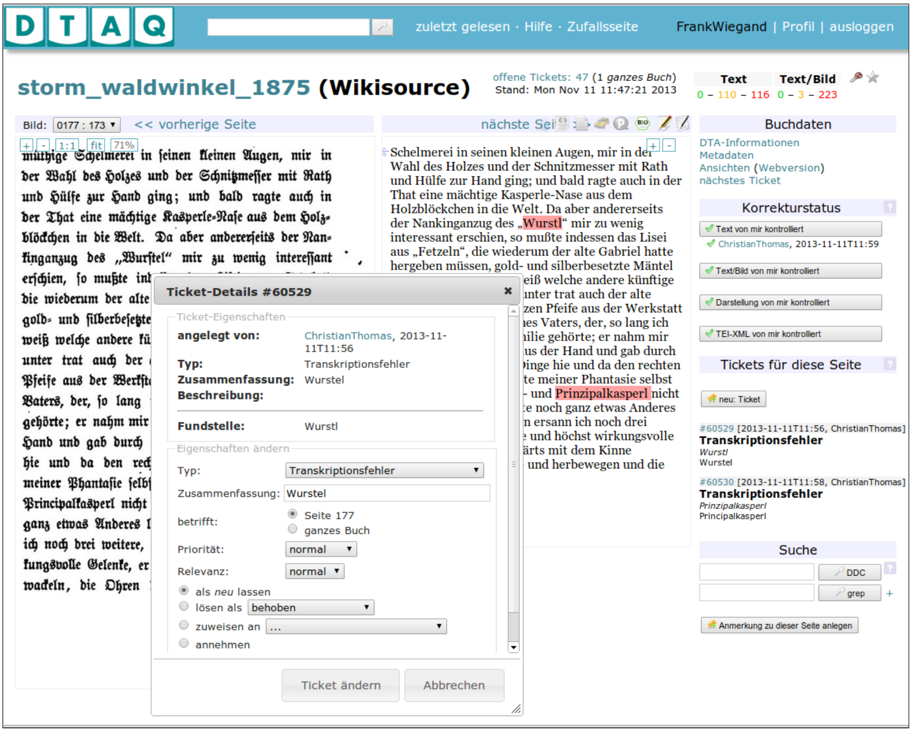
MDZ München: Bereitstellung der Bilddigitalisate (2012-11-15T13:54:31Z)

Frank Wiegand: Konvertierung von Wikisource-Markup nach XML/TEI gemäß

Figure 1: Wikisource-item integrated into the DTA corpus: image, transcription in rendered HTML, metadata and further information 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, Title Page / image 7. In: Deutsches Textarchiv, www.deutschestextarchiv.de/grimmelshausen_michel_1673/7

In parallel, all items can be accessed via DTA's quality assurance platform DTAQ.[29] In DTAQ, texts may be proofread page by page in comparison to their source images (Fig. 2). This way, errors that may have occurred during the former transcription and annotation process, or that were overlooked or not taken care of during integration can be detected and corrected. 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.

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



storm_waldwinkel_1875 (Wikisource)

offene Tickets: 47 (1 *ganzes Buch*)
Stand: Mon Nov 11 11:47:21 2013

müthige Schelmerei in ſeinen kleinen Augen, mir in
der Wahl des Holzes und der Schnitzmeſſer mit Rath
und Hülfe zur Hand ging; und bald ragte auch in
der That eine mächtige Kasperle-Naſe aus dem Holz-
blöckchen in die Welt. Da aber andererſeits der Nan-
kinganzug des „Wurſtel“ mir zu wenig intereſſant
erſchien, ſo mußte ind
die wiederum der alte
gold- und ſilberbeſetzte
weiß welche andere kü
unter trat auch der
Pfeife aus der Werkſt
Vaters, der, ſo lang
gehörte; er nahm mir
Hand und gab durch
hie und da den rech
meiner Phantaſie ſelbſ
Principalkasperl nicht
ganz etwas Anderes
ich noch drei weitere,
kungsvolle Gelenke, er
wackeln, die Ohren

Schelmerei in seinen kleinen Augen, mir in der
Wahl des Holzes und der Schnitzmesser mit Rath
und Hülfe zur Hand ging; und bald ragte auch in der
That eine mächtige Kasperle-Nase aus dem
Holzblöckchen in die Welt. Da aber andererseits
der Nankinganzug des „Wurstl“ mir zu wenig
interessant erschien, so mußte indessen das Lisei
aus „Fetzeln“, die wiederum der alte Gabriel hatte
hergeben müssen, gold- und silberbesetzte Mäntel
eiß welche andere künftige
unter trat auch der alte
zen Pfeife aus der Werkstatt
nes Vaters, der, so lang ich
nilie gehörte; er nahm mir
us der Hand und gab durch
Dinge hie und da den rechten
te meiner Phantasie selbst
- und Prinzipalkasperl nicht
te noch ganz etwas Anderes
n ersann ich noch drei
e und höchst wirkungsvolle
ärts mit dem Kinne
und herbewegen und die

Ticket-Details #60529

Ticket-Eigenschaften

angelegt von: ChristianThomas, 2013-11-11T11:56
Typ: Transkriptionsfehler
Zusammenfassung: Wurstel
Beschreibung:

Fundstelle: Wurstl

Eigenschaften ändern

Typ: Transkriptionsfehler
Zusammenfassung: Wurstel
betrifft: ● Seite 177 ○ ganzes Buch
Priorität: normal
Relevanz: normal
● als *neu* lassen
○ lösen als behoben
○ zuweisen an …
○ annehmen

✓ ChristianThomas, 2013-11-11T11:59

#60529 [2013-11-11T11:56, ChristianThomas]
Transkriptionsfehler
Wurstl
Wurstel

#60530 [2013-11-11T11:58, ChristianThomas]
Transkriptionsfehler
Prinzipalkasperl
Principalkasperl

Figure 2: Quality Assurance in DTAQ: image, transcription in rendered HTML; 'ticket' system to report findings, e.g. transcription errors, printing errors, and inconsistencies in annotation. Storm, Theodor: Waldwinkel, Pole Poppenspäler. Novellen. Braunschweig, 1875, p. 173 / image 177. In: Deutsches Textarchiv – Qualitätssicherung, www.deutschestextarchiv.de/dtaq/book/view/storm_waldwinkel_1875?p=177 [retrieved 2013-11-11; the transcription errors highlighted in the illustration have been corrected in the meantime]

4. Conclusion

In the course of the CLARIN-D curation project described here, the equivalent of more than 79,000 pages was integrated into a large corpus for the written German language between the 15th and the 19th centuries. From the large text repository consisting of the DTA's, HAB's, IDS' and many other partner's corpora available under the roof of CLARIN-D, more balanced reference corpora can now be derived. In this respect, the curation project helped to improve the situation for corpus-based research, particularly in historical linguistics, but also in the humanities in general. By applying a consistent, interoperable encoding based on the recommendations of the TEI[30] and by integrating the resources into the CLARIN-D infrastructure, the data can now be explored in a broader context. Access to and sustainability of these resources were thereby improved substantially. A formerly dispersed, large variety of corpus texts can now 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 the integration of resources of different origin, CLARIN-D contributes significantly to the scholarly community. The idea of curating and sharing corpus resources in a collaborative manner was put into practice with excellent results – which hopefully will encourage similar initiatives.

Affiliation

CLARIN-D curation project "Integration und Aufwertung historischer Textressourcen des 15.-19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur", cf. www.deutschestextarchiv.de/clarin_kupro for an overview on the project and for a list of texts integrated.

References

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 7). www.balisage.net/Proceedings/vol7/cover.html, doi:10.4242/BalisageVol7.Bauman01.

Bailey, Jr., Charles W. (2012): Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works. http://digital-scholarship.org/dcpb/dcb.htm.

CLARIN-D AP 5 (2012): CLARIN-D User Guide. Version: 1.0.1, Publication date: 2012-12-19. www.clarin-d.de/en/language-resources/userguide.html.

Digital Curation Centre (DCC) (2007): What is digital curation? www.dcc.ac.uk/digital-curation/what-digital-curation; DCC curation lifecycle model. www.dcc.ac.uk/resources/curation-lifecycle-model.

Geyken, Alexander/Haaf, Susanne/Wiegand, Frank (2012): The DTA 'base format': A TEI-subset for the compilation of interoperable corpora. In: Jancsary, Jeremy (ed.): 11th Conference on Natural Language Processing (KONVENS): Empirical Methods in Natural Language Processing. Proceedings of the Conference on

[30] Cf. Bauman (2011) and Unsworth (2011).

Natural Language Processing 2012. (= Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5). Wien: ÖGAI, 383-391. www.oegai.at/konvens2012/proceedings.pdf#page=383.

Geyken, Alexander/Haaf, Susanne/Jurish, Bryan/Schulz, Matthias/Thomas, Christian/Wiegand, Frank (2012): TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv. In: Jahrbuch für Computerphilologie, online version. www.computerphilologie.de/jg09/geykenetal.html.

Haaf, Susanne/Wiegand, Frank/Geyken, Alexander (2013): Measuring the correctness of double-keying: error classification and quality control in a large corpus of TEI-annotated historical text. In: Journal of the Text Encoding Initiative 4. http://jtei.revues.org/739, doi:10.4000/jtei.739.

Jurish, Bryan (2010): More than words: using token context to improve canonicalization of Historical German. In: Journal for Language Technology and Computational Linguistics (JLCL) 25/1: 23-39. http://media.dwds.de/jlcl/2010_Heft1/bryan_jurish.pdf.

Jurish, Bryan (2011): Finite-state canonicalization techniques for Historical German. PhD thesis, Universität Potsdam. http://opus.kobv.de/ubp/volltexte/2012/5578/; urn:nbn:de:kobv:517-opus-55789.

Lee, Christopher A./Tibbo, Helen R. (2007): Digital curation and trusted repositories: steps toward success. In: Journal of Digital Information (JoDI) 8(2): Digital Curation & Trusted Repositories. http://journals.tdl.org/jodi/article/view/229/183.

Rusbridge, Chris/Burnhill, Peter/Ross, Seamus/Buneman, Peter/Giaretta, David/Lyon, Liz/Atkinson, Malcolm (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. http://eprints.erpanet.org/82/.

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

Unsworth, John (2011): Computational work with very large text collections. In: Journal of the Text Encoding Initiative 1. http://jtei.revues.org/215, doi:10.4000/jtei.215.

Whyte, Angus/Wilson, Andrew (2010): How to Appraise and Select Research Data for Curation. (= DCC How-to Guides). Edinburgh: Digital Curation Centre. www.dcc.ac.uk/resources/how-guides.