Journal of the Text Encoding Initiative

Issue 10 | December 2016 - July 2019
Selected Papers from the 2015 TEI Conference

Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format

Susanne Haaf and Christian Thomas

Electronic version
URL: http://journals.openedition.org/jtei/1650
DOI: 10.4000/jtei.1650
ISSN: 2162-5603

Publisher
TEI Consortium

Electronic reference
Susanne Haaf and Christian Thomas, « Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format », *Journal of the Text Encoding Initiative* [Online], Issue 10 | December 2016 - July 2019, Online since 08 August 2017, connection on 03 July 2019. URL : http://journals.openedition.org/jtei/1650 ; DOI : 10.4000/jtei.1650

Enabling the Encoding of Manuscripts within the DTABf: Extension and Modularization of the Format

Susanne Haaf and Christian Thomas

ERRATA

This paper was revised on 2018-01-12 to correct an error in figure 2. The previous version is archived at https://journals.openedition.org/jtei/1734.

1. Introduction

1 Since its beginning in 2007, the *Deutsches Textarchiv* project (DTA)[1] has been creating a corpus of historical German printed works dating from the seventeenth to the nineteenth century. The underlying bibliography comprises works of various text types and domains in order to allow for comprehensive insights into the whole range of written material available in the German language at different points in time. The corpus allows for research on the development of the New High German language as well as research from various other disciplines, such as historical or literary studies.

2 An important goal in this context was to ensure interoperability of all corpus data, which requires homogeneous TEI annotation of the whole corpus. For this purpose, the DTA "Base Format" (DTABf) was developed, a TEI format for the unambiguous, homogeneous annotation of corpus data (Geyken et al. 2012; Haaf, Geyken, and Wiegand 2014–15).[2] It was continuously adapted to tagging requirements of the DTA core corpus texts as well as of external texts curated from other projects (Thomas and Wiegand 2015), and successfully applied to more than 2,800 historical works contained in the DTA. Moreover, the DTABf has attracted interest beyond the borders of the DTA project (e.g., from the projects *Hamburger Schlüsseldokumente zur deutsch-jüdischen Geschichte*[3] and *ePoetics*[4]). It is now, among other formats, recommended by the CLARIN infrastructure project for the annotation of historical printed sources (CLARIN-D AP 5 2012, II.6), as well as by the German Research Foundation for basic TEI encoding of linguistic corpora (DFG 2015a) and as an archival format for edition projects in literary studies (DFG 2015b).

3 While until recently the DTA corpus and hence the DTABf have been focused on printed works, there is an increasing interest in the integration of manuscripts into the DTA corpus, as well, for example by scholars who are providing transcriptions and are interested in re-using the DTA infrastructure in terms of linguistic analysis or collaborative work, or, more generally, for publishing and sharing their project output with the scholarly community. Specifications and cautious extensions of the DTABf tagset for certain text types have been provided before, based on, for instance, findings by cooperating projects and within subcorpora integrated into the DTA platform (e.g., for historical newspapers and funeral sermons: see Haaf and Schulz 2014). The category "manuscripts," however, represents an extensive and diverse quantity of texts and text types, containing a significant amount of new and sometimes complex structural and textual phenomena.

4 Thus, efforts were made to create a pure TEI subset for the unambiguous annotation of manuscripts, based on the DTABf.[5] This paper describes the development of this DTABf for Manuscripts, or DTABf-M, its current data basis, its present components, and the efforts towards the reorganization of the DTABf system as a whole in order to include the DTABf-M and previous extensions to the DTABf, but also to prevent the existing DTABf for printed texts from losing its central values of restrictedness and clarity through overly extensive growth.

2. Problem Statement

5 Until recently, the textual basis of the DTA corpora has almost exclusively consisted of printed texts of the historical New High German period (~1600 to ~1900). Nevertheless, now that a solid corpus basis has been created it is desirable to extend the DTA workflow to manuscripts as well, mainly for two reasons:

6 First, a growing number of scholars dealing with historical German texts (e.g., digital scholarly editions) and focusing on the digitization of manuscripts (letters, diaries, etc.) would like to use the restrictive DTA guidelines as a starting point for their annotation, and to use the DTA platform, which requires DTABf conformity, as a publication and working environment for their projects. Thus, adapting the DTABf for manuscript-specific use and thereby enabling access to the DTA infrastructure (including linguistic analysis[6] as well as collaborative text correction and annotation[7]) is a matter of supporting scholarly projects in their usage of the DTA infrastructure, which is part of the DTA's mission. Second, while the DTA corpus is fairly diverse with regard to (printed) text types and disciplines, the absence of manuscripts causes certain "core text types"[8] of writing to be underrepresented within the corpus (e.g., private letters, diaries, *alba amicorum*). This is due to the fact that some text types were rarely or never distributed in print. Therefore—in addition to the benefit of extending the corpus base—it is in the DTA's own interest to include (historical) manuscripts in the corpus.

7 While the DTABf covers a wide range of phenomena in printed texts, there are manuscript-specific features which are not covered by this format. Thus, in order to include manuscripts in the DTA, it is necessary to create a TEI format which provides solutions for manuscript-specific phenomena. To be suitable for the DTA, its workflows, and its processes, this format needs to address three prerequisites: First, it should be based on the original DTABf tagset, mainly reusing it and only extending it if unavoidable (see section 4.1). Second, like the DTABf, it should be created in a data-driven way, that is, based on actual phenomena found in handwritten texts which are transcribed and encoded using the DTABf (see sections 3 and 4.2). Third, the format should complement the DTABf, not replace it. Hence, it is necessary to find a modular way of integrating the DTABf-M into the DTABf (see section 5).[9]

3. Data Basis

8 The DTABf for Manuscripts (DTABf-M) has been developed primarily in cooperation with the project *Hidden Kosmos*, hosted by the Humboldt-Universität zu Berlin.[10] Hidden Kosmos publishes handwritten lecture notes by attendees of Alexander von Humboldt's world-famous "Kosmos-Lectures" in 1827/28. Since June 2014, nine complete volumes with a total of more than 3,500 manuscript pages have been manually transcribed, annotated in TEI XML, and published via the DTA infrastructure. Most of these manuscripts were keyed manually by a vendor and published at an early stage in the web-based quality assurance platform DTAQ. There, the transcription as well as the annotation of each document was checked and corrected, if necessary; DTAQ also provided the means to add additional markup, such as the tagging of person names (`<persName>`), directly at page level. After the process of quality control has been completed, the manuscripts were released on the DTA website.[11] While the *Hidden Kosmos* project is now complete, development of the DTABf-M continues through work on other manuscripts.

9 The DTABf-M is being used and further enriched for *Travelling Humboldt—Science on the Move*,[12] a long-term project at the Berlin-Brandenburg Academy of Sciences and the Humanities. The project aims to provide a digital critical edition of handwritten documents by Alexander von Humboldt connected to his American and Siberian journeys, including his travel journals, which display a very complex order of text segments on each page and were composed and altered in several writing stages.

10 Though these two data sources are homogeneous with regard to time of creation (early nineteenth century), they already included various manuscript-specific phenomena, both frequent (e.g., additions and deletions to the text basis) and rare (e.g., inline notes), that the DTABf-M tagset could be based upon.

11 A third addition to the DTA manuscript corpus is provided by the project *Digitale Edition der Briefe Erdmuthe Benignas von Reuß-Ebersdorf*,[13] where a corpus of letters and other handwritten documents of this pietistic regent from the early eighteenth century is being prepared for a digital scholarly edition.

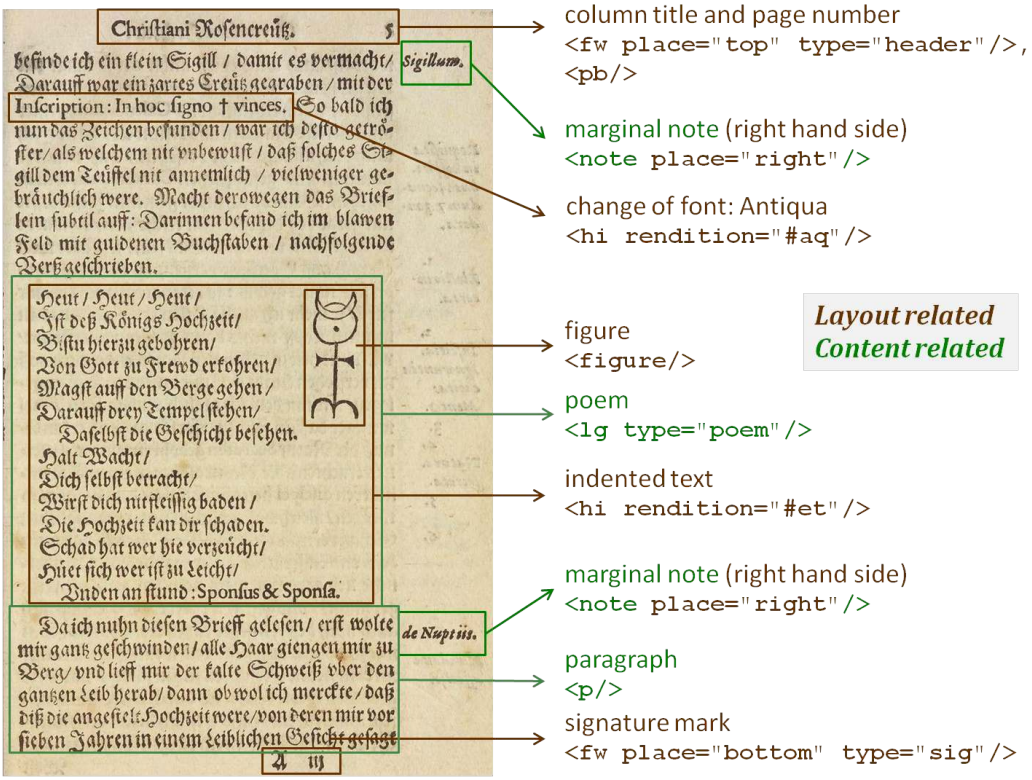
12 Plans to further develop DTABf-M will rely on additional historical sources in the course of their inclusion in the corpus.

4. Steps toward the DTABf for Manuscripts (DTABf-M)

4.1 From Annotating Printed Texts to Annotating Manuscripts

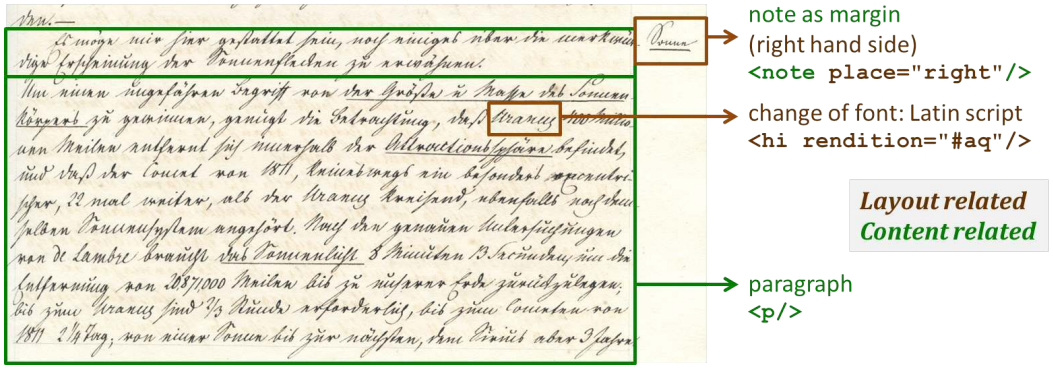
13 Our approach is to base the DTABf for manuscripts on the established DTABf for printed texts, reusing the existing tagset wherever possible and only making changes (enhancements or reductions) when new tags are needed to represent phenomena exclusively found in manuscripts or printed texts, respectively. This is rather straightforward, since, as exemplified by figures 1, 2, 3, and 4, a large number of textual phenomena can be found in both handwritten and printed texts. Figure 1 illustrates a printed page containing some common text structures together with the DTABf elements that apply in those cases. Phenomena on this page include marginal notes, figures, poems, and paragraphs. Figures 2, 3, and 4 show examples of manuscript pages containing similar text structures, showing that the same DTABf elements can be applied there. In fact, much of the DTABf tagset for printed texts can similarly be applied to manuscripts.

Figure 1. Example of textual structures in printed texts and annotation conventions within the DTABf.



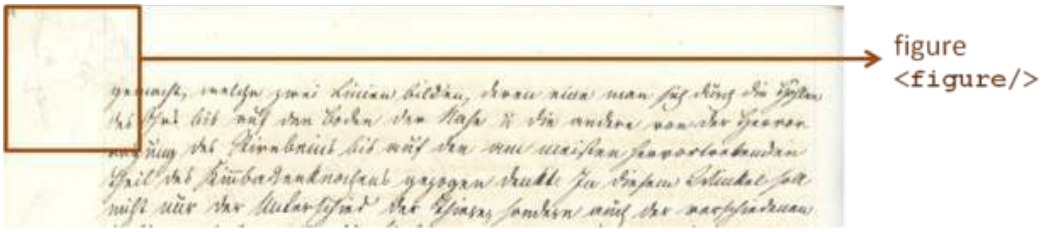
(Andreä 1616, http://www.deutschestextarchiv.de/valentin_hochzeit_1616/9).

Figure 2. Example of textual structures in manuscripts and possible ways to annotate them with DTABf tags.



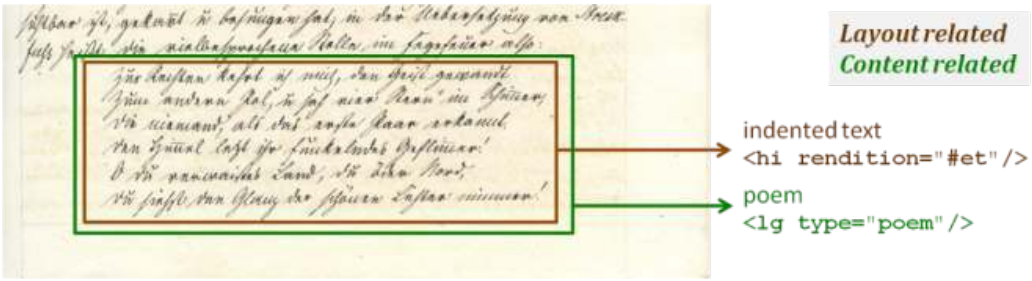
(Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/149).

Figure 3. Example of textual structures in manuscripts and possible ways to annotate them with DTABf tags.



(Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/88).

Figure 4. Example of textual structures in manuscripts and possible ways to annotate them with DTABf tags.

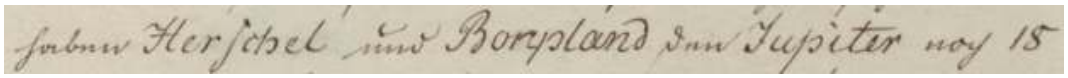


(Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/139).

14 As illustrated by figures 1, 2, 3, and 4, there are many structural similarities between handwritten and printed texts with regard to document structure, the arrangement of the text, and the meaning of layout specifics. Furthermore, inline phenomena may also display similarities. For instance, even certain types of emphasis are at least comparable, if not identical, between printed texts and manuscripts.

15 One example of change in the style of handwriting is illustrated in figure 5. The style changed from old German script (Kurrent) for the general text body to Latin script for distinct terms like proper names or foreign language material. We consider this to be analogous to the change of font from Blackletter (Fraktur) to Antiqua types in print. The DTABf solution for changes from Fraktur to Antiqua typeface is to use the `<hi>` element with `@rendition="#aq"`. This tagging can thus be used similarly for changes from Kurrent to Latin script in manuscripts (figure 5).[14]

Figure 5. The proper names "Herschel," "Bonpland," and "Jupiter" are distinguished by the use of Latin script from the remainder of the text written in Kurrent.

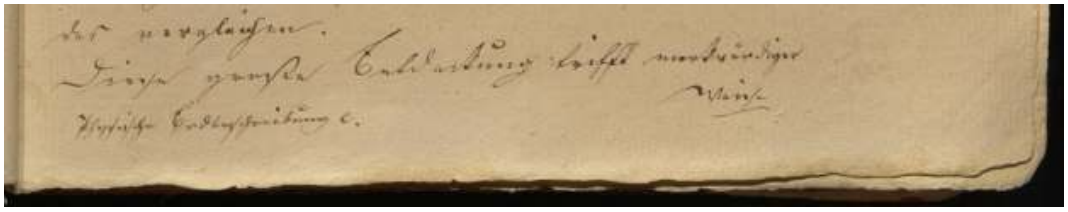


(Anonymous [1827/28a], http://www.deutschestextarchiv.de/nn_msgermqu2345_1827/58).

```
  […] haben <hi rendition="#aq">Herschel</hi> und <hi rendition="#aq">Bonpland</
hi> den <hi rendition="#aq">Jupiter</hi> noch 18<lb/>[…]
```

16 Another example for similar inline phenomena in manuscripts and printed texts is the underlining of important phrases or keywords, represented in the DTABf as `<hi rendition="#u">` for printed texts and manuscripts alike. Furthermore, though this feature is far more frequent in prints, manuscripts may also contain catchwords or signature marks at the bottom of the page, which we tag as `<fw>` with `@type="catch"` or `@type="sig"`, respectively.

Figure 6. The last two lines of running text, followed by a signature mark and a catchword at the bottom of the page.



(Anonymous [1827/28b], http://www.deutschestextarchiv.de/nn_n0171w1_1828/41).

```
  <p>[…] des vergleichen.</p><lb/>
  <p>Dieſe große Entdeckung trifft merkwürdiger<lb/>
   <fw type="sig" place="bottom">Phyſiſche Erdbeſchreibung <hi rendition="#aq">e</
hi>.</fw> <fw type="catch" place="bottom"><hi rendition="#u">Weiſe</hi></fw><lb/
>[…]</p>
```

4.2 DTABf-M: Manuscript-specific Text Annotation

17 Despite the abovementioned similarities, there are certain features that are specific to manuscripts and which require additional tagging solutions within the DTABf-M. In this section, we will give some examples of features which occur in our manuscript corpus and introduce their structural

representation within the DTABf-M. Subsequently, we will give an overview of those manuscript-specific TEI elements which we have identified so far as necessary to be included in a DTABf extension for manuscripts (DTABf-M). Compared to the DTA corpus of printed texts with its more than 3,000 works,[15] the data basis for the DTABf-M is quite small (129 works).[16] Thus, the number of manuscript-specific elements, attributes, and values illustrated in the examples is not exhaustive and will not fit any given manuscript precisely. However, in comparing the texts of the manuscript corpus it is possible to differentiate common from less common features. Thus, though the DTABf-M tagset might still have to be augmented in the future, the proposed tagging solutions do cover phenomena common in handwritten sources and therefore should be applicable to other manuscripts and for the integration of further handwritten sources into the DTA corpora.

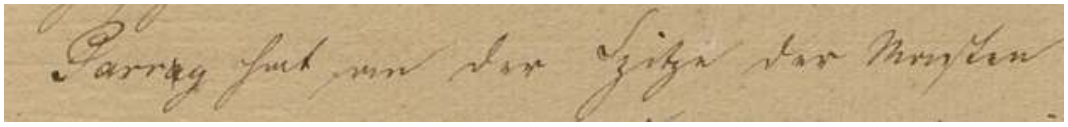
18 The DTABf had to be extended especially with respect to traces of the writing process, such as *ad hoc* corrections by substitution, deletion, or addition of characters, words, or passages, and change of hands or writing devices. These phenomena can be observed in many manuscripts. Manuscript-specific extensions to the DTABf were applied following the principles of simplicity, consistency, and avoidance of ambiguousness, as established for the DTABf. To achieve the latter, DTABf-M specifications were created not only on element and attribute level, but also with regard to attribute values that utilize specified vocabularies for phenomena in manuscripts (e.g., for typical methods of deletion or addition).

4.2.1 Deletions and Additions

19 Common features of manuscripts are corrections to the original text, carried out by adding or deleting textual material. Unlike in printed texts, where manual corrections are usually made on a text which has previously been finalized for the printed publication, corrections in manuscripts are part of the text creation and amelioration process. The TEI elements relevant to these phenomena are "Core Elements for Transcriptional Work" within the TEI Guidelines section on "Representation of Primary Sources" (TEI Consortium 2016, 11; 11.3.1.1). They include `<add>` (for additions) and `<del>` (for deletions), which can be grouped within `<subst>` to represent the substitution of a correct character or phrase for an erroneous one (TEI Consortium 2016, 11.3.1.4). Thus, while it was possible to leave these elements unconsidered for the DTABf for printed material, they became an immensely important part of the DTABf-M.

20 Superfluous textual material in manuscripts may be deleted without substitute. Similarly, characters, words, or phrases missing in a written text may be added with no need to erase text. Thus, the elements `<add>` and `<del>` may be used separately without the element `<subst>`, for instance, where an erroneous character, word, or passage was crossed out (as illustrated in figures 7 and 8) or erased (figure 9), or where missing text was added (figures 10 and 11).

Figure 7. The name "Parry,"[17] initially misspelled as "Parray," was corrected by crossing out the extra "a."



(Anonymous [1827/28b] http://www.deutschestextarchiv.de/nn_n0171w1_1828/291).

```
<p><hi rendition="#aq">Parr<del rendition="#s">a</del>y</hi> hat an der Spitze
der Maften […].</p>
```

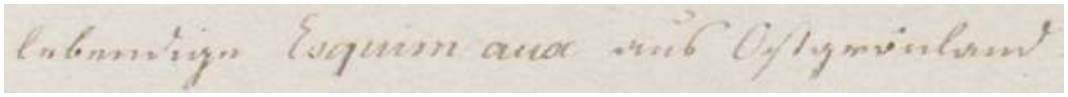
Figure 8. The (apparently) superfluous word "einander" ("each other") was crossed out.[18]

(Parthey [1827/28], http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828/131).

```
[…] die man mit grossen Winden <del rendition="#s">einander</del> näherte […]
```

21 Besides crossing out misspellings or mistakenly-notated characters or words, it is also a common method within manuscripts to erase characters or passages by rubbing or scraping them out (figure 9).

Figure 9. The term initially spelled “Esquimeaux” was altered to the more common spelling “Esquimaux” by erasing the superfluous “e” (which has become almost invisible in the manuscript).[19]

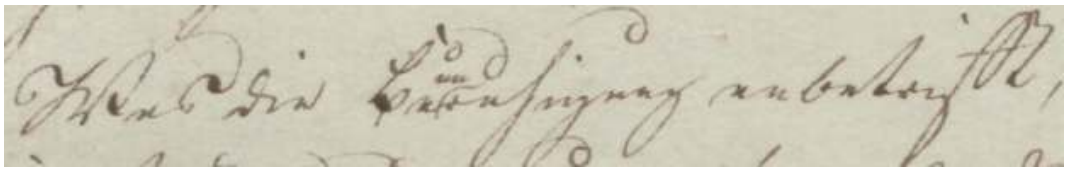


(Anonymous [1828], http://www.deutschestextarchiv.de/nn_msgermqu2124_1827/62).

```
  […] lebendige <hi rendition="#aq">Esquim<del rendition="#erased">e</del>aux</hi>
aus Ostgrönland […]
```

22 Additions of characters or words can be realized within, above, or under the line (figures 10 and 11), or in the right or left margin of a page.

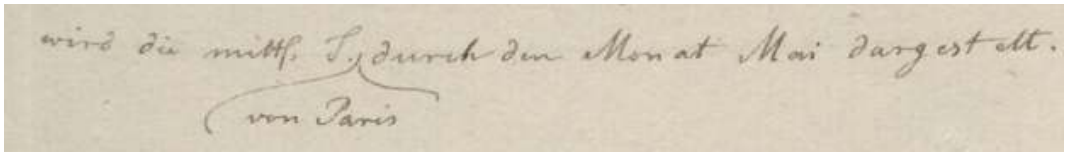
Figure 10. The word initially written, “Beruhigung,” was altered by adding the syllable “un,” changing its meaning to the direct opposite “Be*un*ruhigung” (from “reassurance” to “disturbance”).



(Anonymous [1827/28c], http://www.deutschestextarchiv.de/nn_oktavgfeo79_1828/220).

```
  <p>Was die Be<metamark/><add place="superlinear">un</add>ruhigung anbetrifft,
[…].</p>
```

Figure 11. The words “von Paris” (“of Paris”) are added below the line, to mark clearly that the “mittl[ere] T[emperatur]” (“average temperature”) of this specific place is meant.



(Parthey [1827/28], http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828/619).

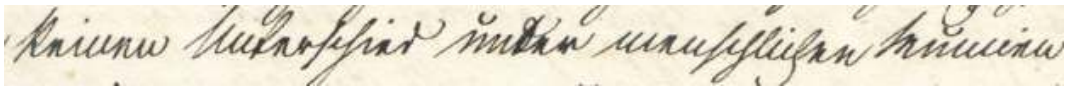
```
  […] wird die mittl. T. <metamark/><add place="sublinear">von Paris </add> durch
den Monat Mai dargestelt.
```

23 The place where added text is meant to be inserted is often marked with some sign or arrow (see figures 10, 11, and 16). To enable the encoding of such signs in the manuscript text we additionally included the TEI element `<metamark>` in the DTABf-M tagset; the abovementioned figures provide transcriptions illustrating the usage of the `<metamark>` element.

4.2.2 Substitutions

24 Deletions and additions may also be parts of a substitution process, where erroneous text was deleted in favor of a correct version which was added to the original text. In such cases, `<add>` and `<del>` are grouped inside a `<subst>` element according to the TEI P5 Guidelines, as shown in figures 12, 13, 14, and 15 (TEI Consortium 2016, 11.3.1.5). Moreover, the following two examples illustrate the necessity of adding a value `@rendition="#ow"` (for "overwritten") to the DTABf-M which can be used in this context within `<del>`.

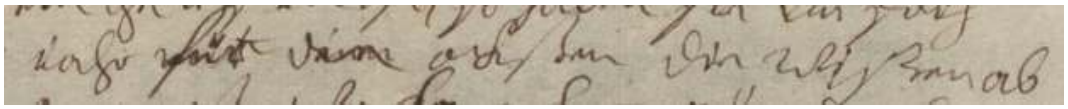
Figure 12. Substitution of the characters "d" and "t" by overwriting the former with the latter.[20]



(Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/28).

```
[…] keinen Unterschied un<subst><del rendition="#ow">d</del><add
place="across">t</add></subst>er menschlichen Mumien […]
```

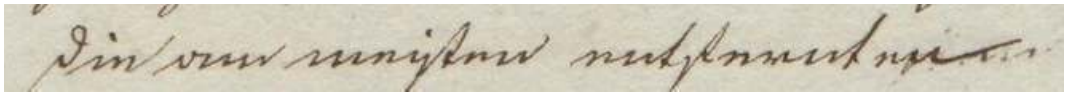
Figure 13. "vor den" (which in the historical New High German language in this case means "for the," indicating a singular form) was overwritten with "für die" ("for the," here indicating plural form).



(Reuß-Ebersdorf 1717, http://www.deutschestextarchiv.de/reuss_paragiatsherrschaftabiv15_1717/4).

```
[…] iahr <subst><del rendition="#ow">vor den</del><add place="across">für die</
add></subst> ockßen die <choice><orig>wißen</orig><reg>wiesen</reg></choice>
ab<lb/>[…]
```

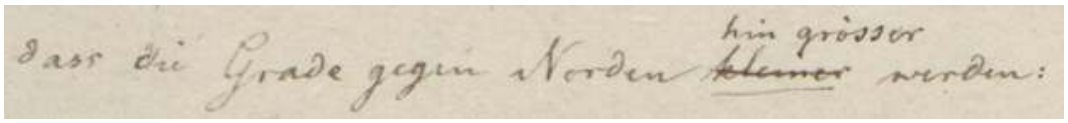
Figure 14. Here, the superlative “entferntesten” (“most distant”) is taken back by scratching out the last four letters and replacing them with “en,” changing the word to “entfernten” (“distant”), obviously because the comparative form was already indicated by the pronoun “meisten” (“the most”) and the additional use of the superlative would have been grammatically wrong.



(Anonymous [1827/28], http://www.deutschestextarchiv.de/nn_msgermqu2345_1827/83).

```
[…] die am meisten entfernte<subst><del rendition="#erased">sten</del><add
place="across">n</add></subst> […]
```

Figure 15. The word “kleiner” (“smaller”), emphasized by an underline, is struck out and replaced by its direct opposite, “(hin) grösser” (“greater”), notated above the line.[21]



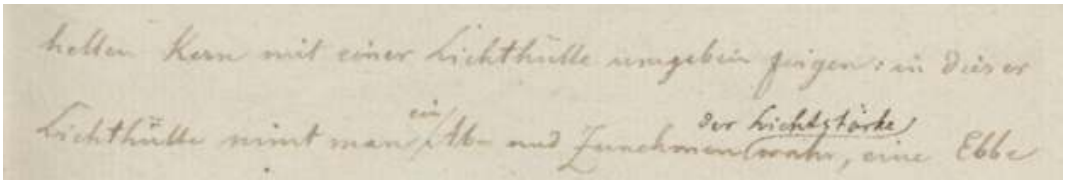
(Parthey [1827/28], http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828/293).

```
[…] dass die Grade gegen Norden <subst><del rendition="#s"><hi
rendition="#u">kleiner</hi></del><add place="superlinear">hin grösser</add></
subst> werden: […]
```

4.2.3 Different Hands or Writing Devices

25 Another manuscript-specific phenomenon is the change of hands and/or writing. To distinguish the change of *hands* (i.e., different scribes) in the course of a writing process from the change of *writing devices* (i.e., the same scribe using, e.g., a pencil instead of their regular ink for certain alterations of the text), we use the `<handNote>` element in the TEI Header.[22] There, an `@xml:id` is assigned to the respective scribe, scribal act,[23] or writing device (in case the scribe cannot be identified). This `@xml:id` is then used as a referencing value of `@hand` in the transcription.

Figure 16. In the second line to be seen in the image scan, the word “ein” (“a”) and the words “der Lichtstärke” (“of the light’s intensity”) have been inserted. Whereas the first addition was written using the same ink and by the same hand as the remainder of the passage, the second addition was clearly written with a different, darker ink, but still by the same hand, i.e., by the same scribe, in this case Gustav Parthey (therefore labelled as `@hand="#Parthey_darkInk"`).[24]



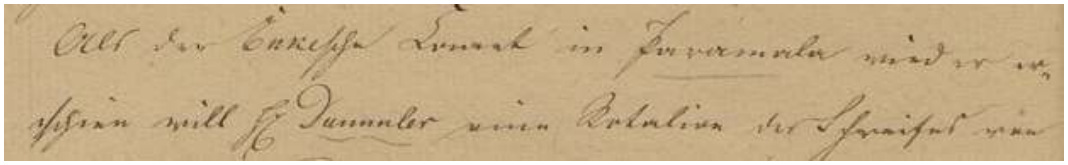
(Parthey [1827/28], http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828/7).

```
  […] hellen Kern mit einer Lichthülle umgeben zeigen: in dieser<lb/>
  Lichthülle nimt man <metamark/><add place="superlinear">ein </add> Ab-
und Zunehmen <metamark/><add place="superlinear" hand="#Parthey_darkInk">der
Lichtstärke </add>wahr, eine Ebbe<lb/>[…]
```

26 The encoding of these phenomena may help differentiate whether certain alterations are *ad hoc* corrections, carried out during the writing process, or later interventions performed when proofreading or commenting on the finished text or the completed draft.

27 Furthermore, we found in our manuscript corpus various occurrences for the phenomenon that words or phrases from one scribe were underlined by another, using a different writing device, as shown in figure 17. Here, the person responsible for the underlining could not be identified. Therefore, the hand has the identifier `"#pencil"`, referring to both the scribe and the writing device.[25]

Figure 17. The main body of the text was written with dark ink and two terms were underlined (here, presumably to mark them as questionable) by a different scribe using a pencil.[26]

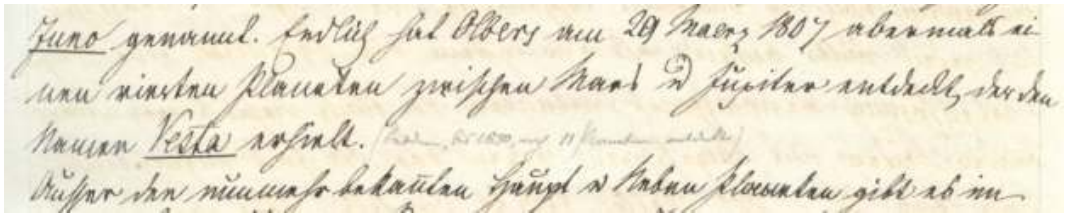


(Anonymous [1827/28b], http://www.deutschestextarchiv.de/nn_n0171w1_1828/216).

```
  Als der <hi rendition="#aq">Enke</hi>ſche Comet in <hi rendition="#u"
hand="#pencil">Paramala</hi> wieder er-<lb/>
  ſchien will <choice><abbr>H </abbr><expan>Herr</expan></choice> <hi
rendition="#aq"><hi rendition="#u" hand="#pencil">Dummler</hi></hi> eine Rotation
des Schweifes von<lb/>[…]
```

28 In the following example (figure 18) it is obvious from the content of the note that this passage was added only in 1850, that is, much later than the original text was written (in 1829).[27] Like the previous example, this one also contains a change of the writing device, from ink to pencil. The addition in this case was, corresponding to its function, tagged as a <note>.

Figure 18. The additional information that between the time of Humboldt's lectures (1827/28) and 1850, eleven more planets had been discovered, was inserted by the same scribe directly into the line, using a pencil (@hand="#Hufeland_pencil").



(Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/19).

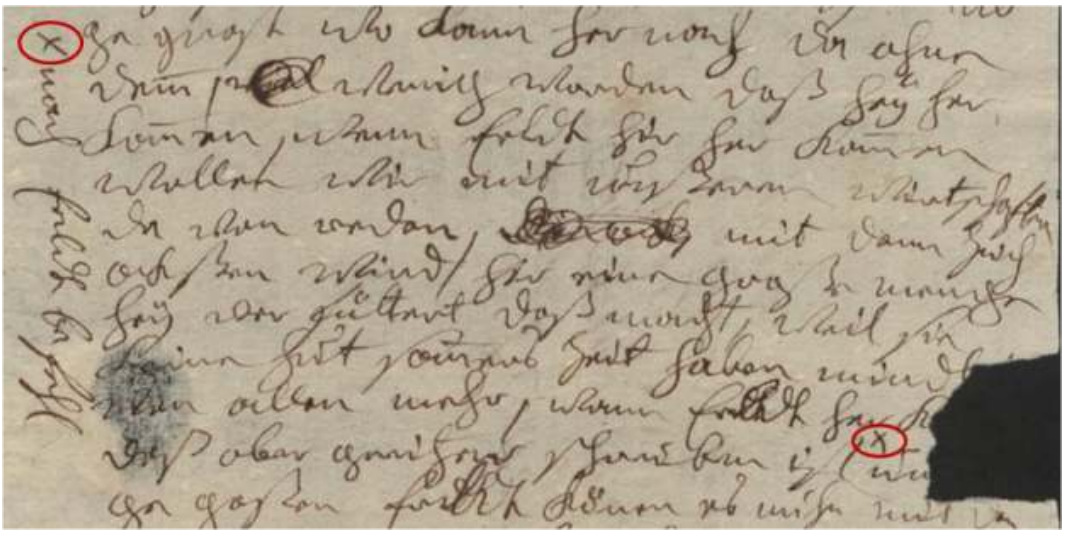
```
  […] Namen <hi rendition="#aq #u">Vesta</hi> erhielt. <note place="mInline"
hand="#Hufeland_pencil">(Seitdem, bis 1850, noch 11 Planeten entdeckt)</note><lb/
>[…]
```

4.2.4 Different Types of Notes

29 Manuscripts—just like printed materials—may contain different types of notes, providing comments on the text or further information about it. While notes at the bottom of a page (footnotes) or at the end of a chapter or the text body (endnotes) are more typical for printed text types than for manuscripts, marginal notes at the right or left margin of a page may occur in both manuscripts and printed texts (figure 19).

30 In addition, notes in manuscripts may occur at several places other than the ones common for printed texts. They may, for example, be inserted inline or at the top or bottom of the text area of a page (figures 19, 20, and 21). We therefore introduced the additional `@place` values `"mInline" | "mTop" | "mBottom"` to the `<note>` element.[28] The usage of `@place="mInline"` within `<note>` was already illustrated in figure 18, while figures 20 and 21 provide examples for the usage of `@place="mBottom"` and `@place="mTop"` within `<note>`.

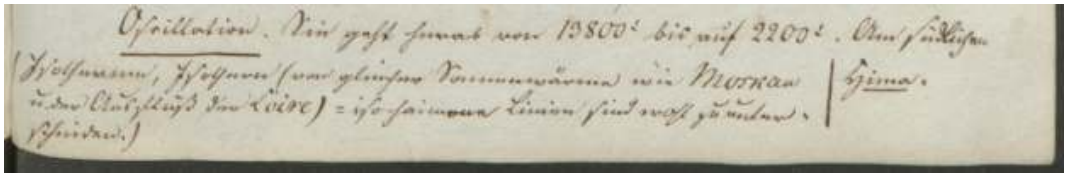
Figure 19. A note has been added at the left-hand side of the manuscript stating that the document in question has been prepared according to the addressee's command.



(Reuß-Ebersdorf 1717, http://www.deutschestextarchiv.de/reuss_paragiatsherrschaftabiv15_1717/4).

```
[…] schreiben ist <note rendition="#v" place="left" n="×">nach Erldt.
<choice><orig>be fehl</orig><reg>befehl</reg></choice><lb/></note>
```

Figure 20. This note, stating that *isotherm*, *isothere*, and *isocheim* must be well differentiated from one another, is recorded at the very bottom of the page. However, it is not referring to a certain point in the text as a footnote would, but is rather a general comment on the topic of the page. Therefore, we use `@place="mBottom"`.



(Anonymous [1827/28a], http://www.deutschestextarchiv.de/nn_msgermqu2345_1827/281).

```
  […]<note place="mBottom">(Isotherme, Isothere (von gleicher Sonnenwärme wie <hi
rendition="#aq">Moskau</hi><lb/>u. der Ausfluß der <hi rendition="#aq">Loire</hi>)
= isochaimone Linien sind wohl zu unter-<lb/>scheiden.)</note><lb/>[…]
```

Figure 21. On the very first page of his text, this scribe made a note in the top right corner, stating that the table of contents for this document is to be found on page 155.

(Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/9).

```
  <note place="mTop"><hi rendition="#right">Die Übersicht des Inhaltes siehe
<ref target="#f0159"><hi rendition="#aq">pag</hi>: 155</ref>. <choice><orig><gap
reason="illegible"/></orig><reg>pp.</reg></choice>
   </hi><lb/></note><lb/>
```

4.3 DTABf-M Metadata Annotation

31 As for metadata, the DTABf already makes quite extensive use of the TEI Guidelines. Thus, most of the metadata information necessary for manuscripts in text corpora was already covered by the existing DTABf metadata tagset. There was only one significant change to make: instead of `<typeDesc>` we added the `<handDesc>` element with its child `<handNote>`.[29] The `@xml:id` in `<handNote>` identifies the hand or scribal act described and can be referred to from the `@hand` attribute within the document. The writing device of each hand or scribal act is specified within a `@medium` attribute.

4.4 Overview: DTABf Extension for Manuscripts

32 In section 4.2 we presented some examples of typical structural and textual phenomena in manuscripts. Table 1 provides an overview of the elements, attributes, and values which were introduced to the DTABf for manuscript encoding and of how they are related to already existing DTABf tags.

33 The proposed format, DTABf-M, is still a work in progress and will be subject to continuous development based on further manuscripts added to the DTA corpus. Development will occur in a manner similar to the approach used with the DTABf for printed texts: extensions will be performed cautiously, in a restrictive and minimalistic manner, and based on actual phenomena observed in the historical text sources (Haaf, Geyken, and Wiegand 2014–15, §60–61).

Table 1. Manuscript-specific additions to the DTABf; * = new; + = new in this context.

Elements	Attributes	Values
Text		
*<add>	+@place	*"superlinear" | *"intralinear" | *"across" | *"sublinear" | "left" | "right"
	*@hand	e.g., "#JaneDoe", "#JohnDoe"
<corr> | <expan> | <reg> | <supplied> | <unclear>	*@resp	
*<del>	+@rendition	*"#ow" | *"#s" | *"#erased"
	*@hand	e.g., "#JaneDoe", "#JohnDoe"
	+@cert	"low" | "high"
	@type	*"session"
<figure>	@type	*"stamp"

<fw>	@type	*"folNum"
<head>	+@type	*"leftMargin" | *"rightMargin"
*<metamark>		
<note>	+@place	*"mTop" | *"mBottom" | "mInline"
	*@hand	
<space>	+@unit	"chars" | "lines" | "pages" | "words"
	+@quantity	data.count
*<subst>		
<supplied>	+@reason	*"covered" | *"damage"
*<unclear>	+@reason	"illegible" | *"covered"
Metadata		
<bibl>	@type	*"MAN" (i.e., manuscript)
*<handDesc>		
*<handNote>	+@xml:id	
	*@medium	*"pencil" | *"ink" | *"altInk"

5. Inclusion of the DTABf-M in the DTABf System

5.1 Consideration of Modularization Possibilities

34 So far, when dealing with text type–specific extensions to the DTABf (cf. Haaf and Schulz 2014), the adaptations have been applied directly to the DTABf schema. Until now, this approach has proved feasible since the number of necessary additions to the DTABf subset remained small. Additions that were necessary occurred mainly at value or attribute level and in most cases did not run the risk of generating ambiguities in the format or evoking uncertainties concerning proper application. The practice of extensive documentation, of distinguishing four different annotation levels (required, recommended, optional, proscribed), and of introducing additional validation via Schematron constraints as well as of marking new values as affiliated to certain text types[30] (cf. Haaf, Geyken, and Wiegand 2014–15), was sufficient to incorporate these respective extensions into the DTABf without loss of homogeneity and consistency.

35 For manuscripts, however, as shown above, rather extensive additions to the DTABf became necessary (see table 1), increasing the risk of ambiguities and uncertainties in applying the format. We want to prevent some of these insecurities by providing detailed documentation for DTABf-M–conformant tagging based on the DTABf documentation.[31] However, an additional technical solution on the schema level has proved desirable as well.

36 The maintenance of two separate TEI customizations (one for DTABf, the other for DTABf-M) was to be avoided for reasons of efficiency as well as to prevent workflow errors. In fact, we only recently gave up the maintenance of separate customizations and schemas for said reasons and introduced Schematron constraints to the format instead (see Haaf, Geyken, and Wiegand 2014–15, §25). So, for manuscripts, one possible method would have been the introduction of new Schematron constraints to our existing set of rules.[32] However, Schematron validation is relatively slow and each new rule increases validation runtime. Given the significant number of differences between manuscripts and prints, the set of rules would have grown substantially, this way increasing the effect of slow validation.

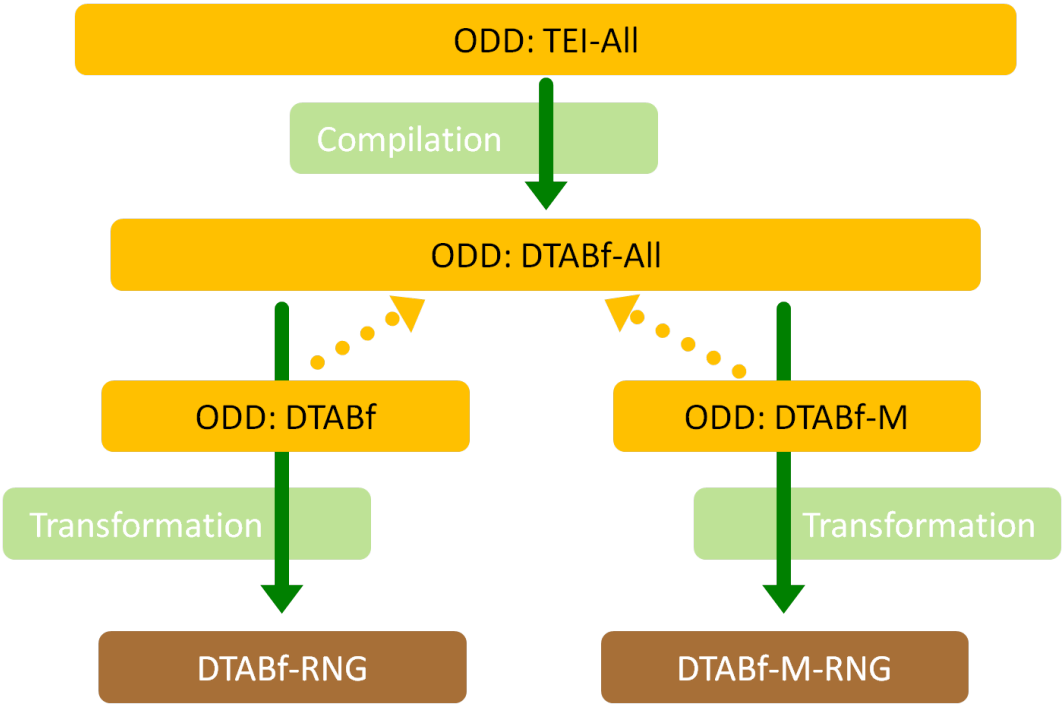
37 Hence, we wanted to generate different RNG schemas for quick validation, but without having to create different separate TEI customizations. Fortunately, the TEI infrastructure already offers a technique for this case: the method of chaining ODDs (Rahtz and Burnard 2014, 16–17), where one creates ODDs which are based on one another, but which can then be transformed into different independent schemas.

5.2 Modular DTABf System of Chained ODDs[33]

38 The concept of chaining ODDs is the creation of a large, comprehensive ODD (here: `ODD-L`) that considers all possible annotations within the whole DTABf system as a derivative of the `TEI_all` ODD and is compiled with regard to `TEI_all`. After the larger ODD is in place, additional smaller ODDs are created (here: `ODD-S1`, `ODD-S2`, … `ODD-Sn`) that determine reductions to `ODD-L` and thus represent subsets of the ODD-L tagset. Schemas can then be created by transformation of an `ODD-S` which is based on a compiled version of a comprehensive source `ODD-L` (expressed as a URI within the `@source` attribute of `<schemaSpec>`).[34] This way, we avoid creating similar parallel customizations that have to be maintained separately. Instead, each feature needed in the DTABf system is described only once, in one specific place, and then reused in the other relevant contexts.

39 More concretely, we created a comprehensive DTABf ODD containing the entire DTABf tagset (`DTABf-All`) and representing a true subset of the TEI_all tagset. Two small ODDs, dependent on `DTABf-All`, specify the encoding of prints or manuscripts: first, the (established) DTABf for printed texts (`DTABf`), and second, the (new) DTABf for manuscripts (`DTABf-M`). Figure 22 illustrates the system.

Figure 22. Chaining ODDs for the harmonization of DTABf and DTABf-M.



5.3 Examples

40 The following section presents some examples for the chaining method at different layers of the DTABf ODDs.

5.3.1 Inclusion and Exclusion of Elements

41 The `DTABf-All` ODD includes the elements `<subst>` (module `transcr`), `<add>` (module `core`), and `<del>` (module `core`) for the tagging of revisions and corrections in the handwritten text.

Example 1. ODD for DTABf-All:[35]

```
<moduleRef key="transcr" include="fw metamark subst ..."/>
<moduleRef key="core" include="add del list p sp ..."/>
```

42 These elements are not part of the DTABf for printed texts, so the `DTABf` ODD excludes these elements.

Example 2. ODD for DTABf:

```
<moduleRef key="transcr" except="metamark subst"/>
<moduleRef key="core" except="add del ..."/>
```

43 The `DTABf-M` ODD, however, includes these elements by referencing the entire modules from the `DTABf-All` ODD. All elements of all the modules included in `DTABf-All` will automatically be included here as well, unless explicitly excluded.

Example 3. ODD for DTABf-M:

```
<moduleRef key="transcr"/>
<moduleRef key="core" except="sp ..."/>
```

5.3.2 Inclusion and Exclusion of Classes

44 For DTABf-M, we make use of the new TEI class `att.written`, providing us with the attribute `@hand` to encode a change of hands within the source text. For printed materials we do not need this attribute.[36] Hence, the class `att.written` is included in the `DTABf-All` ODD by inclusion of the class `att.transcriptional`, of which `att.written` is a sub-class.

Example 4. ODD for DTABf-All:[37]

```
<classSpec ident="att.transcriptional" mode="change" type="atts">
 <attList>
  <attDef ident="status" mode="delete"/>
  <attDef ident="cause" mode="delete"/>
  <attDef ident="seq" mode="delete"/>
 </attList>
</classSpec>
```

45 The class `att.written` is also tacitly included in the `DTABf-M` ODD by inheritance from the `DTABf-All` ODD, whereas the `DTABf` ODD explicitly excludes this class.

Example 5. ODD for DTABf:

```
<classSpec ident="att.written" module="tei" type="atts" mode="delete"/>
```

5.3.3 Changes within Classes

46 For the `@rendition` attribute of class `att.global.rendition`, a fixed list of possible values is determined within the DTABf. For manuscripts, we had to add two values to this list: `"#mPrint"` for printed material in manuscript text (e.g., boilerplate material) and `"#mRetrace"` for retracing of characters in order to highlight them.

47 `DTABf-All` contains all possible values of `@rendition`.

Example 6. ODD for DTABf-All:[38]

```
<classSpec ident="att.global.rendition" module="tei" type="atts" mode="change">
 <attList>
  <attDef ident="rendition" mode="change" usage="opt">
   <valList type="closed" mode="replace">
    <valItem ident="#c"/>
    <valItem ident="#b"/>
    <valItem ident="#i"/>
    [...]
    <valItem ident="#mPrint"/>
    <valItem ident="#mRetrace"/>
   </valList>
  </attDef>
 [...]
 </attList>
</classSpec>
```

48 `DTABf-M` does not change that list. `DTABf`, however, deletes the named values from the list.

Example 7. ODD for DTABf:

```
<classSpec ident="att.global.rendition" module="tei" type="atts" mode="change">
 <attList>
  <attDef ident="rendition" mode="change" usage="opt">
   <valList type="closed" mode="change">
    <valItem ident="#mRetrace" mode="delete"/>
    <valItem ident="#mPrint" mode="delete"/>
   </valList>
  </attDef>
 </attList>
</classSpec>
```

5.3.4 Changes of Element Features

49 In manuscripts the title of a chapter or section is sometimes written on the right or left margin of a page. Those instances cannot be considered marginal notes, but represent a type of heading. Therefore, we introduced `@type="rightMargin"|"leftMargin"` in `<head>`. Since this is a phenomenon we haven't (yet) encountered in printed texts, the `@type` attribute in `<head>` is only needed for manuscript annotation.[39]

50 To address this phenomenon, the element `<head>` is provided with the necessary attribute-value-pairs within its element specification in `DTABf-All`:

Example 8. ODD for DTABf-All:

```
<elementSpec ident="head" module="core" mode="change">
 <attList>
  <attDef ident="type" mode="change" usage="opt">
   <valList type="closed" mode="replace">
    <valItem ident="rightMargin"/>
    <valItem ident="leftMargin"/>
   </valList>
  </attDef>
 [...]
 </attList>
</elementSpec>
```

51 Again, within the DTABf-M ODD things are left as they are. In the DTABf ODD, however, @type is deleted from <head>:

Example 9. ODD for DTABf:

```
<elementSpec ident="head" module="core" mode="change">
 <attList>
  <attDef ident="type" mode="delete"/>
 </attList>
</elementSpec>
```

52 These examples illustrate several ways in which ODD chaining is used within the DTABf system. While the DTABf-All ODD is quite extensive, the ODDs for DTABf and DTABf-M could be kept quite narrow and relatively simple. Redundancies are generally avoided with this method. Thus, maintenance of the ODDs is easily manageable.

6. Conclusion and Further Prospects

53 The DTA project is an example of the application of the TEI Guidelines to large-scale corpora. Our primary goal is to be as inclusive as possible, allowing for other projects to benefit from our resources (i.e., our comprehensive guidelines and documentation as well as the technical infrastructure that includes Schemas, ODDs, and XSLT scripts) and contribute to our corpora. We also want to ensure interoperability of all data within the DTA corpora. The underlying TEI format has to be continuously maintained and adapted to new necessities with these two premises in mind.

54 In this paper, we presented the workflow implemented at the DTA for the adaptation of the DTABf for manuscripts. The challenge here was twofold: On the one hand, we wanted to identify frequent structural phenomena one might want to annotate within manuscripts, apply DTABf tagging as far as possible, and provide new markup solutions where the DTABf was too limited. On the other hand, in case adaptations are needed, we have tried to keep the format as unambiguous, robust, and user-friendly as possible in order to ensure truly interoperable texts. An important issue is how to include this new format in the existing DTABf system. To address this, the approach of chaining ODDs has proved practicable and appropriate.

55 Interoperability issues are of great interest not only to the DTA but also to large infrastructure projects such as CLARIN,[40] where corpora are being combined, exchanged, and made available in new contexts. The DTABf and its extensions are meant to address these issues.

BIBLIOGRAPHY

Literature

CLARIN-D AP 5. 2012. "CLARIN-D User Guide." Version 1.0.1. http://www.clarin-d.de/de/hilfe/benutzerhandbuch.

Deutsche Forschungsgemeinschaft (DFG). 2015. "Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora." http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf.

———. 2015. "Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft." http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/foerderkriterien_editionen_literaturwissenschaft.pdf.

Gansel, Christina. 2011. *Textsortenlinguistik*. Göttingen: Vandenhoeck & Ruprecht.

Geyken, Alexander, Susanne Haaf, and Frank Wiegand. 2012. "The DTA 'base format': A TEI-Subset for the Compilation of Interoperable Corpora." In *Proceedings of the 11th Conference on Natural Language Processing (KONVENS)*, edited by Jeremy Jancsary, 383–91. Vienna: Österreichische Gesellschaft für Artificial Intelligence. http://www.oegai.at/konvens2012/proceedings/57_geyken12w/57_geyken12w.pdf.

Haaf, Susanne, Alexander Geyken, and Frank Wiegand. 2014–15. "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* 8. http://jtei.revues.org/1114. doi:10.4000/jtei.1114.

Haaf, Susanne, Frank Wiegand, and Alexander Geyken. 2013. "Measuring the Correctness of Double-Keying: Error Classification and Quality Control in a Large Corpus of TEI-Annotated Historical Text." *Journal of the Text Encoding Initiative* 4. http://jtei.revues.org/739. doi:10.4000/jtei.739.

Haaf, Susanne, and Matthias Schulz. 2014. "Historical Newspapers & Journals for the DTA." In *Proceedings of the LRT4HDA Workshop, Held at the 9th LREC Conference*, edited by Kristín Bjarnadóttir, Mathew Driscoll, Steven Krauwer, Stelios Piperidis, Cristina Vertan, and Martin Wynne, 50–54. N.p.: European Language Resources Association. http://www.lrec-conf.org/proceedings/lrec2014/workshops/LREC2014Workshop-LRT4HDA%20Proceedings.pdf#page=57.

Rahtz, Sebastian, and Lou Burnard. 2014. "Advanced Topics in ODD." In *ODD: One Document Does it All*. Workshop at the Text Encoding Initiative Conference and Members Meeting, Oct 22–24 Evanston, IL. http://tei.it.ox.ac.uk/Talks/2014-10-odds/talk-05-advanced.xml.

Thomas, Christian. 2015. "Hidden Kosmos—Humboldts 'Kosmos-Vorträge' als Probe der Digital Humanities." Presentation at the DHd Annual Conference "Von Daten zu Erkenntnissen: Digitale Geisteswissenschaften als Mittler zwischen Information und Interpretation," February 23–27, Graz, Austria. https://www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/media/c-thomas-dhd-graz-paper-hidden-kosmos-20150126.pdf.

Thomas, Christian, Benjamin Fiechter, and Marius Hug. 2016. "Methoden und Ziele der Erschließung handschriftlicher Quellen zu Alexander von Humboldts Kosmos-Vorträgen: Das Projekt Hidden Kosmos der Humboldt-Universität zu Berlin." In *Horizonte der Humboldt-Forschung: Natur, Kultur, Schreiben*, edited by Ottmar Ette and Julian Drews, 287–318. Hildesheim: Georg Olms.

Thomas, Christian, and Frank Wiegand. 2015. "Making Great Work Even Better: Appraisal and Digital Curation of Widely Dispersed Electronic Textual Resources (c. 15th–19th centuries) in CLARIN-D." In *Historical Corpora: Challenges and Perspectives*, edited by Jost Gippert and Ralf Gehrke, 181–96. Tübingen: Narr.

TEI Consortium. 2016. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*. Version 3.0.0. Last updated March 29. N.p.: TEI Consortium. http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/.

Historical Sources

Andreä, Johann Valentin. 1616. "Chymische Hochzeit Christiani Rosencreutz Anno 1459." Straßburg. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/valentin_hochzeit_1616.

Anonymous. [1827/28a]. "Alexander von Humboldts Vorlesungen über phÿsikalische Geographie nebst Prolegomenen über die Stellung der Gestirne. Berlin im Winter von 1827 bis 1828." [Berlin]. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/nn_msgermqu2345_1827.

———. 1827/28b]. "Physikalische Geographie von Heinr. Alex. Freiherr v. Humboldt. [V]orgetragen im Wintersemester 1827/8." [Berlin]. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/nn_n0171w1_1828.

———. 1827/28c]. "Die physikalische Geographie von Herrn Alexander v. Humboldt, vorgetragen im Semestre 1827/28." [Berlin]. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/nn_oktavgfeo79_1828.

———. 1828]. "Physikalische Geographie. Vorgetragen von Alexander von Humboldt." [Berlin]. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/nn_msgermqu2124_1827.

Hufeland, Otto. [Ca. 1829]. "Vorlesungen über physicalische Geographie von A. v. Humboldt. [G]eschrieben im Sommer 1829 durch Otto Hufeland." [Berlin]. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829.

Parthey, Gustav. [1827/28]. "Alexander von Humboldt[:] Vorlesungen über physikalische Geographie. Novmbr. 1827 bis April,[!] 1828. Nachgeschrieben von G. Partheÿ." [Berlin]. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/parthey_msgermqu1711_1828.

Reuß-Ebersdorf, Erdmuthe Benigna von. 1717. *[Letter to Heinrich XXIV. (Reuß-Köstritz); Ebersdorf (Thuringia, Germany), October 8, 1717]*. Ebersdorf. In *Deutsches Textarchiv*. http://www.deutschestextarchiv.de/reuss_paragiatsherrschaftabiv15_1717.

NOTES

1 *Deutsches Textarchiv: Grundlage für ein Referenzkorpus der neuhochdeutschen Sprache*, Berlin-Brandenburg Academy of Sciences and Humanities, accessed January 28, 2017, http://www.deutschestextarchiv.de/.

2 See *DTA-Basisformat: TEI-Format für die Auszeichnung historischer Texte*, Berlin-Brandenburg Academy of Sciences and Humanities, accessed January 28, 2017, http://www.deutschestextarchiv.de/doku/basisformat.

3 *Hamburger Schlüsseldokumente zur deutsch-jüdischen Geschichte: Eine Online-Quellenedition*, Institut für die Geschichte der deutschen Juden, accessed January 28, 2017, http://juedische-geschichte-online.net/projekt/edition/.

4 See *ePoetics: Korpuserschließung und Visualisierung deutschsprachiger Poetiken (1770–1960) für den "Algorithmic criticism"*, Technische Universität Darmstadt, accessed January 28, 2017, http://www.epoetics.de/.

5 See *DTA-Basisformat für Manuskripte*, Berlin-Brandenburg Academy of Sciences and Humanities, accessed January 28, 2017, http://www.deutschestextarchiv.de/doku/basisformat/manuskript.html.

6 See "Software im Deutschen Textarchiv, 8: Linguistische Analyse historischer Texte" (Linguistic Analysis for Historical Texts) (CAB), accessed January 28, 2017, http://www.deutschestextarchiv.de/doku/software#cab.

7 See "DTAQ: Kollaborative Qualitätssicherung im Deutschen Textarchiv" (Collaborative Quality Assurance within the DTA), accessed January 28, 2017, http://www.deutschestextarchiv.de/dtaq/about. On the process of quality assurance in the DTA, see, for example, Haaf, Wiegand, and Geyken 2013.

8 See Gansel 2011, 53, for a reflection on the term "Kerntextsorte" (i.e., core text type).

9 These premises lead to the method of cautious extensions to the existing DTABf rather than the adoption of a new format within the DTA. Of course, there are other TEI formats for manuscript tagging, varying with regard to their scope, premises, and character of specifications and documentation. These include (to only name a few) the guidelines of the project *Briefe und Texte aus dem intellektuellen Berlin um 1800* (*Letters and Texts: Intellectual Berlin around 1800*; "Edition-specific TEI Encoding Guidelines," accessed February 12, 2016, http://www.berliner-intellektuelle.eu/encoding-guidelines.pdf), of the Shelley-Godwin Archive ("Encoding the S-GA," accessed January 28, 2017, http://shelleygodwinarchive.org/about/#encodingthesga), and of the digital edition *Theodor Fontane: Notizbücher* (*Theodor Fontane's Notebooks*; accessed January 28, 2017, https://fontane-nb.dariah.eu/doku.html?id=gesamtdokumentation).

10 *Hidden Kosmos—Reconstructing Alexander von Humboldt's "Kosmos-Lectures,"* accessed January 28, 2017, https://www.culture.hu-berlin.de/de/forschung/projekte/hidden-kosmos/. For further information on the project's aims and methods, and on the cooperation between DTA and Hidden Kosmos, see, for example, Thomas 2015 and Thomas, Fiechter, and Hug 2016.

11 For the ongoing publication of the Hidden Kosmos subcorpus, see the DTA search page, AvHKV subcorpus, accessed July 13, 2017, http://www.deutschestextarchiv.de/search/metadata?corpus=avhkv.

12 German Title: *Alexander von Humboldt auf Reisen—Wissenschaft aus der Bewegung*; project homepage, accessed January 28, 2017, http://www.bbaw.de/en/research/avh-r.

13 *Digitale Edition der Briefe Erdmuthe Benignas von Reuß-Ebersdorf (Digital Edition of Erdmuthe Benigna of Reuß-Ebersdorf's letters)*, project at the Department of History at Friedrich-Schiller-University Jena, accessed January 28, 2017, http://www.histinst.uni-jena.de/Bereiche/Geschlechtergeschichte/Projekte/Digitale+Edition+der+Briefe+Erdmuthe+Benignas+von+Reuß_Ebersdorf+.html.

14 Note that the snippets provided below the figures within the following examples illustrate the DTABf-M-conformant transcription and annotation of the text passage under consideration. The URL provided in each example leads to the page of the document from which the example was taken. The tagging is simplified in some instances by leaving out annotations that are not central to the purpose of this paper. As described in the introduction, some manuscripts are still being processed for quality assurance, and therefore only available in DTAQ, where registration and password are required to access the documents.

15 Approx. 670,000 pages as of December 2016.

16 Approx. 4,500 pages as of December 2016.

17 Referring to the English naval officer and Arctic explorer William Edward Parry (1790–1855); for an authority file reference, see http://d-nb.info/gnd/116048166 (accessed January 30, 2017). Parry is one of the persons most frequently mentioned in the course of Humboldt's Kosmos-Lectures: see the overview on personal names retrieved from the attendee's lecture notes of the two courses at the Berlin University and the Sing-Akademie building, accessed January 30, 2017, http://deutschestextarchiv.de/kosmos/person.

18 Translation: "[...] that were brought near to each other with big winches [...]."

19 Translation: "[...] living Eskimos [i.e., Inuit] from Eastern Greenland [...]."

20 Translation: "[...] no difference between human mummies [...]." Either the writer initially set out to write "und" ("and") instead of "unter" ("under"), then corrected his mistake immediately by substituting a "t" for the "d" and continuing to complete the word as "unter," or he initially misspelled the German word "unter" as "under," and fixed his mistake after the word was written.

21 Translation: "[...] that the grades [of longitude/latitude] become {*smaller*greater} northwards: [...]."

22 TEI Consortium 2016, http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-handNote.html.

23 That is, different `<handNote>` elements can represent the same scribe using different writing devices.

24 Translation: "[...] show a bright center surrounded by an illuminated shell: in this illuminated shell an increasing and decreasing of the light's intensity can be observed, an ebb [...]."

25 This consistent usage of `@hand` within `<hi>` and `<note>` and several other elements was made possible only recently (March 29, 2016) with the TEI P5 release 3.0.0: see TEI Consortium 2016, "TEI P5 version 3.0.0 release notes," http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/readme-3.0.0.html.

26 Translation: "By the time Encke's Comet reappeared in Paramala [i.e., Paramatta, then a settlement of its own, today a suburb of Sydney], Mister Dummler [probably referring to James Dunlop (1793–1848; see the authority record, accessed January 30, 2017, http://d-nb.info/gnd/102456828)] claims to have observed a rotation of the comet tail of [...]."

27 See the title page of the document from which the example was taken, Hufeland [ca. 1829], http://www.deutschestextarchiv.de/hufeland_privatbesitz_1829/7: "geschrieben im Sommer 1829 durch Otto Hufeland" ("written in the summer of 1829 by Otto Hufeland"). Otto Hufeland, presumably born in 1806, was a son of Gottlieb Hufeland (1760–1817; see the authority record, accessed January 30, 2017, at http://d-nb.info/gnd/117053961), a relative of the well-known German physician Christoph Wilhelm Hufeland (1762–1836; see the authority record, accessed January 30, 2017, http://d-nb.info/gnd/118554514). His manuscript documenting the Kosmos-Lectures at the Sing-Akademie was written in the year after their closure in 1828 and contains several later additions. It is not an original account of the lectures, but a copy of another volume of lecture notes (cf. Anonymous [1828], http://www.deutschestextarchiv.de/nn_msgermqu2124_1827).

28 Note that the method of commencing a value with a letter (combination) indicating the types of text where that value may be used has already been pursued in the context of DTABf extensions for other text types, e.g., funeral sermons (`"fs..."`) and journals/newspapers (`"j..."`) (see Geyken, Haaf, and Wiegand 2012, 390; Haaf and Schulz 2014, 53). We use `"m..."` in this case to ensure that the `@place` values used for manuscripts are not confused with the DTABf-conformant `@place` values for notes in printed texts.

29 Beyond that, there is only one further small adaptation of the metadata concerning the value list of `@type` within `<bibl>`: it is extended by the document type `"MAN"` (i.e., manuscript). For documentation on the specifics of DTABf-M metadata annotation, see the DTABf documentation section on manuscript metadata, accessed July 12, 2017, http://www.deutschestextarchiv.de/doku/basisformat/msMetadata.html.

30 E.g., by commencing the values with identifiers for certain text types; see note 28 for details.

31 See "DTA-Basisformat—Auszeichnung von Manuskripten" ("DTA Base Format—Annotation of Manuscripts"), accessed July 13, 2017, http://www.deutschestextarchiv.de/doku/basisformat/manuskript.html.

32 For the DTABf Schematron rule set, see "Schematron Extension of the DTA 'Base Format'," accessed July 13, 2017, http://www.deutschestextarchiv.de/basisformat.sch.

33 Thanks to James Cummings (Academic IT Services, University of Oxford) for very helpful advice concerning this method.

34 See Rahtz and Burnard 2014, 16–17, and http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/TD.html#TDbuild for more information on the implementation of this system.

35 Note that for reasons of clarity in these examples the `<moduleRef>`s only contain a small extract of the element set included by the `<moduleRef>`s within the original DTABf ODDs.

36 According to the DTA guidelines, corrections and notes inserted into printed texts manually are not considered for transcription, which means that there is no necessity for `@hand` in printed texts of the DTA corpora.

37 Note that, due to class inheritance, `att.written` is included in the schema by including `att.transcriptional` and by *not excluding it* explicitly.

38 The values `"#c"`, `"#b"`, and `"#i"` represent centered text, bold typeface, and italics, respectively.

39 With regard to the overall consistency, the usage of `@place` instead of `@type` might seem more appropriate for the current scenario. However, we decided to use `@type` in order to remain compliant with the TEI Guidelines, which do not provide `@place` for `<head>`: see TEI Consortium 2016, http://www.tei-c.org/Vault/P5/3.0.0/doc/tei-p5-doc/en/html/ref-head.html.

40 CLARIN: project Common Language Resources and Technology Infrastructure, http://clarin.eu/; German partner CLARIN-D, http://www.clarin-d.de/.

ABSTRACT

This paper presents work in progress on the DTA "Base Format" for Manuscripts (DTABf-M), an extension to the DTA "Base Format" (DTABf) for the TEI-conformant annotation of manuscripts. The DTABf is a TEI-subset for the consistent, yet unambiguous, annotation of large amounts of historical text. During our work on the

DTA corpora, the DTABf has continuously been subject to further adaptations to specific annotation needs. The latest addition, the DTABf-M, contains elements, attributes, and values necessary for the annotation of (historical) handwritten documents. The goal is to provide a TEI format for diverse manuscripts in large text corpora.

While the DTABf covers a wide range of phenomena found not only in printed texts but also in manuscripts, there are certain manuscript-specific features which have to be additionally represented by the DTABf-M. There are several prerequisites for DTABf-M to be suitable for the DTA and its workflows and processes: First, it should be based on the original DTABf tagset, and only extend it if unavoidable. Second, like the DTABf, the DTABf-M should be created in a bottom-up approach, that is, based on actual phenomena found in handwritten texts which are transcribed and encoded using the DTABf. Third, the format should complement the DTABf, not replace it. Hence, it is necessary to find a modular way of integrating the DTABf-M into the DTABf. This paper describes how we deal with these issues in the process of developing the DTABf-M.

INDEX

Keywords: annotation of manuscripts, TEI corpora, chaining ODDs, interoperability, interchange, TEI customization, standardization

AUTHORS

SUSANNE HAAF

Susanne Haaf studied German philology and computational linguistics (M.A.) at the universities of Heidelberg and Zürich. From 2007 to 2010 she worked as a research assistant for the edition project Martin Bucers Deutsche Schriften at the Heidelberg Academy of Sciences and Humanities. Since 2010 she has been working as a research assistant at the Berlin-Brandenburg Academy of Sciences and Humanities for the projects Deutsches Textarchiv (DTA) and CLARIN-D.

CHRISTIAN THOMAS

Christian Thomas has studied German literature and philosophy at the Humboldt-Universität zu Berlin. Since 2010 he has been working as a research assistant at the Berlin-Brandenburg Academy of Sciences and Humanities for the projects Deutsches Textarchiv (DTA) and CLARIN-D. Additionally, from 2014 until 2016 he has been working as a research assistant in the project Hidden Kosmos at Berlin's Humboldt-University.