**Journal of the Text Encoding Initiative**

**Issue 8 | December 2014 - December 2015**
**Selected Papers from the 2013 TEI Conference**

# The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources

**Susanne Haaf, Alexander Geyken and Frank Wiegand**

**Electronic version**
URL: http://journals.openedition.org/jtei/1114
DOI: 10.4000/jtei.1114
ISSN: 2162-5603

**Publisher**
TEI Consortium

**Electronic reference**
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* [Online], Issue 8 | December 2014 - December 2015, Online since 09 April 2015, connection on 21 April 2019. URL : http://journals.openedition.org/jtei/1114 ; DOI : 10.4000/jtei.1114

# *The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources*

**Susanne Haaf, Alexander Geyken, and Frank Wiegand**

## 1. Introduction

1 Until recently the creation of large historical reference corpora was, from the point of view of its encoding, a rather project-specific activity. Although reference corpora were built from texts of various origins, the texts had to be converted into a tailor-made format. For example, corpora like the well-known British National Corpus[1] and the DWDS core corpus (Geyken 2007) are both annotated on the basis of the Guidelines of the Text Encoding Initiative (most recent release: P5; see TEI Consortium 2014). However, the encoding in these cases was typically carried out specifically for the creation of these corpora—that is, it was a unidirectional process. The interchange, and more importantly the interoperability, of corpora with other corpora played only a minor role.

2 In recent years the picture has dramatically changed. With the availability of more and more digitized texts in TEI P5 format and the advent of many academic and non-academic corpus-building projects, the task of creating reference corpora has shifted from a project-specific task to a more general task, requiring joint efforts by many stakeholders. In this new situation, individual

projects creating corpora must ensure that their resulting annotation scheme is valid against a TEI P5 schema, and also enables easy reuse of both their metadata and text data in other project contexts. In other words, corpus projects are required to provide interoperable TEI data, so that the resulting corpora compiled from different sources can be made exploitable by common methods and tools. Interoperability issues affect different aspects, from metadata exchange through the extraction and analysis of document components up to (at least for historical texts) the creation of a uniform stylesheet in order to present all corpus texts in a similar way.[2]

3 A main prerequisite for interoperability between corpora is the homogeneity of text structure. Thus, because of the intentionally high flexibility of the TEI tag set,[3] it is no longer sufficient to base the annotation on the TEI Guidelines in their entirety. Rather, the TEI tag set has to be narrowed down to subsets that are both extensive, considering the structural phenomena they must document, and unambiguous about how similar phenomena may be encoded. Given these requirements, the desirability of creating some common agreed-upon TEI formats for certain editing purposes—and sharing these formats with the community in order to achieve homogeneously tagged TEI texts across project borders—is evident and has already been attempted several times (see, e.g., Pytlik Zillig 2009; Unsworth 2011).

4 Interoperability problems with TEI-encoded documents can occur on several levels: the exchange of metadata, the consistent extraction of document components, and the creation of a uniform stylesheet. It is well known that the flexibility of the TEI may lead to significant structural variation in TEI-conformant headers. This forces computational methods to deal with an enormous number of different cases for semantically consistent information extraction. Within the transcription itself, similar problems may occur with the extraction of structural phenomena in a text collection. For example, letters or quotations can be annotated differently across different document collections. Therefore it can be very difficult or even impossible to formulate a query that retrieves "all letters" across all document collections without knowing all the solutions adopted by all the individual collections. For complex queries the problem becomes even harder. Obviously, a standardized encoding across collections would be very helpful in solving this problem.

5 Last but not least, structural variation poses difficulties which should not be underestimated to the creation of a satisfactory rendering stylesheet that maps similar structural phenomena to identical styles. As a consequence, at present, document collections encoded in TEI can be exchanged only by accepting the loss of interoperability on one or several of the above-mentioned levels.

6 Existing TEI P5 subsets either were provided by the TEI Consortium, such as TEI Lite (Burnard and Sperberg-McQueen 2012) and TEI Tite (Trolard 2011), or originate from text digitization and curation projects, for example the Best Practices for TEI in Libraries (TEI SIG on Libraries 2011), TEI Analytics (Unsworth 2011; Pytlik Zillig 2009), I5 (Lüngen and Sperberg-McQueen 2012), or TextGrid's Baseline Encoding for Text Data in TEI P5 (TextGrid 2007–2009).

7 A shared goal of these formats is to allow for the encoding of basic structural properties such as the markup of chapters and paragraphs. These formats agree on many of the most common structural features (e.g., a paragraph is marked up using the element <p>). However, each of these formats is designed for particular audiences and goals, and hence deals with the range of TEI possibilities for the markup of similar structures according to its own needs.

8 In this article we describe the DTA "Base Format" (DTABf), a strict subset of the TEI P5 tag set. Its purpose is to provide a balance between expressiveness and precision as well as an interoperable annotation scheme for a large variety of text types in historical corpora of printed text from multiple sources. The DTABf shares properties and tagging solutions with the TEI P5 subsets mentioned above, but differs from those formats in several aspects which we described comprehensively in Geyken, Haaf, and Wiegand 2012. For instance, unlike TEI Lite and "Best Practices for TEI in Libraries," the DTABf uses controlled vocabularies for the specification of attribute values, and unlike TEI Tite, the DTABf forms a strict subset of the TEI tag set (i.e., no extensions of or content changes to the TEI P5 tag set are made) and offers a detailed specification for metadata recording within the TEI header.

9 The DTABf was created in a bottom-up process alongside the corpus compilation of the DTA project, thus covering most of the predictable structural phenomena of historical texts. It was applied to 15 text corpora from scholarly digitization projects. In addition, there is a powerful infrastructure implemented by the DTA project in order to process, correct, annotate, and publish texts in DTABf. These principles have helped establish the DTABf as a common, shared TEI format for the annotation of historical German texts.

10 The remainder of this paper starts with a short presentation of the project background, the Deutsches Textarchiv (DTA), which will show that a common base format was required to integrate text collections from 15 external corpus projects (section 2). In section 3, we describe the DTABf in more detail. Section 4 focuses on the dissemination of the DTABf; we explain how comprehensive documentation, continuous training courses, and the development of customized software tools interact to create a user community for the DTABf. In section 5, we present examples of good practice that illustrate how different external corpora can be converted into the DTABf, making such corpora interoperable in a wider context—for example, as part of the text corpora provided by the large European infrastructure project CLARIN.[4] Section 6 discusses how new structural phenomena encountered in new texts are handled within the DTABf by adding new properties to the DTABf. We conclude with a short summary and some ideas about future prospects for the DTABf.

## 2. Project Background

11 The goal of the project Deutsches Textarchiv (DTA)[5] is to create a reference corpus of the historical New High German language (1600–1900) which is balanced with regard to date of creation, text type, and thematic scope, and therefore constitutes the basis of a reference corpus for the New High German language. The DTA corpora contain printed historical works from different genres and text types (fictional texts in prose, poems, dramas, scientific texts of numerous different disciplines, and functional literature such as cookbooks, handbooks, sermons, or travel books). The DTA acquires texts in two complementary ways: via digitization of new texts for the DTA core corpus of around 1,500 historical works (approximately 150 million tokens) and via the curation of existing historical text collections, digitized in other project contexts, which are integrated into the DTA infrastructure (currently 120 million tokens from 15 projects).[6]

12 Even though text digitization for the entire core corpus is carried out by the DTA project itself and therefore can be based on similar guidelines, the DTA core corpus texts are heterogeneous due to structural differences between text types, variation in printing habits over time and in different regions, and peculiarities of different printing and publishing houses. External texts from different sources being integrated into the DTA infrastructure introduce additional heterogeneity in terms of project-specific markup conventions and different primary formats.

13 In order to cope with this heterogeneous text collection, the DTABf was developed as an annotation scheme that allows for collective processing by software tools including metadata harvesting, retrieval of complex document structures, and presentation of the text data in various formats, including HTML, Text, and ePUB. The DTABf is completely based on the TEI P5 tag set, reducing and further constraining the stock tag set, but not extending it in any way. The goal of the DTABf is to provide solutions for the tagging of all structural phenomena occurring in historical printed texts down to a certain annotation depth while remaining consistent and unambiguous to ensure consistent markup for the DTA corpora as a whole. Thus the DTABf is the backbone of the DTA and guarantees that all DTABf texts are interoperable within the DTA context as well as for re-use in other projects. The DTABf is a "living" TEI format. It is carefully adjusted when new texts containing new structural phenomena are integrated into the DTA corpora.

# 3. Description of the DTA "Base Format" (DTABf)

## 3.1 History and Scope

14 The DTABf emerged from the TEI format used for the annotation of the DWDS corpus, a balanced corpus for twentieth-century German (Geyken 2007). With the beginning of the DTA project, the DWDS format was adapted to the requirements of the encoding of historical texts. The DTABf was applied continuously to all texts digitized during the first phase of the DTA project (2007–2010, approximately 700 texts dating from 1780 to 1900). In this period it was successively adapted to new phenomena which occurred in the respective texts and which had not previously been covered by the DTABf. With the beginning of DTA phase 2 (2010–2014, approximately 600 texts dating from 1600 to 1780), the DTABf was extensively revised on the basis of the annotated historical data resulting from phase 1: the treatment of structural phenomena was reconsidered and consistent solutions were determined. In the course of these efforts the formal description of the DTABf as given in a corresponding ODD[7] was further substantiated and the DTA annotation guidelines were compiled.

15 The revised version of the DTABf has been successfully applied not only to the phase 2 texts of the DTA core corpus, but also to texts originating from external project contexts which were curated by the DTA in the course of the module DTAE (DTA-Erweiterungen, i.e., DTA Extensions) as well as

the curation project of WG 1 in CLARIN-D.[8] Continuous adjustments of the DTABf remain necessary in order to account for new phenomena, but we take great care to preserve consistency and avoid ambiguity. In addition, since the DTABf is now based upon a large amount of text, we are mostly able to avoid changes which disturb backward compatibility.[9]

16 As a result, the DTABf forms a TEI customization which is based on a large corpus of historical texts and thus offers solutions for most structural phenomena encountered within historical printed texts.

## 3.2 Design and Components

17 The DTABf tag set not only offers tagging solutions for text structuring but also provides a specification for the description of metadata in the TEI header. As of May 2014, the DTABf consists of 50 TEI header elements and 75 text elements accompanied by limited element- or class-specific attributes and, where applicable, attribute values. See Appendix 1 and Appendix 2 for more information on the distribution of DTABf elements within the DTA corpus.

### 3.2.1 Recording of Metadata

18 The DTABf TEI header is designed to cover extensive metadata information.[10] First, DTABf-conformant metadata records contain bibliographic information about (1) the digital document as published by the DTA, (2) all instances which preceded the current digital edition together with (3) the persons or organizations responsible for those instances, and (4) all licenses relevant for the digital object at hand. Second, descriptions of the physical text source upon which the current edition is based are required, including information about its constitution as well as its physical location (institution, repository, and shelfmark). Finally, general information is given about the content and design of the document (e.g., language, typeface, document type) and the DTA sub-corpus it belongs to.

### 3.2.2 Formal and Semantic Structuring of Text

19 The tag set for text encoding contains tagging solutions for formal as well as semantic text structures.[11] The former include page breaks, lists, tables, and figures, as well as physical layout information like forme work and different types of highlighting. The latter include chapters or text sections with titles, paragraphs, notes, opening or closing text parts, special text types

such as poems, letters, and indices, and inline phenomena such as proper nouns or citations. Furthermore, documented editorial interventions are possible (e.g., the correction of printing errors, the expansion of abbreviations, normalizations, and editorial comments).

### 3.2.3 Linguistic Tagging

20 Linguistic information—tokenization, lemmatization of historical forms, Part-of-Speech (POS) tagging—is acquired automatically by various tools and applied to the DTA texts via the standoff method.[12] We decided not to adopt an inline encoding for linguistic annotations for two reasons. First, the integration of token-based linguistic analyses in the texts leads to an enormous increase in the number of tags, which hinders manual editing of the transcriptions. The second reason is that postprocessing TEI texts—including postprocessing of linguistic annotations—often requires a conversion of the TEI text into a version of the text in which the reading order has been re-established (serialization). We provide such a solution for texts encoded in the DTABf schema (DTA-Tokwrap) and we prefer to provide users with linguistic annotations for our texts by converting the TEI texts into the Text Corpus Format (TCF),[13] the standoff format used within the CLARIN project.

### 3.2.4 Components of the DTABf

21 The DTABf now consists of five components:

1. an ODD file[14] specifying constraints on TEI elements, attributes, and attribute values
2. a RelaxNG schema[15] generated from the ODD
3. a set of Schematron rules complementing the schema
4. comprehensive documentation[16] explaining the treatment of structuring requirements
5. the TEI text instances provided by the DTA[17]

## 3.3 Ensuring Consistency: The DTABf ODD, Schema, and Schematron

22 In order to ensure homogeneous text annotation over the entire DTA corpus, the tag set has to remain unambiguous; that is, for each phenomenon there should be only one possible method of encoding.

23 Thus, we made use of the possibilities of the ODD source format to restrict annotations down to the attribute value level. First, from all possible TEI modules only a subset needed for our purposes was chosen. Second, from each of the included TEI modules, only a subset of available elements

needed to encode the DTA corpus texts was selected. Likewise, attribute classes or single attributes within certain classes were eliminated from the schema if they turned out to be unnecessary for our purposes. And finally, if applicable, each attribute (at the class or element level) was provided with a fixed selection of permitted values. In cases where value lists could not be restricted (e.g., `@n` on `<lg>` containing the number of a stanza), we set fixed data types for the respective attribute values wherever possible. There are only a few cases remaining where the restriction of fixed attribute values would not be reasonable (e.g., `@quantity` on `<gap>` specifies the amount of text left out in the transcription for whatever reason and thus can be filled with any numeral). With the restrictions provided by the DTABf schema the flexibility of the TEI P5 tag set is reduced in favor of unambiguous though still fully TEI P5 compliant solutions.

24 As stated above, the DTABf provides not only a vocabulary for text annotation but also a specification for the TEI header in order to allow for consistent metadata recording. Although those two vocabularies—the `<text>` tag set and the `<teiHeader>` tag set—are mutually exclusive within the DTABf to a large extent, the underlying TEI P5 schema allows for quite a number of the elements to be valid in both the `<text>` and the `<teiHeader>` areas.

**Example 1: Tagging of Notes and Remarks**

The `<note>` element may have different attribute-value pairs depending on where it is used: Within the `<text>` area notes may be marginal notes (`<note place="right|left">`), footnotes (`<note place="foot">`), endnotes (`<note place="end">`), or editorial remarks of the person working on the digital edition of the text (`<note type="editorial">`). Within the `<teiHeader>` of a document, however, other kinds of notes are relevant, e.g., remarks about responsibilities for certain instances of the digital document (`<note type="remarkResponsibility">`), about the digital document as a whole (`<note type="remarkDocument">`), or about the constitution of its physical source (`<note type="remarkSource">`).

**Example 2: DTABf `<teiHeader>` elements within `<text>`**

There is quite a significant number of DTABf `<teiHeader>` elements which the DTABf would not allow in the `<text>` area but which are allowed within `<text>` according to the TEI Guidelines. Examples are `<biblFull>`, `<msDesc>` and their descendants, `<respStmt>` and `<resp>` or the children of `<persName>` (e.g., `<addName>`, `<nameLink>`, or `<genName>`).

25 With the ODD vocabulary in itself we cannot change these constraints for elements while remaining fully TEI-compliant. Therefore, to solve this problem, until recently we provided two separate schemas: one representing the DTABf in its entirety, the other excluding the DTABf metadata tagset.[18] The latter formed an interim schema to facilitate DTABf-compliant text annotation irrespective of the recording and structuring of metadata, whereas the former constitutes the final schema as it is applied to the complete finalized DTABf documents. But the method of maintaining two separate schemas was time-consuming and thus only provisional. It has now been replaced by a set of Schematron rules which are defined on top of the DTABf schema and which specify contextual constraints.[19]

**Example 3: Schematron rules for `<teiHeader>` only elements**

This Schematron rule, which deals with elements which could be used in the `<teiHeader>` or `<text>` area but should, according to the DTABf, be restricted to the `<teiHeader>`, is:

```
  <rule context="tei:addName | tei:address | tei:addrLine | tei:email |
tei:biblFull | tei:country | tei:forename | tei:genName | tei:measure |
tei:msDesc | tei:nameLink | tei:publicationStmt | tei:resp | tei:respStmt |
tei:roleName | tei:surname | tei:titleStmt | tei:title">
    <report test="ancestor::tei:text" role="ERROR">
     [E0001] Element "<name/>" not allowed anywhere within element "text".
    </report>
  </rule>
```

26 In addition, Schematron is used for further quality assurance, such as for checking the content of certain elements or for specifying constraints on attribute values which cannot be expressed with conventional content models.[20]

**Example 4: Schematron rules for `@facs` values**

For example, the DTABf constraints for the `@facs` attribute of the element `<pb>` cannot be expressed within ODD using regular datatypes, but can be described by Schematron rules: The value of `@facs` is a string starting with `"#f"`, followed by a four-digit number; the `@facs` value of the first `<pb>` element in a document should be `"#f0001"`; the following `@facs` values should increase successively by 1.

The corresponding Schematron rules are:

```
  <rule context="tei:pb[1][not(preceding::tei:pb)]">
   <assert test="@facs[matches(., '^#f0001$')]" role="ERROR">
   [E0015] Value of @facs within first "pb" incorrect; expected value:
#f0001.
   </assert>
  </rule>
```

and

```
  <rule context="tei:pb[@facs]">
   <assert test="if (matches(@facs, '^#f\d\d\d\d') and
matches(preceding::tei:pb[1]/@facs, '^#f\d\d\d\d') and (preceding::tei:pb))
then xs:integer(substring(@facs, 3)) = preceding::tei:pb[1]/
xs:integer(substring(@facs, 3)) +1 else 1" role="ERROR">
   [E0014] Value of @facs within "pb" incorrect; @facs-values of "pb"-
elements have
   to increase by 1 continually starting with #f0001.
   </assert>
  </rule>
```

27 The more the DTABf grows, the more cautious we have to be with adding material (elements, attributes, or attribute values) to the tag set, considering that certain tagging solutions might be too flexible and/or overlap in their scope. Both cases might lead to misinterpretations and thus to inconsistencies in the application of the DTABf. Whenever changes to the DTABf become necessary, we have to carefully consider the DTABf tag set as a whole in order to ensure its consistency.

## 3.4 Annotation Levels

28 With the growth of the DTABf it gets increasingly difficult and time-consuming to apply the whole range of possible DTABf annotations to each DTA corpus text. Therefore, there must be a way to communicate the degree of conformity of a given TEI text with the other texts in the DTA corpus. For such scenarios, the TEI Guidelines propose to divide the set of elements used within a certain TEI format into different levels according to the necessity of their usage.[21] We introduced several levels of annotation, each containing a set of elements which have to be used consistently where applicable in order to achieve conformity with the respective level. The first three annotation levels are based on one another.

29 *Level 1 (required)* covers the minimal amount of text structuring which is required to be applied to a text in order to achieve DTABf conformity. Elements used at this level include `<div>`, `<head>`, `<p>`, `<lg>`, `<figure>`, `<pb>`, and `<cb>`.

30 *Level 2 (recommended)* contains additional elements (such as `<cit>`, `<opener>`, `<closer>`, `<lb>`, and `<hi>`) which are not required but still recommended by the DTABf. All DTA core corpus texts are consistently annotated up to level 2, hence all other level 2 texts are entirely interoperable with the DTA core corpus.

31 *Level 3 (optional)* defines optional elements (like `<persName>`, `<placeName>`, `<choice>` and its possible descendants, or `<foreign>`) which are part of the DTABf though not consistently applied to the DTA core corpus.

32 All elements contained in the first three annotation levels, together with their respective attributes and values, constitute the DTABf as a TEI P5 subset. Therefore, interoperability is completely ensured down to the optional level.

33 *Level 4 (proscribed)* contains TEI P5 elements which are proscribed by the DTABf because they represent text structures for which a different TEI P5 solution was chosen for the DTABf.[22]

34 We are currently working on a way to document the DTABf level of a text within its TEI header. As a consequence, the annotation level can be used as a criterion for dividing the DTA corpus into subsets of variable annotation depth. The corpus query tool DDC[23] allows for element-based queries of the DTA corpus. Future plans include the combination of those conditions so that queries

within DTABf-conformant sub-corpora, which are created based on level as well as on the elements represented by them, are possible via the DTA website. Similar search options are to be offered by CLARIN's Federated Content Search.[24]

## 3.5 Transformation of DTABf Texts into Other Formats

### 3.5.1 Presentation of DTABf Texts

35 The high granularity of the DTABf tag set and its coherent application to the DTA text corpora enable us to create stylesheets for the transformation of the TEI/XML documents into many other formats, including HTML, plain text with basic structural information, or ePUB. These stylesheets in turn are able to deal with the whole range of tagging scenarios the DTABf allows. In fact, the possibility of uniform presentations of all DTABf texts was one specific goal for the design of the DTABf, and is continually considered when making adjustments to it.

36 In most cases semantic annotations can be represented on the presentation level as long as they are unambiguous, consistent, and well-documented. For instance, division titles, list items, speakers, and stage directions in a drama can without any difficulty be presented in such a way that users of the respective reading versions are able to recognize these textual and structural phenomena at a glance (e.g., in our case titles are presented as bold text of larger size; stage directions are printed in italics; list items are outdented; tables are rendered as tables with dividing rules between rows and cells). Also, the specifications of the DTABf down to the attribute value level support the presentation of contents in underspecified elements. For example, the values `"foot"`, `"end"`, `"left"`, and `"right"` within the `@place` attribute of the element `<note>` define the position of a note in the source text. In addition, since the presentation of DTA texts is meant to come as close to the original layout as possible, the DTABf also includes tagging solutions for pure layout information (such as centered text, italics, changes of fonts, and boldface), which additionally support the presentation.

37 However, sometimes the semantics of structures interfere with the presentation, which should ideally approximate the layout of the source text. An example of this semantic interference is found in the tagging of title pages. The TEI definition of the `<titlePage>` element is quite limited in that it restricts the usage of this element to complete pages.[25] There are cases, though, that are not met by this definition where the `<titlePage>` element would still be reasonable. For example,

usually the heading on the first page of a newspaper edition contains bibliographic information about the edition, but does not span an entire page. We therefore considered refraining from using the <titlePage> element in newspapers and instead using <docTitle> and analogous elements for the tagging of title information in newspapers. This solution would be fully TEI-conformant. However, it would lead to different encodings of semantically similar structures (in both cases we want to encode title information of a document as given in the source text; the layout is only of secondary interest here). Furthermore, the presentation of information on title pages within the DTA corpora is based on the possibility to define the <titlePage> element as block element and hence to homogeneously render all text occurring within <titlePage> as centered, etc. So, if we left out <titlePage> in newspapers, we would lose the ability to present title information in this text type (newspapers) similarly to title information in all other text types of the DTA corpus. We therefore decided to use <titlePage> for newspapers as well, but introduced the attribute-value pair @type="heading" to the DTABf to differentiate the area of title information in newspapers from title pages in the narrower sense.

38 The fixed vocabulary of the DTABf permits several coherent presentations of DTABf texts to users who can set the parameters of their preferred text presentation themselves.[26]

39 We did not make use of the TEI standard stylesheets for two reasons. The first reason is pragmatic. The TEI stylesheet library[27] is a very large project on its own, consisting of complex and deeply structured XSL files which try to cover most of the TEI tag set (even though some elements like <cb> are completely ignored at least by the TEI to HTML conversion stylesheets). In addition, the standard stylesheets are too generic for an adequate visual representation of historical texts in the DTA context where we attempt to achieve a presentation as close as possible to the original print sources. We distinguish some elements with regard to their presentation depending on the context of their appearance. As an example, the <p> element describes a paragraph in prose text, but within <sp> (speech in a performance text) the same element denotes a speaker's utterance (which may not be represented as a common paragraph in the original printing). Our own stylesheet library is complemented by an extensive test suite. The stylesheets are available at https://github.com/haoess/dta-tools.

### 3.5.2 Conversion into other Common Formats

40 The DTABf TEI header can be converted automatically into other common metadata formats. Currently, the DTA provides a Dublin Core and a Component Metadata Infrastructure (CMDI)[28] version of all metadata records which can be harvested via the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).[29] In this way, DTA metadata can be reused by other platforms and Online Public Access Catalogues (OPACs). In addition, the CMDI metadata records for the DTA corpus texts can be interpreted by the Virtual Language Observatory (VLO) of CLARIN[30] and thus become visible together with other language resources within the CLARIN infrastructure.

41 The plain text versions of all DTA documents together with their corresponding linguistic annotations are provided as TCF files. The Text Corpus Format (TCF)[31] is a standoff format for linguistic annotation and forms the input format for CLARIN's WebLicht[32] infrastructure. Thus, it becomes possible to further analyze DTA texts using linguistic tools provided within WebLicht.

# 4. Dissemination of the DTABf

## 4.1 Documentation

42 All decisions concerning the DTABf are described in comprehensive documentation which explains the treatment of common structuring requirements as well as special cases and illustrates each phenomenon with examples from the corpus.[33] The DTABf documentation is thematically subdivided into different sections (formal/semantic annotation, metadata annotation, special encoding of certain text types such as journals and newspapers). Not only does it explain formal customizations of the TEI tag set as realized within the ODD and schema as well as the Schematron rule set, but it also specifies transcription guidelines as well as rules which could not be formalized without changing the content models of the TEI tag set and thus going beyond the DTABf schema.

43 The prose documentation is supplemented by tables containing all DTABf elements, attributes, and attribute values for text encoding on the one hand[34] and for metadata annotation on the other.[35]

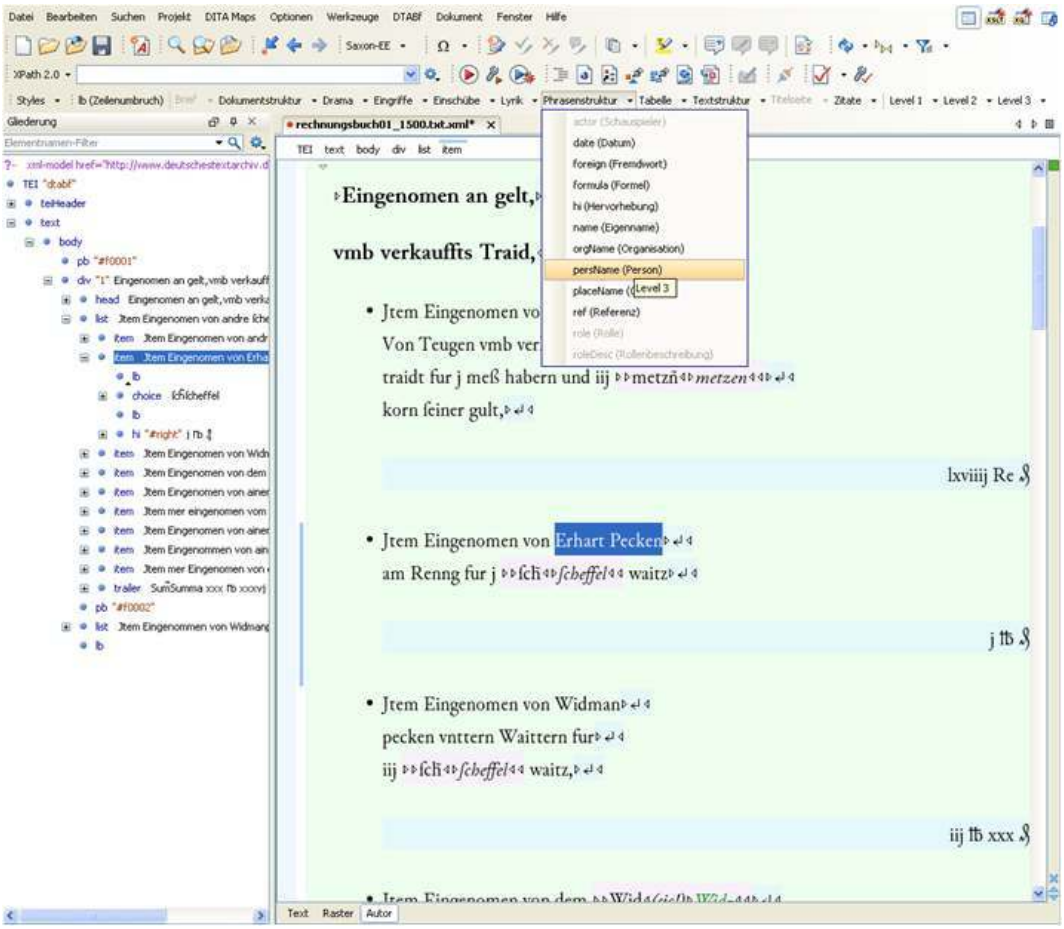
44 The DTABf documentation provides a description of the work completed so far, but more importantly it serves as a guideline for users working with the DTABf for the TEI-compliant structuring of historical texts.

## 4.2 Training and Tools for Users

45 The existence of comprehensive documentation is a necessary prerequisite for the usability of the DTABf and thus for its acceptance by a larger user community. In addition, the DTA offers workshops and tutorials where users learn to apply the DTABf in a consistent way.

46 Furthermore, work on text editions according to the DTABf is supported by the DTA oXygen Framework DTAoX, a framework for the Author mode of the oXygen XML Editor. DTAoX provides an ad hoc, WYSIWYG-like visualization of DTABf-compliant annotations in tagged text passages, of the annotation levels they belong to, and of conflicts with the DTABf. Additionally, to support DTABf-compliant text annotation, elements can easily be inserted in the text using the DTAoX toolbar, which groups elements according to semantic categories and annotation levels.[36]

**Figure 1. DTAoX oXygen XML Editor framework, with toolbar for DTABf elements and color scheme for different annotation levels (green: level 1; blue: level 2; purple: level 3).**



47 All DTABf-compliant texts added to the DTA corpus are integrated into the quality assurance platform DTAQ[37] where their transcriptions and annotations may be proofread.

# 5. The DTABf at Different Stages of Digitization

## 5.1 Digitization by Way of Transcription – DTABf-born Texts

48 Homogeneity within a corpus does not only concern its tagging but also crucially depends on the way transcriptions are realized. Therefore, the DTABf is complemented by extensive transcription guidelines which can be referred to and further specified within the DTABf-compliant TEI header.[38] These guidelines are intended for transcriptions which are as close to the source text as possible, the guiding idea being that the historical language and writing should be preserved. Normalizations, if at all needed, can often be conducted on the presentation level rather than within the primary transcriptions where they might adulterate the results of research on historical spelling, printing habits, historical abbreviations, and the like. Wherever possible, special characters are transcribed as such, using the corresponding Unicode code points. Generally, editorial interventions are marked, and the conditions of the source text are carefully documented as well.

## 5.2 Interchange of TEI Documents

49 The history of the DTABf coincides with the history of documents which are exchanged between collaborating projects and the DTA.

50 An example of such a collaboration may be found in the project *Johann Friedrich Blumenbach—online*.[39] This project is working on a digital edition of Blumenbach's printed works in German and Latin as well as his handwritten texts. All texts are prepared in a TEI P5 format.

51 The DTA integrates the digitized full texts of German monographs and selected journal articles of Blumenbach into its platforms. This process is not unidirectional but involves several processing steps in the course of which the digital documents are exchanged between and enriched by the two projects.

52 First, the texts are transcribed by the Blumenbach project and annotated according to the Blumenbach project's TEI P5 format. The resulting TEI P5-compliant documents are automatically converted into the DTABf, analyzed by the linguistic tools of the DTA (including tokenization, lemmatization, POS tagging, normalization of historical spellings), and integrated into DTAQ where they can be proofread and corrected. At this stage, the texts are already made available to the interested community; that is, they can be read online, downloaded in different formats (such as HTML or plain text versions both derived from the DTABf-XML or the TCF version containing the corresponding linguistic information), and queried within the DTA.

53 After the completion of the first correction phase within DTAQ, the DTABf-compliant texts are further processed by the DTA's tools for automatic Named-Entity Recognition (NER). The documents are then manually corrected by the Blumenbach project in light of the NER results, and the corrected texts are again returned to the DTA where they are re-integrated into DTAQ, published on the DTA website, and integrated into the CLARIN-D infrastructure.

54 The TEI P5 formats used by both projects (DTABf and Blumenbach's TEI P5 format) agree semantically, so documents can be converted automatically between the two formats in the course of the exchanges described above.

55 This workflow is a good example of interoperability and interchange of TEI documents, while it also shows that manual efforts remain necessary to create TEI formats compatible with one another.

## 6. Life History of the DTABf

56 Although the text corpus upon which the DTABf is built is already very large, new structural phenomena may still be encountered, mainly because of individual printing habits, especially in earlier works. In addition, the structuring of external texts integrated into the DTA corpus may differ from the DTABf formally (if varying TEI solutions were chosen for similar phenomena) or semantically (if tagging not provided by the DTABf was applied). Therefore, the DTABf continually comes under scrutiny. The main challenges are twofold: (1) to decide whether adaptations to the format are unavoidable in order to meet new requirements, and (2) to ensure that any such adaptations do not lead to inconsistencies in the structural markup within the corpus.

## 6.1 Phenomena within the Scope of the DTABf

### 6.1.1 Treating New and Known Phenomena with Established Solutions

57 By this point, the DTABf has achieved such comprehensive coverage of historical texts that substantial changes are quite unusual. Novelties do not necessarily lead to changes in the DTABf. Often the DTABf already offers solutions which may also be applied to newly encountered phenomena.

58 For example, the DTABf provides solutions for annotating quotations and corresponding bibliographic references as well as for concatenating discontinuous text passages. The example below shows discontinuous citations where the quotation is presented inline, whereas the bibliographic citation occurs beforehand within a marginal note. Though it was new to us, we were able to handle this scenario without any extensions to the existing DTABf.

Figure 2. Discontinuous Parts of Quotations.

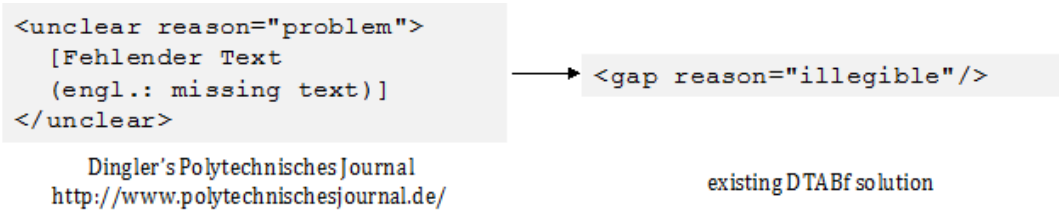

```
nicht ſagen/ das Chriſtus allein im Himmel vnd nicht bey vns<lb/>
  <cit xml:id="bibl9" next="#quote12">
    <bibl>
      <note place="left">Pſal. 139.</note>
    </bibl>
  </cit>
auff Erden ſey/ ſintemahl er ſitzet zu der Rechten der krafft Got-<lb/>
tes/ welche ſich nicht theilen leſt/ ſondern allenthalben iſt/ vnd alles<lb/>
erfüllet. Deñ alſo ſtehet geſchrieben im 139 Pſalm.
<cit xml:id="quote12" prev="#bibl9">
  <quote>Wo<lb/>
    ſoll ich hingehn für deinem Geiſt/ vnd wo ſol ich hinfliehen für<lb/>
    deinem Angeſicht? Fuhre ich gehn Himmel/ ſo biſtu da/ bettet<lb/>
    ich mir in die Helle/ ſiehe ſo biſtu auch da/ Nehme ich flügel der<lb/>
    Morgenröte/ vnd bliebe am euſſerſten Meer/ ſo würde mich doch<lb/>
    deine Handt daſelbſt führen/ vnd deine Rechte mich halten:</quote>
</cit>
```

Lucas Bacmeister, Leichpredigt ... Tessen Von Parsow (Rostock, 1614), facsimile 18.[40]

### 6.1.2 Converting New Solutions into Established Solutions

59 TEI tagging solutions which were applied to external texts might be compliant with the DTABf in their thematic scope, though encoded using different TEI vocabulary. In such cases the original tagging can easily be replaced with the DTABf solution.

Figure 3. Tagging of Text Loss.




## 6.2 Changes to the DTABf

60 However, there are cases where new requirements cannot be handled within the DTABf and changes to the format become necessary. Possible scenarios are the following:

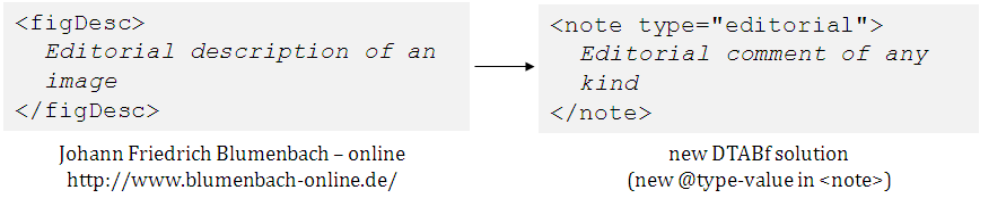
1. New texts may contain structures which are new to the DTABf, for instance, within new document types such as newspapers or manuscripts.

   **Example 5: Newspapers**

   The DTA core corpus consists of texts from various disciplines, text types, and genres to allow for insights about the New High German language as it was used in different contexts and discourses at different points of time in its history. However, an important text type—which because of its enormous extent has not been included in the DTA corpora—is newspapers. We are currently extending the DTA corpus by adding historical newspapers in the course of different project partnerships (see Haaf and Schulz 2014). Newspaper texts from external projects are converted into the DTABf. We adjusted the DTABf to allow for the tagging of structures which are significant for or even limited to the text type `"newspaper"`. Most importantly, we added division types for text passages which are typical for newspapers like political news (`@type="jPoliticalNews"`), weather reports (`@type="jWeatherReports"`), financial news (`@type="jFinancialNews"`), the feuilleton (`@type="jFeuilleton"`), and articles within those named categories (`@type="jArticle"`). Specifics of the DTABf for newspapers are covered by separate documentation.[41]

2. Structuring applied within texts from external projects (e.g., editorial comments) might not yet have an equivalent within the DTABf. Therefore, new components (here, a new attribute-value-pair `@type="editorial"`) are introduced.

**Figure 4. Editorial Comments.**

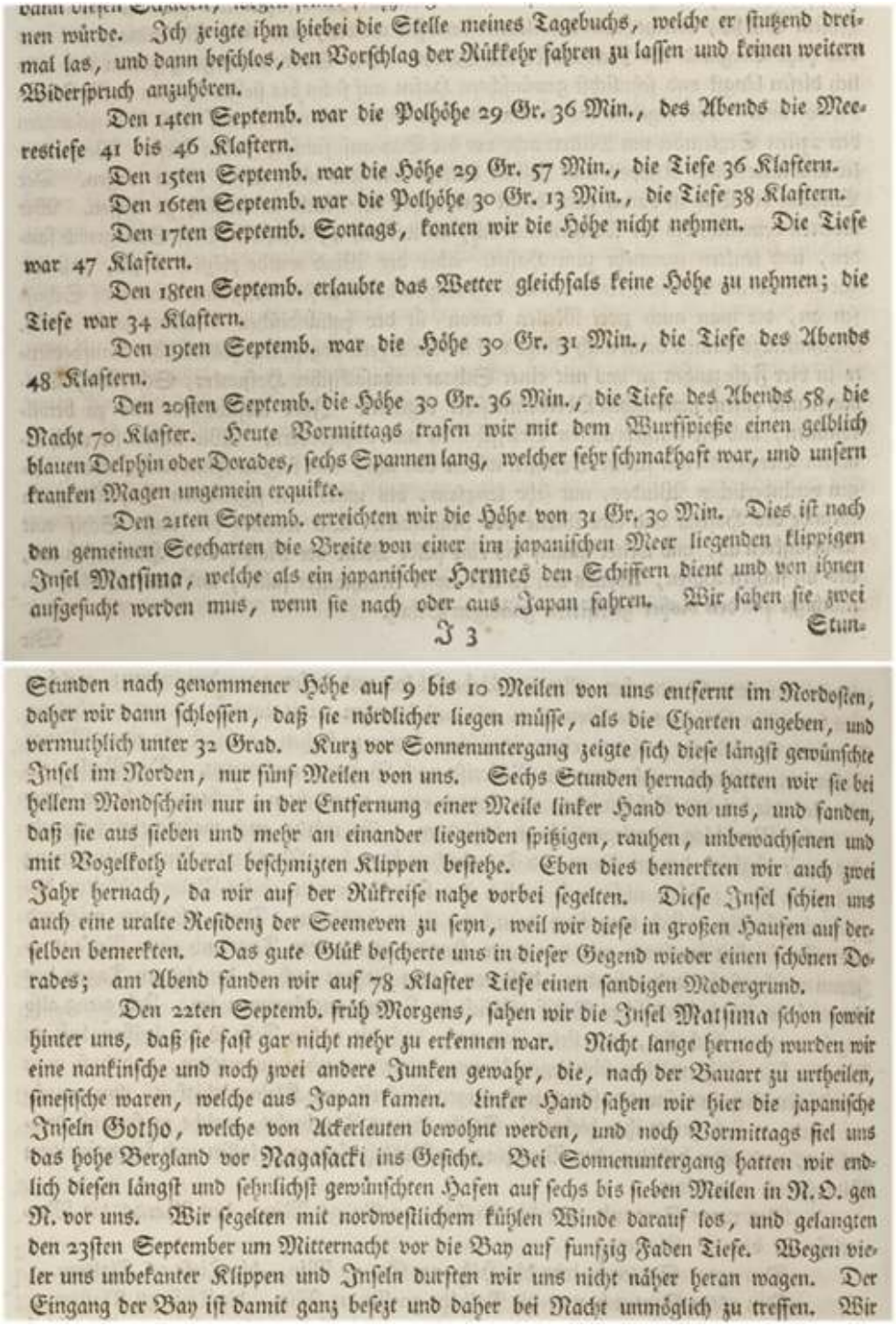



Johann Friedrich Blumenbach – online
http://www.blumenbach-online.de/

new DTABf solution
(new @type-value in <note>)

3. The documentation might lack precise examples, leading to uncertainties about the applied text structuring.

Figure 5. List Items or Paragraphs?



dann dieſen Capitein, [illegible]
nen würde. Ich zeigte ihm hiebei die Stelle meines Tagebuchs, welche er ſtuzend drei-
mal las, und dann beſchlos, den Vorſchlag der Rükkehr fahren zu laſſen und keinen weitern
Widerſpruch anzuhören.

Den 14ten Septemb. war die Polhöhe 29 Gr. 36 Min., des Abends die Mee-
restiefe 41 bis 46 Klaftern.

Den 15ten Septemb. war die Höhe 29 Gr. 57 Min., die Tiefe 36 Klaftern.

Den 16ten Septemb. war die Polhöhe 30 Gr. 13 Min., die Tiefe 38 Klaftern.

Den 17ten Septemb. Sontags, konten wir die Höhe nicht nehmen. Die Tiefe
war 47 Klaftern.

Den 18ten Septemb. erlaubte das Wetter gleichfals keine Höhe zu nehmen; die
Tiefe war 34 Klaftern.

Den 19ten Septemb. war die Höhe 30 Gr. 31 Min., die Tiefe des Abends
48 Klaftern.

Den 20ſten Septemb. die Höhe 30 Gr. 36 Min., die Tiefe des Abends 58, die
Nacht 70 Klafter. Heute Vormittags trafen wir mit dem Wurfſpieße einen gelblich
blauen Delphin oder Dorades, ſechs Spannen lang, welcher ſehr ſchmakhaft war, und unſern
kranken Magen ungemein erquikte.

Den 21ten Septemb. erreichten wir die Höhe von 31 Gr. 30 Min. Dies iſt nach
den gemeinen Seecharten die Breite von einer im japaniſchen Meer liegenden klippigen
Inſel **Matſima**, welche als ein japaniſcher **Hermes** den Schiffern dient und von ihnen
aufgeſucht werden mus, wenn ſie nach oder aus Japan fahren. Wir ſahen ſie zwei

J 3 Stun-

Stunden nach genommener Höhe auf 9 bis 10 Meilen von uns entfernt im Nordoſten,
daher wir dann ſchloſſen, daß ſie nördlicher liegen müſſe, als die Charten angeben, und
vermuthlich unter 32 Grad. Kurz vor Sonnenuntergang zeigte ſich dieſe längſt gewünſchte
Inſel im Norden, nur fünf Meilen von uns. Sechs Stunden hernach hatten wir ſie bei
hellem Mondſchein nur in der Entfernung einer Meile linker Hand von uns, und fanden,
daß ſie aus ſieben und mehr an einander liegenden ſpißigen, rauhen, unbewachſenen und
mit Vogelkoth überal beſchmizten Klippen beſtehe. Eben dies bemerkten wir auch zwei
Jahr hernach, da wir auf der Rükreiſe nahe vorbei ſegelten. Dieſe Inſel ſchien uns
auch eine uralte Reſidenz der Seemeven zu ſeyn, weil wir dieſe in großen Haufen auf der-
ſelben bemerkten. Das gute Glük beſcherte uns in dieſer Gegend wieder einen ſchönen Do-
rades; am Abend fanden wir auf 78 Klafter Tiefe einen ſandigen Medergrund.

Den 22ten Septemb. früh Morgens, ſahen wir die Inſel **Matſima** ſchon ſoweit
hinter uns, daß ſie faſt gar nicht mehr zu erkennen war. Nicht lange hernach wurden wir
eine nankinſche und noch zwei andere Junken gewahr, die, nach der Bauart zu urtheilen,
ſineſiſche waren, welche aus Japan kamen. Linker Hand ſahen wir hier die japaniſche
Inſeln **Gotho**, welche von Ackerleuten bewohnt werden, und noch Vormittags fiel uns
das hohe Bergland vor **Nagaſacki** ins Geſicht. Bei Sonnenuntergang hatten wir end-
lich dieſen längſt und ſehnlichſt gewünſchten Hafen auf ſechs bis ſieben Meilen in N. O. gen
N. vor uns. Wir ſegelten mit nordweſtlichem kühlen Winde darauf los, und gelangten
den 23ſten September um Mitternacht vor die Bay auf funfzig Faden Tiefe. Wegen vie-
ler uns unbekanter Klippen und Inſeln durften wir uns nicht näher heran wagen. Der
Eingang der Bay iſt damit ganz beſezt und daher bei Nacht unmöglich zu treffen. Wir

**Engelbert Kaempfer, Geschichte und Beschreibung von Japan, ed. Christian Wilhelm von Dohm, vol. 1 (Lemgo, 1777), pp. 69–70.**[42]

4. New TEI P5 releases may introduce changes to `tei_all` which affect the DTABf.

**Figure 6.** `@unit` **vs.** `@type` **within** `<biblScope>` **(Release 2.3.0).**

```
<biblScope type="pages">
  volume number of the
  publication within a series
</biblScope>
```

DTABf before release 2.3.0 of TEI/P5

```
<biblScope unit="pages">
  volume number of the
  publication within a series
</biblScope>
```

DTABf after release 2.3.0 of TEI/P5

61 Changes to the DTABf are carried out only if they are consistent with the existing tag set and do not introduce ambiguities to the format. The changes mainly concern attributes or values and only rarely TEI elements or modules.

## 7. Conclusion and Further Prospects

62 In this paper we described the DTABf as a "living" TEI format for the annotation of historical written texts for the creation of large reference corpora. The DTA corpus base is still growing either through digitization carried out by the DTA team or through the addition of text collections originating from external cooperating projects. Therefore the DTABf is not static but is constantly checked and adjusted to new structural phenomena.

63 Future work will involve the adaptation of the DTABf to manuscripts. Currently, the DTA corpora almost exclusively contain printed works; only a couple of manuscripts have been integrated so far for evaluation purposes.[43] However, some important text types usually exist in handwritten form rather than as printed documents (e.g., letters, diaries, and financial records). In order to improve the balance of the DTA corpus, it would therefore be interesting to integrate manuscripts from some of these widespread text types into our collection. Our tests with manuscripts showed that most of the structural phenomena which occur in manuscripts are similar to those in printed texts and hence can already be treated within the DTABf. However, there are some additional characteristics of handwritten texts which might be usefully encoded (e.g., ad hoc additions, deletions, or insertions of the writer, or the change of hands within one document). These additional phenomena will necessitate adaptations to the DTABf. Furthermore, they are likely to

affect the presentation of the TEI transcriptions—we might not be able to imitate the manuscript facsimile on the presentation level to the same extent as we do for printed text sources. In addition, some aspects of the metadata needed for manuscripts differ from the metadata needed for print documents. The adaptations to the DTABf and its corresponding tools and services necessary for the integration of manuscripts into the DTA will be performed in a document-based way through the integration of further manuscripts into the DTA corpus.

64 Other adaptations of the DTABf will be necessary in order to integrate texts obtained via Optical Character Recognition (OCR). OCR software only recognizes basic (text) zones which eventually have to be mapped to semantically meaningful structures. Semi-automatic subsequent structuring of OCR texts is possible, but becomes increasingly complex and error-prone with greater sophistication, detail, and granularity in the target markup. Therefore, based on experiences with automatic post-structuring of OCR texts in the course of the DFG-funded project *Die Grenzboten*,[44] we are planning to create an additional basic structuring level for OCR texts within the DTABf, onto which semi-automatic text structuring can be implemented, and upon which further manual post-structuring can be performed.

65 As the best practice format for the structural annotation of historical printed texts within CLARIN-D,[45] the DTABf has been brought in line with the CLARIN infrastructure. In this context, the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), as one of the CLARIN-D service centers,[46] provides tools and services which allow for the integration of DTABf texts into the CLARIN infrastructure where they are preserved and made available for the long term. In addition, routines for the conversion of DTABf texts and metadata into the CLARIN formats TCF and CMDI have been developed, so that re-use of DTABf texts within the CLARIN infrastructure is possible.

## BIBLIOGRAPHY

Burnard, Lou, and C. M. Sperberg-McQueen. 2012. "TEI Lite: Encoding for Interchange: An Introduction to the TEI." Final revised edition for TEI P5, August. Accessed February 7 2014. http://www.tei-c.org/Guidelines/Customization/Lite/.

Geyken, Alexander. 2007. "The DWDS Corpus: A Reference Corpus for the German Language of the 20th Century." In *Idioms and Collocations: Corpus-Based Linguistic and Lexicographic Studies*, edited by Christiane Fellbaum, 23–41. London: Continuum.

Geyken, Alexander, Susanne Haaf, and Frank Wiegand. 2012. "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, 383–91. Schriftenreihe der Österreichischen Gesellschaft für Artificial Intelligence 5. Wien: ÖGAI. Accessed February 7 2014. http://www.oegai.at/konvens2012/proceedings/57_geyken12w/.

Haaf, Susanne, and Matthias Schulz. 2014. "Historical Newspapers & Journals for the DTA." In *Language Resources and Technologies for Processing and Linking Historical Documents and Archives – Deploying Linked Open Data in Cultural Heritage – LRT4HDA. Proceedings of the workshop*, held at the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 26–31, 2014, Reykjavik (Iceland), 50–54. Accessed February 7 2014. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRT4HDA%20Proceedings.pdf#page=57.

Jurish, Bryan. 2012. "Finite-state Canonicalization Techniques for Historical German." PhD diss., Universität Potsdam. Accessed February 7 2014. http://opus.kobv.de/ubp/volltexte/2012/5578/.

Jurish, Bryan, and Kay-Michael Würzner. 2013. "Word and Sentence Tokenization with Hidden Markov Models." *Journal for Language Technology and Computational Linguistics* 28, no. 2: 61–83. Accessed February 7 2014. http://www.jlcl.org/2013_Heft2/3Jurish.pdf.

Jurish, Brian, Christian Thomas, and Frank Wiegand 2014. "Querying the deutsches textarchiv." *CEUR Workshop Proceedings*. 1131: 25–30. Accessed February 7 2014. http://ceur-ws.org/Vol-1131/mindthegap14_7.pdf.

Lüngen, Harald, and C. M. Sperberg-McQueen. 2012. "A TEI P5 Document Grammar for the IDS Text Model." *Journal of the Text Encoding Initiative* 3. Accessed February 7 2014. http://jtei.revues.org/508. doi:10.4000/jtei.508.

TEI Consortium. 2014. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 2.6.0. Last updated January 20. Accessed February 7 2014. http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/.

TEI SIG on Libraries. 2011. "Best Practices for TEI in Libraries: A TEI Project." Edited by Kevin Hawkins, Michelle Dalmau, and Syd Bauman. Version 3.0. October. Accessed February 7 2014. http://www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html.

TextGrid. 2007–2009. "TextGrid's Baseline Encoding for Text Data in TEI P5." Accessed February 7 2014. http://www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf.

Thomas, Christian, and Frank Wiegand. 2012. "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 Historical Corpora 2012 International Conference, Goethe University, Frankfurt, Germany, December 6–9. Accessed February 7 2014. http://edoc.bbaw.de/volltexte/2012/2308/.

Trolard, Perry. 2011. "TEI Tite: A Recommendation for Off-site Text Encoding." Version 1.1. September. Accessed February 7 2014. http://www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html.

Unsworth, John. 2011. "Computational Work with Very Large Text Collections: Interoperability, Sustainability, and the TEI." In *Journal of the Text Encoding Initiative* 1 (June). Accessed February 7 2014. http://jtei.revues.org/215. doi:10.4000/jtei.215.

Pytlik Zillig, Brian. 2009. "TEI Analytics: Converting Documents into a TEI Format for Cross-collection Text Analysis." In *Literary and Linguistic Computing* 24, no. 2: 187–92. doi:10.1093/llc/fqp005.

## APPENDIXES

# Appendix 1. DTABf Elements of Level 1 within `<text>` with Frequencies, Attributes, and Values

| Element | Frequency* | Attributes | Values |
|---|---|---|---|
| `<p>` | 1,225,251 | | |
| `<l>` | 811,758 | `@n` | `"[verse number (in the textsource)]"` |
| `<pb>` | 558,763 | `@n` | `"[page number]"` |
| | | `@facs` | `"[link to facsimile]"` |
| `<item>` | 456,195 | | |
| `<note>` | 368,227 | `@n` | `"[note sign/number]"` |
| | | `@place` | `"left", "right", "foot", "end"` |
| | | `@type` | `"editorial"` |
| `<head>` | 247,030 | | |

| | | | |
|---|---|---|---|
|  | 223,885 | @n | "[depth of text structure]" |
| | | @type | "abbreviations", "act", "advertisement", "appendix", "bibliography", "chapter", "contents", "copyright", "corrigenda", "dedication", "diaryEntry", "edition", "figures", "frontispiece", "imprimatur", "imprint", "index", "letter", "lexiconEntry" "poem", "postface", "preface", "recipe", "scene" |
| <lg> | 117,431 | @n | "[(for stanzas:) number of stanza]" |
| | | @type | "poem" |
| <cell> | 65,179 | @cols | "[number of colums occupied]" |
| | | @rows | "[number of rows occupied]" |
| | | @role | "label" |
| <cb> | 63,521 | @n | "[column number]" |
| | | @type | "start", "end" |
| <list> | 46,989 | | |
| <figure> | 31,731 | @type | "notatedMusic" |
| | | @facs | "[URL pointing to an image of the tagged figure]" |
| <formula> | 30,172 | @notation | "MathML", "TeX" |

| | | | |
|---|---|---|---|
| | | @facs | "[pointer to a graphic of the transcribed formula]" |
| <row> | 21,749 | | |
| <gap> | 21,251 | @reason | "fm", "illegible", "insignificant", "lost" |
| | | @unit | "chars", "words", "lines", "pages" |
| | | @quantity | "[amount of text missing]" |
| <supplied> | 19,292 | | |
| <table> | 6,426 | | |
| <titlePart> | 5,324 | @type | "copyright", "dedication", "desc", "main", "price", "series", "sub", "volume" |
| <body> | 2,621 | | |
| <titlePage> | 2,418 | @type | "halftitle", "heading", "main", "series" |
| <text> | 2,110 | | |
| <front> | 2,073 | | |
| <back> | 1,054 | | |

* Frequency of occurrence within the entire DTA corpus (> 2,100 works) in May 2014.

## Appendix 2. DTABf Elements of Level 2 within <text> with Frequencies, Attributes, and Values

| Element | Frequency* | Attributes | Values |
|---|---|---|---|

| | | | |
|---|---|---|---|
| <lb> | 15,728,726 | @n | "[printed line number]" |
| <hi> | 4,188,984 | @rendition | "[the medium for highlighting]" |
| <fw> | 554,541 | @place | "bottom", "top" |
| | | @type | "catch", "header", "pageNum", "sig" |
| <milestone> | 77,873 | @unit | "section" |
| | | @rendition | "#hr", "#hrBlue", "#hrRed" |
| <sp> | 61,057 | @who | "[speaker’s id (as assigned to the role)]" |
| <speaker> | 60,970 | | |
| <space> | 46,401 | @dim | "horizontal", "vertical" |
| | | @unit | "[amount of space concerned]" |
| | | @quantity | "[unit measuring the amount of space concerned]" |
| <stage> | 20,470 | | |
| <bibl> | 11,062 | | |
| <quote> | 10,384 | @type | "translation" |
| <cit> | 9,847 | | |
| <salute> | 5,470 | | |
| <argument> | 4,978 | | |
| <dateline> | 4,503 | | |

| | | | |
|---|---|---|---|
| <closer> | 3,930 | | |
| <docTitle> | 2,374 | | |
| <docImprint> | 2,127 | | |
| <byline> | 2,105 | | |
| <pubPlace> | 1,963 | | |
| <publisher> | 1,944 | | |
| <docDate> | 1,903 | | |
| <docAuthor> | 1,799 | | |
| <castItem> | 1,038 | | |
| <role> | 993 | | |
| <opener> | 745 | | |
| <floatingText> | 514 | | |
| <signed> | 340 | | |
| <epigraph> | 318 | | |
| <postscript> | 238 | | |
| <roleDesc> | 144 | | |
| <imprimatur> | 111 | | |
| <castList> | 97 | | |
| <castGroup> | 64 | | |
| <docEdition> | 52 | | |
| <spGrp> | 36 | | |
| <actor> | 12 | | |

* Frequency of occurrence within the entire DTA corpus (> 2,100 works) in May 2014.

## NOTES

**1** BNC, http://www.natcorp.ox.ac.uk/.

**2** For a discussion of this issue see Geyken, Haaf, and Wiegand 2012, 383ff.

**3** See TEI Consortium 2014, 23.3, "Using the TEI: Personalization and Customization", http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/USE.html#MD: "[T]he TEI scheme supports a variety of different approaches to solving similar problems, and also defines a much richer set of elements than is likely to be necessary in any given project.... For these reasons, it is almost impossible to use the TEI scheme without customizing or personalizing it in some way."

**4** CLARIN: Common Language Resources and Technology Infrastructure, https://www.clarin.eu/.

**5** The DTA (http://www.deutschestextarchiv.de) at the Berlin-Brandenburg Academy of Sciences and Humanities has been funded by the Deutsche Forschungsgemeinschaft (German Research Foundation, DFG) since 2007.

**6** See DTA-Erweiterungen (DTA Extensions), http://www.deutschestextarchiv.de/dtae.

**7** ODD: "One document does it all": see TEI Consortium 2014, "Customization: Getting Started with P5 ODDs," http://www.tei-c.org/Guidelines/Customization/odds.xml; TEI Consortium 2014, 23.3, "Personalization and Customization", http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/USE.html#MD.

**8** "Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-D-Infrastruktur: Kurationsprojekt 1 der Facharbeitsgruppe 1 Deutsche Philologie," http://www.deutschestextarchiv.de/doku/clarin_kupro_index; see also Thomas and Wiegand 2012.

**9** See section 6.2 below for more information about extensions and modifications to the DTABf.

**10** See the documentation for the DTABf-compliant recording of metadata, "Dokumente zum DTA-Basisformat: DTA-Basisformat—Header," http://www.deutschestextarchiv.de/doku/basisformat_header.

**11** See the DTABf documentation for the formal and semantic text annotation: "Dokumente zum DTA-Basisformat: Formale Erschließung des Volltextes," http://www.deutschestextarchiv.de/doku/basisformat_texterschliessung_formal; "Dokumente zum DTA-Basisformat: Inhaltliche Erschließung des Volltextes," http://www.deutschestextarchiv.de/doku/basisformat_texterschliessung_inhaltlich.

**12** DTA-Tokwrap: "Software im Deutschen Textarchiv: 6. Tokenizer für komplexe Texte (DTA-Tokwrap)," http://www.deutschestextarchiv.de/doku/software#dtatw; Jurish and Würzner 2013. CAB: "Software im Deutschen Textarchiv: 8. Linguistische Analyse historischer Texte (CAB)," http://www.deutschestextarchiv.de/doku/software#cab; Jurish 2012; DDC: "Software im Deutschen Textarchiv: 7. Indizierung und linguistische Suche (DDC)," http://www.deutschestextarchiv.de/doku/software#ddc; Jurish, Thomas, and Wiegand 2014.

**13** A "Tool Chaining Format used by WebLicht," http://www.clarin.eu/category/glossary/tcf.

**14** See http://www.deutschestextarchiv.de/basisformat.odd.

**15** See http://www.deutschestextarchiv.de/basisformat.rng.

**16** See http://www.deutschestextarchiv.de/doku/basisformat.

**17** All DTA texts as well as the DTA corpora as a whole are available for download and are freely accessible on the DTA website, http://www.deutschestextarchiv.de.

**18** The full DTABf RNG is available for download at http://www.deutschestextarchiv.de/basisformat.rng, the reduced schema is accessible under: http://www.deutschestextarchiv.de/basisformat_ohne_header.rng.

**19** See http://www.deutschestextarchiv.de/basisformat.sch.

**20** However, it is possible to embed Schematron rules in the ODD file by using a different namespace.

**21** See TEI Consortium 2014, 15.5, "Recommendations for the Encoding of Large Corpora",: http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/CC.html#CCREC.

**22** Ibid. For a description of the DTABf levels, see http://www.deutschestextarchiv.de/doku/basisformat_table?lang=en.

**23** DDC: Dialing DWDS Concordancer; see "Software im Deutschen Textarchiv: 7. Indizierung und linguistische Suche (DDC)," http://www.deutschestextarchiv.de/doku/software#ddc.

**24** http://www.clarin.eu/content/federated-content-search.

**25** See TEI Consortium 2014, element specification of `<titlePage>`: "(title page) contains the title page of a text, appearing within the front or back matter," http://www.tei-c.org/Vault/P5/2.6.0/doc/tei-p5-doc/en/html/ref-titlePage.html.

**26** See the customizable HTML presentation which can be accessed from the starting page of each book on the DTA website.

**27** TEI XSL Stylesheets, https://github.com/TEIC/Stylesheets/.

**28** See CLARIN, "Component Metadata," http://www.clarin.eu/content/component-metadata. For the CMDI profile of the DTA, see http://catalog.clarin.eu/ds/ComponentRegistry/?item=clarin.eu:cr1:p_1345180279115; for a list of differences between the DTABf TEI header and DTA's CMDI profile, see http://www.deutschestextarchiv.de/doku/cmdi.

**29** See "DTA—APIs," http://www.deutschestextarchiv.de/api.

**30** See http://catalog.clarin.eu/vlo/?theme=CLARIN-D.

**31** See WebLicht, "The TCF Format," http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format.

**32** See the WebLicht main page, http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

**33** See "DTA-Basisformat—Einführung," http://www.deutschestextarchiv.de/doku/basisformat.

**34** English version: "DTA 'base format'—overview (elements within text)," http://www.deutschestextarchiv.de/doku/basisformat_table?lang=en; German version: "DTA-Basisformat—Überblick (Elemente innerhalb von text)," http://www.deutschestextarchiv.de/doku/basisformat_table?lang=de.

**35** English version: "DTA 'base format'—overview (elements within teiHeader)," http://www.deutschestextarchiv.de/doku/basisformat_table_header?lang=en; German version: "DTA-Basisformat—Überblick (Elemente innerhalb von teiHeader)," http://www.deutschestextarchiv.de/doku/basisformat_table_header?lang=de.

**36** DTAoX is available for download at http://www.deutschestextarchiv.de/doku/software#dtaox.

**37** Deutsches Textarchiv—Qualitätssicherung; see "Kollaborative Qualitätssicherung im Deutschen Textarchiv," http://www.deutschestextarchiv.de/dtaq/about.

**38** For the documentation of the DTA transcription guidelines, see "DTA-Richtlinien zur Texterfassung," http://www.deutschestextarchiv.de/doku/richtlinien.

**39** http://www.blumenbach-online.de/.

**40** Deutsches Textarchiv Qualitätssicherung, http://www.deutschestextarchiv.de/dtaq/book/view/bacmeister_predigt_1614?p=18.

**41** "DTA-Basisformat—Auszeichnung von Zeitungen," http://www.deutschestextarchiv.de/doku/basisformat_zeitungen.

**42** Deutsches Textarchiv, http://www.deutschestextarchiv.de/kaempfer_japan01_1777/157, http://www.deutschestextarchiv.de/kaempfer_japan01_1777/158.

**43** See, for example, Georg Gustav Erbkam, Tagebuch meiner egyptischen Reise, Teil 1, Ägypten, 1842–43, in Deutsches Textarchiv, http://www.deutschestextarchiv.de/erbkam_tagebuch01_1842; Theresia Lindnerin, Koch Buch zum Gebrauch der Wohlgebohrenen Frau (um 1780), in Deutsches Textarchiv Qualitätssicherung, http://www.deutschestextarchiv.de/dtaq/book/show/lindnerin_kochbuch_1780.

**44** "Die Grenzboten—Digitalisierung, Erschließung und Volltexterkennung einer der herausragenden deutschen Zeitschriften des 19. und 20. Jahrhunderts," joint project of the Staats- und Universitätsbibliothek Bremen and the Berlin-Brandenburgische Akademie der Wissenschaften (BBAW), http://gepris.dfg.de/gepris/projekt/196492153.

**45** See the CLARIN-D User Guide, version 1.0.1 (December 19), part II, ch. 6, subsection "Text Corpora," http://www.clarin-d.de/en/language-resources/userguide.html.

**46** "CLARIN-D Service Centres," http://www.clarin-d.de/en/clarin-d-centres.html.

## ABSTRACT

In this article we describe the DTA "Base Format" (DTABf), a strict subset of the TEI P5 tag set. The purpose of the DTABf is to provide a balance between expressiveness and precision as well as an interoperable annotation scheme for a large variety of text types of historical corpora of printed text from multiple sources. The DTABf has been developed on the basis of a large amount of historical text data in the core corpus of the project Deutsches Textarchiv (DTA) and text collections from 15 cooperating projects with a current total of 210 million tokens. The DTABf is a "living" TEI format which is continuously adjusted when new text candidates for the DTA containing new structural phenomena are encountered. We also focus on other aspects of the DTABf including consistency, interoperability with other TEI dialects, HTML and other presentations of the TEI texts, and conversion into other formats, as well as linguistic analysis. We include some examples of best

practices to illustrate how external corpora can be losslessly converted into the DTABf, thus enabling third parties to use the DTABf in their specific projects. The DTABf is comprehensively documented, and several software tools are available for working with it, making it a widely used format for the encoding of historical printed German text.

## INDEX

**Keywords:** TEI customization, historical corpora, corpus annotation, interoperability, interchange, schema design, standardization

## AUTHORS

### SUSANNE HAAF

Susanne Haaf works as a research assistant for the German Text Archive (DTA) and CLARIN-D at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW).

### ALEXANDER GEYKEN

Alexander Geyken works at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW), and is head of the project groups of the Digital Dictionary of German Language (DWDS, a long-term BBAW-project) and the German Text Archive (DTA).

### FRANK WIEGAND

Frank Wiegand works as a research assistant/software developer for the German Text Archive (DTA) and the Digital Dictionary of German Language (DWDS) at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW).