The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora

Alexander Geyken
BBAW
Jägerstr. 22/23
D-10117 Berlin
geyken@bbaw.de

Susanne Haaf
BBAW
Jägerstr. 22/23
D-10117 Berlin
haaf@bbaw.de

Frank Wiegand
BBAW
Jägerstr. 22/23
D-10117 Berlin
wiegand@bbaw.de

Abstract

This article describes a strict subset of TEI P5, the DTA 'base format', which combines the richness of encoding non-controversial structural aspects of texts while allowing only minimal semantic interpretation. The proposed format is discussed with regard to other commonly used XML/TEI schemas. Furthermore, the article presents examples of good practices showing how external corpora can either be converted into the DTA 'base format' directly or after cautiously extending it. Thus, the proposed encoding schema contributes to the paradigm shift recently observed in corpus compilation, namely from private encoding to interoperable encoding.

1 Introduction

Up to the end of the 1990s corpus compilation on the basis of the *Guidelines of the Text Encoding Initiative* (TEI; most recent release: P5, cf. Burnard and Baumann, 2012) was mainly a project specific activity. Corpus documents were validated against a project specific document grammar, (possibly) private character encodings were used, and the documents were transformed into proprietary formats in order to get indexed for full text retrieval. In that era of project specific encoding, exchange of documents across projects was not a primary goal, and, in general, character encoding problems as well as differences in the document type grammar (DTD) were obstacles to a broader exchange of data. With the advent of XML and Unicode, documents encoded according to the recommendations of the TEI became interchangeable, or, more precisely, documents encoded in TEI P5 could be safely exchanged on different platforms without worrying about incompatibilities of character encoding. However, differences in structural encodings still remained. The large flexibility of using the TEI Guidelines to encode similar semantic phenomena with different XML elements is one major reason for this problem: for example, there are several ways to encode the hierarchy of sections in documents, either with numbered division elements (`<div1>` ... `<div7>`) elements or by enumerating the hierarchy with numeric `@n`-values: `<div n="1">`, `<div n="2">`, etc. Likewise, there are different ways to encode information about person names (`<persName>`, `<name type="person">`, `<rs type="person">`), several ways to link text passages (`<ref>`, `<ptr>`, `@corresp`, `@next`/`@prev`, ...) etc. However, the main reason for differences in structural encoding resides in the fact that different projects use different subsets of the TEI according to their needs. Problems like these become apparent when the attempt is made to carry out specific tasks with the exchanged data on another platform together with another document collection.

Problems occur on several levels, the first one being the difficulty to create a common style sheet across different document collections encoded in different TEI P5 schemas in order to present all document collections uniformly on the web. Another problem concerns the exchange of TEI metadata: Due to the flexibility of the TEI tag set, the structure of TEI Headers may dif-

fer considerably, which forces harvesting mechanisms exploiting this information in a uniform way to deal with a lot of different cases. This is obviously not the idea of a standard. Examples 1 and 2 illustrate this fact: the information about an author of a work can be either underspecified (Ex. 1) or very detailed (Ex. 2). However, both are valid according to the TEI Header specification.

Example 1

```
<author>Ernst, Ferdinand</author>
```

Example 2

```
<author>
  <persName>
    <forename>Ferdinand</forename>
    <surname>Ernst</surname>
  </persName>
</author>
```

Furthermore, machine-exploitable extraction of document components such as 'retrieve all letters of the document collection' or 'display all quotations in a chapter' pose an enormous problem since division types or entity encoding for quotes do not have to be realized in an ubiquitous way across document collections. Clearly the problem is even worse for complex XPath queries or for data mining tasks where ubiquitous encoding is a necessary prerequisite. To sum up: 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. These problems are widely acknowledged (cf. e.g. Ramsay et al., 2011: p. 1-4; Pytlik Zillig, 2009: p. 187 seq.; Unsworth, 2011: p. 1 seq.; Stührenberg, 2012: p. 141 seq.).

More recently, several attempts were made to increase the interoperability among different document collections by creating common formats. Therefore, subsets of TEI P5 were created reducing the tei_all tag set to a considerably smaller number of elements and attributes (cf. Day, 2010: p. 1). The TEI consortium recommends such customizations of the TEI inventory according to the individual needs of projects instead of taking the whole TEI tagset as a basis for the annotation of a corpus (Burnard and Baumann, 2012: ch. 15.5, 23.2). TEI formats like TEI Lite (Burnard and Sperberg-McQueen, 2006), TEI Tite (Trolard, 2011) or the *Best Practices for TEI in Libraries* (TEI SIG on Libraries, 2011; henceforth: TEI-Lib) are promoted.[1] In addition, several corpus and data curation projects have developed other TEI- or TEI-related formats according to their particular purposes, e.g. TEI Analytics developed by the MONK project (cf. Unsworth, 2011; Pytlik Zillig, 2009; Pytlik Zillig et al., 2012; henceforth: TEI-A), IDS-XCES by the Institute for the German Language in Mannheim and *TextGrid's Baseline Encoding for Text Data in TEI P5* (TextGrid, 2007–2009; henceforth: TextGrid's BE). These formats have been designed to allow for the basic structuring of all written texts and therefore serve as a starting point from which more detailed, possibly project specific text structuring could start.

The remainder of this paper starts with a short presentation of the above-mentioned subsets of the TEI (section 2). In section 3, we motivate the creation of a TEI format for the "Deutsches Textarchiv" (DTA), the DTA 'base format'[2] (henceforth: DTA-BF). Section 4 presents examples of good practice illustrating how different external corpora can be converted into the DTA-BF, thus being interoperable in a wider context, e.g. as part of the text corpora provided by the large European infrastructure project CLARIN.[3] We conclude with a short summary and some ideas about future prospects.

2 Comparison between existing TEI Encoding Formats

In this section, we compare some (well-known) existing XML annotation formats, which are fully or partially based on the TEI Guidelines, namely the above-mentioned formats TEI Tite, TEI Lite, TEI-Lib, TEI-A, IDS-XCES, and TextGrid's BE. The formats are evaluated with respect to their applicability for the annotation of historical corpora such as the "Deutsches Textarchiv".

All of the mentioned encoding formats have in common their attempt to unify large amounts of – possibly different – texts. However, considerable differences persist. TEI Tite and TEI-

[1] Cf. www.tei-c.org/Guidelines/Customization.

[2] www.deutschestextarchiv.de/doku/basisformat.

[3] Common Language Resources and Technology Infrastructure; www.clarin.eu.

Lib are complementary in the sense that they provide annotation guidelines for text digitization undertaken by libraries. While TEI Tite was created to allow for basic text structuring undertaken by external vendors, therefore intending "to prescribe *exactly* one way of encoding a particular feature of a document in as many cases as possible" (Trolard, 2011: ch. 1; Day, 2010: p. 16), TEI-Lib is intended "to support in-house encoding that adheres as closely as possible to common TEI practice and library standards yet still leaves room for variation in local practice" (TEI SIG on Libraries, 2011: ch. 2; cf. Dalmau/Schlosser, 2010: p. 355 seq.). Both formats are therefore especially suited to the task of annotating large amounts of heterogeneous text material in a library context. TEI Lite pursues a similar goal, being meant to "meet 90 % of the needs of 90 % of the TEI user community" (Burnard and Sperberg-McQueen, 2006: Prefatory note), but without being restricted to library usage. TEI-A results in a customization which is supposed to be suitable for the annotation of diverse texts from variable sources, as well, but has a different starting point than TEI-Lib, TEI Tite and TEI Lite, since it was created as a format to bring together texts which were already annotated individually (Pytlik Zillig, 2009: p. 188 seq.). Similarly, TextGrid's BE is intended as basic encoding format enabling the intertextual search within TextGrid (TextGrid, 2007-2009: p. 6.). Finally, IDS-XCES serves as an encoding scheme for the IDS corpus texts. It is originally based on XCES, the XML adaption of CES, which was extended, partially with respect to the TEI Guidelines, according to the requirements of the IDS corpora (Institute for the German Language Mannheim, 2012; Stührenberg, 2012: p. 175-180).

Despite their individual genesis and purpose there is a set of structuring elements common to all of the named formats. E.g. the text of a document is divided into a `<front>`, a `<body>`, and a `<back>` area, paragraphs are structured as such using the element `<p>`, verse (at some point) should be encoded using `<lg>` and `<l>`, speech acts in a drama are encoded with the `<sp>` element etc. Such analogies show that there is a commonplace structuring level, which might be classified as level-1-encoding, or as, what the TEI P5 *Recommendations for the Encoding of Large Corpora* subsume under "required" elements, demanding that "texts included within the corpus will always encode textual features in this category, should they exist in the text" (Burnard and Baumann, 2012: ch. 15.5). Still, in some cases the selections of TEI P5 elements differ. E.g. only TEI-A offers tagging solutions for screenplays, such as `<view>` and `<camera>`. Furthermore, the flexibility of the TEI specification allows that semantically similar phenomena are addressed differently by the encoding formats. E.g. TEI Lite, TEI Tite, and TEI-Lib allow for the encoding of additions and deletions which were performed on the source document by providing the elements `<add>` and `<del>`, whereas TextGrid's BE offers the elements `<addSpan>` and `<delSpan>` for this purpose.

The appropriate selection of elements is just one factor for the evaluation of annotation formats. Almost equally important is the appropriate choice of attributes and their corresponding values, ideally expressed as a fixed value set. In addition, there are more general factors to be taken into account with regard to the practical applicability in specific project contexts, namely the determination of annotation levels, solutions to the provision of metadata, comprehensive guidelines on text transcription and editorial interventions as well as the documentation of the format itself. Last but not least annotation formats differ in their degree of conformance to the TEI Guidelines - is the format a strict subset of TEI-P5 or does it make use of extensions.

Tables 1 and 2 summarize the commonalities and differences between the annotation formats considered here with respect to the abovementioned factors. These factors serve as a guideline for the discussion of the DTA base format that is presented in the next section.

3 The DTA 'base format'

This section discusses the DTA-BF, a customization of the TEI P5 tag set, created for the encoding of (historical) German text in large text corpora. The format emerged from previous work on a TEI P5 corpus project, the DWDS corpus (Geyken, 2007). The DWDS corpus is a cor-

	TEI P5 subset	documentation	element-wise attribute selection	fixed/recommended attribute values	levels
TEI Tite	no	yes, mainly element-wise	class-wise	no	no
TEI Lite	yes	yes	class-wise; some element-wise recommendations for attributes	no	no
TEI for Libraries	yes	yes	selection of generally recommended attributes	no	yes[a]
TEI Analytics	no	yes, element-wise; but examples include undocumented elements	yes	in some cases (e.g. recommended values for `@unit` and `@part`; fixed values for `@scope`)	no
TextGrid's Baseline Encoding	yes	yes, but examples include undocumented elements	in some cases (e.g. for inline elements)	in some cases	no
IDS-XCES	no	only changes to XCES are communicated; no documentation of the usage of elements	yes	in some cases	no
DTA 'base format'	yes	yes	yes	yes	yes

[a]levels are not strictly cumulative

Table 1: Comparison of annotation formats – part 1

pus of the 20th century German language of written text. It is roughly equally distributed over time and over five genres: journalism (approx. 27 % of the corpus), literary texts (26 %), scientific texts (approx. 22 %) and functional texts (approx. 20 %), as well as a smaller number of transcripts of spoken language (5 %). The focus of encoding was put on the non-controversial structural aspects of the documents with the goal to facilitate cross-document full text retrieval for linguistic purposes.[4] With the start of the project *Deutsches Textarchiv*[5] (DTA) in 2007, the TEI P5 compliant schema had to be extended considerably for two main reasons: faithful page per page presentation of the entire works, and the necessity to deal with older prints, thus having to cope with additional structural variation. The DTA project works on building a text corpus for the historical New High German. Within seven years of work, a selection of 1,300 texts of different text types, originating from the 17th to 19th century, are being digitized and annotated according to the TEI P5 Guidelines. Linguistic analyses are added to the digitized text sources in a stand-off format for further corpus research.

The goal of the DTA-BF is to provide a homogeneous text annotation for a collection of historical texts being heterogeneous with respect to the date of their origin (1650–1900) and text types (literary texts, functional texts, scientific texts). To achieve this, the DTA-BF follows some overall restrictions, this way combining the different benefits of the named formats.

In the remainder of this section we show how the DTA-BF deals with the factors mentioned in section 2 ensuring the quality of corpora encoded according to the DTA-BF as well as the applicability of the DTA-BF for other projects.

3.1 Selection of Elements, Attributes, and Values

The selection of DTA-BF elements corresponds to a large extent to the tagset of TEI Lite. However, unlike TEI Lite, the DTA-BF also provides a restricted set of attribute values in order to minimize the possibility of using different tagging solutions for similar structural phenom-

[4]Cf. www.dwds.de.

[5]Cf. www.deutschestextarchiv.de. The DTA is funded by the *German Research Foundation* between 2007 and 2014.

	inline metadata	solutions for editorial interventions[a]	transcription guidelines	text type specific encoding guidelines[b]
TEI Tite	no	no	no	newspapers
TEI Lite	yes	CN; AD-ST; AD-Ed (except <supplied>); AE	no	no
TEI for Libraries	yes	CN (except <reg>, <orig>); AD-ST; AD-Ed (except <supplied>)	instructions for quotation marks and hyphens	interviews
TEI Analytics	yes	CN; AD-ST (except <del>); AD-Ed (no <gap>, <unclear>; <supplied> is not documented, but used in examples)	no	screenplays
TextGrid's Baseline Encoding	yes	CN; AD-ST (<addSpan>, <delSpan> instead of <add>, <del>); AD-Ed (except <supplied>); AE	no	dictionary entries
IDS-XCES	no[c]	CN (except <choice>, <sic>; <corr> with @sic); AD-Ed (except <unclear>, <supplied>); AE (except <expan>; <abbr> with @expan)	no	spoken language (e.g. dialogues, speeches, debates, interviews)
DTA 'base format'	yes	CN; AD-Ed (except <unclear>; usage of <supplied> instead); AE	yes	funeral sermons, newspapers

[a] I.e. correction and normalization (CN; includes <choice>, <sic>, <corr>, <reg>, <orig>); deletions, and additions: in the source text (AD-ST; includes <add>, <del>), editorial (AD-Ed; includes <gap>, <unclear>, <supplied>); abbreviations and expansions (AE; <choice>, <abbr>, <expan>)

[b] Other than prose, verse, drama, letter

[c] A metadata format is provided, which contains TEI Header elements as well as a considerable amount of other elements

Table 2: Comparison of annotation formats – part 2

ena.[6] This goal is explicitely expressed by the TEI Tite guidelines, as well.[7] However, only some recommendations for attribute values are given, whereas no firm value lists are integrated in the TEI Tite schema. Other formats, such as TEI-Lib, explicitely decided against the restriction of attribute values.[8]

In our opinion, it is crucial to provide a detailed specification not only of elements but of corresponding attributes and values as well to minimize ambiguities of the tag set. Therefore, each of the 105 TEI P5 elements currently contained by the DTA-BF tagset[9] is provided with a fixed list of possible attributes and values. The selection of attributes specified for each element is restricted not only class-wise but element-wise. Attribute values may occur within the DTA-BF in three different ways:

1. In general, the DTA-BF prescribes a fixed set of possible values for each attribute, thus being even more restrictive than TEI Tite. E.g. possible values for the @unit attribute of the element <gap> are: "chars", "words", "lines", or "pages". The selection of values in the DTA-BF can either apply for an attribute in every possible context or depend on the surrounding element.

2. In rare cases, where attribute values cannot

[6] The necessity of minimal semantic ambiguity of the tagset has been pointed out by Unsworth (2011: § 7).

[7] "Tite is meant to prescribe exactly one way of encoding a particular feature of a document in as many cases as possible, ensuring that any two encoders would produce the same XML document for a source document" (Trolard, 2011: ch. 1).

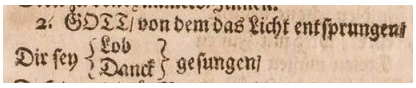
[8] Cf. e.g. the statement of TEI-Lib about possible @type-values: "Constructing a list of acceptable attribute values for the @type attribute for each element, on which everyone could agree, is impossible. Instead, it is recommended that projects describe the @type attribute values used in their texts in the projects ODD file and that this list be made available to people using the texts" (TEI SIG on Libraries, 2011: ch. 3.8.1).

[9] I.e. about 80 elements used for the annotation of the texts themselves plus 25 additional elements needed specifically for the representation of metadata within the TEI Header.

be reduced to a fixed set, restrictions are made with respect to the data type. E.g. the value of the attribute `@quantity` within the element `<gap>` has to be a non-negative integer (`data.count`).

3. Finally, there are cases, in which attribute values cannot be restricted by the schema at all. E.g. the value of `@n` in `<pb>` may consist of alphanumeric characters (e.g. `<pb n="16">`, `<pb n="XVI">`) as well as strings (e.g. `<pb n="[16]">`). In such (rare) cases value restrictions are given in the DTA-BF guidelines.

The DTA-BF has been designed to cover all annotation requirements for a basic structuring of the large variety of historical texts that are dealt with in the DTA. On this common structural level each typographically marked segment in the source text (centered, printed in bold or italics, printed with an individual typeface, smaller or larger letters) is labelled preferably with one basic semantic category (citation, poem, title, note etc.) or, if a semantic function cannot easily be assigned, with the formal category describing the typographical characteristics of the respective text segment. Fig. 1 shows the combination of semantic (`<lg>`, `<l>`) and formal (`<list rendition="#leftBraced #rightBraced">`, `<item>`) tagging.




```
<lg n="2">
  <head>2.</head>
  <l>GOTT/ von dem das Licht
    ent&#x17f;prungen/</l><lb/>
  <l>Dir &#x17f;ey
    <list rendition="#leftBraced
      #rightBraced">
      <item>Lob</item><lb/>
      <item>Danck</item>
    </list>
  ge&#x17f;ungen/</l><lb/>
  [...]
</lg>
```

Figure 1: Friedrich Rudolph Ludwig von Canitz: Gedichte. Berlin, 1700. Image 16. www.deutschestextarchiv.de/canitz_gedichte_1700/16

3.2 Annotation Levels

As stated above, the DTA-BF is supposed to serve as a guideline for the homogeneous annotation of heterogeneous historical corpus texts. However, the necessity of a homogeneous format for the annotation of historical text seems to be opposed to the fact, that different projects usually have different needs as for how detailed text structuring should be. This problem is aggravated by the fact that text structuring becomes more labor intensive the more detailed it is. To address this problem, the TEI *Recommendations for the Encoding of Large Corpora* advise users who wish to create language corpora to define four levels of text annotation (required, recommended, optional, and proscribed) when determining a subset of TEI elements appropriate to the anticipated needs of the project (Burnard and Bauman, 2012: ch. 15.5).

Like TEI-Lib, the DTA-BF defines different encoding levels according to the depth of text structuring, thus categorizing all available DTA-BF elements due to different text structuring necessities and depths. In accordance with the TEI Guidelines but in contrast to TEI-Lib (TEI SIG on Libraries, 2011: ch. 1), the levels 1 to 3 of the DTA-BF are strictly cumulative.[10] The first level represents the least common denominator for text structuring, therefore containing elements that are mandatory for basic semantic text annotation. The elements in level 2 are strongly recommended but not mandatory. Level 3 contains optional elements, which can be used for in-depth semantic text structuring, but are not applied extensively throughout the DTA core corpus.[11] Thus, the obligation to use the provided elements decreases with the increase of levels and, in connection with that, the depth of text structur-

[10] For an overview on the DTA-BF elements and their corresponding levels cf. www.deutschestextarchiv.de/doku/basisformat_table.

[11] However, the DTA-BF remains an annotation format for the structuring of historical corpus texts, esp. serving linguistic purposes. Projects, which aim at providing historical critical editions of texts, will need further annotation possibilities (e.g. an inventory for a critical apparatus as specified in chapter 12 of the TEI guidelines; cf. Burnard and Baumann, 2012: ch. 12). Such projects might want to start off with the DTA-BF for annotation and extend it according to their requirements.

ing. The fourth level contains an exception list of elements which should be avoided in favor of a different solution provided by the DTA-BF.

3.3 Metadata

Like TEI-Lib (TEI SIG on Libraries, 2011: ch. 4.1), the DTA-BF provides a TEI Header customization which allows to express rich bibliographic metadata for each corpus text. The DTA-BF metadata specification focuses on the bibliographic description of written corpora. We provide conversion routines to other standards such as the Component Metadata Infrastructure CMDI[12], which is the recommended metadata format in CLARIN.[13]

3.4 Text Transcription and Editorial Interventions

In addition to the DTA-BF, extensive transcription guidelines are provided in order to support common transcription practices for each text in the DTA corpus.[14] To this end, furthermore, the DTA-BF contains regulations for possible editorial interventions. From the TEI formats mentioned above, only the TEI-Lib guidelines point out the necessity of transcription guidelines, but limit their advice to the handling of punctuation and hyphenation problems (TEI SIG on Libraries, 2011: ch. 3.2).

3.5 Documentation

The DTA-BF comes with a detailed documentation[15] explaining the usage of each element, attribute, and value according to the possible annotation needs in text structuring. The documentation contains examples taken from the DTA corpus and illustrates typical encoding scenarios as well as exception cases.

There is also an ODD[16] specification for the DTA-BF together with a corresponding RNG schema[17] generated with TEI-ROMA.[18]

[12] Cf. www.clarin.eu/cmdi.

[13] Cf. footnote 3.

[14] Cf. www.deutschestextarchiv.de/doku/richtlinien.

[15] Cf. www.deutschestextarchiv.de/doku/basisformat.

[16] One document does it all; cf. www.tei-c.org/Guidelines/Customization/odds.xml.

[17] Cf. www.deutschestextarchiv.de/basisformat.odd; www.deutschestextarchiv.de/basisformat.rng.

[18] Cf. www.tei-c.org/Roma.

3.6 Relation to the TEI Guidelines

Like TEI Lite, TEI-Lib, and TextGrid's BE, the DTA-BF is a strict subset of the TEI P5 tag set. It is therefore entirely compatible with the TEI P5 Guidelines in that they are only customized by selection, but not extended in any way.

4 Lifecycle of the DTA 'base format' within DTAE

DTA Extensions (DTAE) is a module of the DTA project with the goal to integrate digitized historical German texts drawn from external sources into the DTA core corpus. There are two prerequisites for those texts: they need to be considered as influential with respect to the goal of the DTA to compile a historical reference corpus, and they have to dispose of a high transcription quality.

External resources may be transcribed either in a word processing or HTML environment – a case we do not discuss here since it has no effect on the DTA-BF – or more often (as more and more philological projects adopt the TEI) be encoded in a customized TEI P5 format. In this case, a transformation of the customized TEI schema into the DTA-BF has to be specified. In general, all texts are subject to a quality assurance phase before being published in the DTA environment (Geyken et al., 2012; Haaf et al., 2012). For this task, a web-based distributed quality assurance platform has been implemented (Wiegand and Geyken, 2011), where users can proofread texts page by page and report different kinds of errors. As a result of the conversion and correction process, material from heterogeneous corpus formats is made accessible in the context of one homogeneous, high-quality text corpus.

So far, corpus texts from 10 external projects with a total of 200,000 pages were integrated into the DTA corpus after being converted into the DTA-BF, including *Blumenbach online*, *AEDit*, and *Dinglers Polytechnisches Journal*.[19]

We distinguish three cases for the integration of external TEI-encoded corpora into the DTA environment: 1. The conversion of the customized TEI schema into the DTA-BF can be done automatically, since the DTA-BF provides a semanti-

[19] Cf. www.deutschestextarchiv.de/doku/dtae for the full list of DTAE projects.

cally equivalent solution. 2. The solution adopted in the customized TEI schema corresponds to a text phenomenon which has not been considered by the DTA-BF so far and which in turn leads to a modification of the DTA-BF. 3. The external text corpus cannot be automatically converted, either because the underlying TEI schema is too flexible thus leading to structuring ambiguities (cf. section 1), or because the schema is applied inconsistently over the text collection. Since this last case requires manual intervention, it is only considered for external texts which are stable, either because the project is finished, or because the quality of the transcription and the structural encoding is sufficient, which means no additional annotation work is likely to be carried out on the source text.

The customized TEI schema of the project *Dinglers Polytechnisches Journal* may serve as an example for the first case. This schema defines missing transcriptions due to illegibility of the text source as follows:[20]

```
<unclear reason="problem">
 [Fehlender Text
  (engl.: missing text)]
</unclear>
```

Even though the DTA-BF does not include the TEI element `<unclear>`, this expression can easily be converted into the equivalent of the DTA-BF annotation:

```
<gap reason="illegible"/>
```

The following two examples illustrate the second case, i.e. modifications of the DTA-BF according to the requirements of external corpus projects:

The *Blumenbach online* project provides editorial figure descriptions (`<figDesc>`), a kind of additional information about the source text given by the editor. Such additional information was not foreseen by the DTA-BF. Since the `<figDesc>` is only a special case of an editorial comment, the DTA-BF element `<note>` was extended by the attribute-value combination `@resp="editorial"`. With this extension, we were able to preserve the figure descriptions of the edited Blumenbach texts and to generally allow for editorial comments elsewhere.

Furthermore, modifications of the DTA-BF may become necessary due to the integration of new (special) text types in the DTA corpora. E.g. in the context of the DFG funded project *AEDit Frühe Neuzeit (Archiv-, Editions- und Distributionsplattform für Werke der Frühen Neuzeit)* the *Forschungsstelle für Personalschriften* (*Academy of Sciences and Literature in Mainz*) is currently digitizing funeral sermons of the former municipal library in Wrocław. The digitized texts are being annotated according to the DTA-BF. The addition of new specific `@type`-values for `<div>` elements became necessary in order to allow for the naming of different text types within a funeral sermon. The new values added to the existing value selection were prefixed `fs` in order to limit their usage to the document type "funeral sermon" (e.g. `fsSermon`, `fsConsolationLetter`, `fsCurriculumVitae`, `fsEpitaph` etc.).

However, possible modifications of the DTA-BF are considered carefully in order to avoid negative effects on the annotation consistency of the DTA corpus.

5 Conclusion and further prospects

In this article, we have presented the DTA 'base format', a strict subset of TEI P5. The DTA-BF has been designed and developed with the goal to cope with a large variety of text types of written German corpora. It is a reasonable common denominator for a large reference corpus of the historical New High German ranging from 1650 to 1900. It goes without saying that the success of the DTA-BF is largely dependent on its adoption by other projects, namely the number of documents encoded in the format. Establishing the usage of the DTA-BF in a broader context may be supported considerably within a large infrastructure such as provided by CLARIN and, for the German context, CLARIN-D, where major text corpus providers are gathered pursuing the goal to define policies which guarantee the interoperability of resources which are integrated into the infrastructure. The *Berlin-Brandenburg Academy of Sciences and Humanities* (BBAW) as a partner of the CLARIN-D project, as a future CLARIN Center, and as the coordinator of

[20] Cf. dingler.culture.hu-berlin.de/article/pj003/ar003042 for an example.

the work package 'Resources and Tools' plays an important role in the discussion process. In addition, a CLARIN-D corpus project has recently been started with the goal to curate already existing corpus texts of the 15th to 19th century and to integrate them into the CLARIN-D infrastructure by using the DTA-BF as a starting point, thus enabling the DTA-BF to evolve in an environment of even more heterogeneous text resources. The project partners of this CLARIN-D curation project are the BBAW (coordination), the *Herzog August Library of Wolfenbüttel*, the *Institute for the German Language Mannheim*, and the *University of Gießen*.

References

Lou Burnard, and Syd Bauman. 2012. *P5: Guidelines for Electronic Text Encoding and Interchange*, Version 2.1.0, June 17th, 2012. www.tei-c.org/release/doc/tei-p5-doc/en/html.

Lou Burnard, and C. M. Sperberg-McQueen. 2006. *TEI Lite: Encoding for Interchange: an introduction to the TEI – Revised for TEI P5 release* (February 2006). www.tei-c.org/Guidelines/Customization/Lite/.

Michelle Dalmau, and Melanie Schlosser. 2011. *Challenges of serials text encoding in the spirit of scholarly communication.* In: Library Hi Tech 28,3 (2011), pp. 345-359. http://dx.doi.org/10.1108/07378831011076611.

Michael Day. 2010. *IMPACT Best Practice Guide: Metadata for Text Digitisation & OCR.* www.impact-project.eu/uploads/media/IMPACT-metadata-bpg-pilot-1.pdf/.

Alexander Geyken. 2007. *The DWDS corpus: A reference corpus for the German language of the 20th century.* In: Christiane Fellbaum (Hg.): *Collocations and Idioms: Linguistic, lexicographic, and computational aspects.* London, 23–41.

Alexander Geyken, Susanne Haaf, Bryan Jurish, Matthias Schulz, Christian Thomas, and Frank Wiegand. 2012. *TEI und Textkorpora: Fehlerklassifikation und Qualitätskontrolle vor, während und nach der Texterfassung im Deutschen Textarchiv.* In: *Jahrbuch für Computerphilologie*, Jg. 9, 2012.

Susanne Haaf, Frank Wiegand, and Alexander Geyken. 2012. *Measuring the correctness of double-keying: Error classification and quality control in a large corpus of TEI-annotated historical text.* Accepted to be published in the *Journal of the Text Encoding Initiative* 3, 2012.

Institute for the German Language (IDS) Mannheim. 2012. *IDS-Textmodell: Unterschiede gegenüber XCES.* www.ids-mannheim.de/kl/projekte/korpora/idsxces.html (accessed June 24th, 2012).

TEI SIG on Libraries. 2011. *Best Practices for TEI in Libraries. A TEI Project.* Ed. by Kevin Hawkins, Michelle Dalmau, and Syd Bauman, Version 3.0 (October 2011). www.tei-c.org/SIG/Libraries/teiinlibraries/main-driver.html.

Stephen Ramsay, and Brian Pytlik Zillig. 2011. *Code Generation Techniques for Document Collection Interoperability.* Chicago Colloquium on Digital Humanities and Computer Science 2011. http://chicagocolloquium.org/wp-content/uploads/2011/11/dhcs2011_submission_6.pdf.

Maik Stührenberg. 2012. *Auszeichnungssprachen für linguistische Korpora: theoretische Grundlagen, De-facto-Standards, Normen.* Bielefeld: Universität Bielefeld. urn:nbn:de:hbz:361-24927723.

TextGrid. 2007–2009. *TextGrid's Baseline Encoding for Text Data in TEI P5.* www.textgrid.de/fileadmin/TextGrid/reports/baseline-all-en.pdf (accessed June 24th, 2012).

Perry Trolard. 2011. *TEI Tite – A recommendation for off-site text encoding.* Version 1.1, September 2011. www.tei-c.org/release/doc/tei-p5-exemplars/html/tei_tite.doc.html.

John Unsworth. 2011. *Computational Work with Very Large Text Collections. Interoperability, Sustainability, and the TEI. Journal of the Text Encoding Initiative* 1 (June 2011). http://jtei.revues.org/215 (access June 24th, 2012).

Frank Wiegand, and Alexander Geyken. *Quality assurance of large TEI corpora.* Poster, presented at the 2011 Annual Conference and Members' Meeting of the TEI Consortium: *Philology in the Digital Age.* Würzburg, 2011. www.deutschestextarchiv.de/doku/tei-mm-2011_poster.pdf.

Brian Pytlik Zillig. 2009. *TEI Analytics: converting documents into a TEI format for cross-collection text analysis. Literary and Linguistic Computing*, 24(2), 187–192. doi:10.1093/llc/fqp005.

Brian Pytlik Zillig, Syd Bauman, Julia Flanders, and Steve Ramsay. 2012. *MONK TEIAnalytics. Target schema for MONK ingestion transformation.* segonku.unl.edu/teianalytics/TEIAnalytics.html (accessed June 24th, 2012).