Deutsches Textarchiv: TEI-XML anhand des DTA-Basisformats

Susanne Haaf, Alexander Geyken

Deutsches Textarchiv & CLARIN-D

@Berlin-Brandenburgische Akademie der Wissenschaften

Gliederung

- Projekthintergrund
  - Über das DTA
  - Das DTA in der Community
- Problemstellung: viele Texte – ein Format
- Über das DTABf
  - Komponenten, Tagset und Levels
  - Modularisierung
  - Korpusanalyse (Beispiele)
- Ausblick

Projekthintergrund

Deutsches Textarchiv (DTA)

- Historische Werke verschiedenster Textsorten (ca. 1600–1900)
- 3,036 Werke (> 670.000 Seiten, > 162 Mio. Tokens)
- Manuelle Transkription & TEI-Annotation (Double Keying)
- Hohe Textqualität

www.deutschestextarchiv.de

Deutsches Textarchiv (DTA)

- Automatische linguistische Analyse: Tokens, Lemmas, POS, Orthographie
- Zugriff und Korpusanalyse über die DTA-Platform
- Freier Download (CC-Lizenzen)
- Anbindung an CLARINs linguistische Tools

Suchanfragen aufgrund linguistischer Information

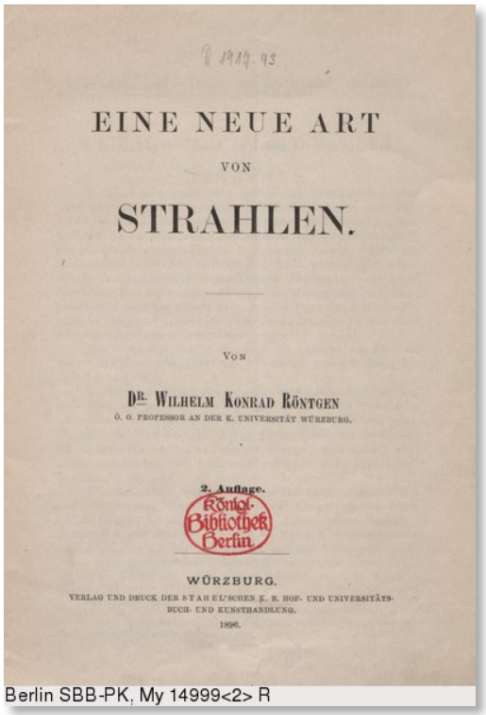
Beispiel: Suche nach Nomen + eine Form des Hilfsverbs „sein“ + „teuer“ als prädikatives Adjektiv

Neue Suche · Ganze Sätze: "$p=NN ist with $p=VAFIN teuer with $p=ADJD" #less_by_date #ha suchen Hilf

10 Treffer pro Seite Sortierung: Datum aufsteigend/absteigend · zufällig

gehe zu: Anfang · -10 · -5 · vorherige · nächste · +5 · +10 · Ende

1:	[boettiger_predigt_1652:37]	Denn ein	**Christen-Mensch**	**ist theuer** für GOtt geachtet/ und sein …
2:	[glauber_opera02_1659:386]	Gesetzt das	**Korn**	**were thewrer**/ die Säfte kosteten mehr an …
3:	[dannhauer_catechismus08_1666:1101]	… sondern nach dem Wort Gottes/ das Wort	**Gottes**	**ist theuer** worden/ darnach man geheulet wie …
4:	[dannhauer_catechismus09_1672:141]	Solche	**Hirten**	**seind theur**/ (schreibet Luth. Tom. VII …
5:	[sandrart_academie0101_1675:26]		**Mahlerey-Arbeit**	**ist theurer**/ als Statuen.
6:	[sandrart_academie0103_1675:205]	2* Ihre	**Arbeit**	**ist theurer**/ als Statuen.
7:	[sandrart_iconologia_1680:269]	…/ hell' Augen/ ein Gemüt/ dem	**redlich-seyn**	**ist theur**.
8:	[frentzel_schauplatz_1744:306]	… wegen Mangel des Futters grosse Noth, das	**Futter**	**war theuer** und noch nicht einmahl ums Geld …
9:	[suppius_oden_1749:179]	Mein Glücke friert! das	**Holz**	**ist theuer**, Jch habe keinen Kreuzer Geld …
10:	[rochow_versuch_1772:51]	Das	**Holz**	**ist theuer**.
11:	[muenter_bekehren_1772:213]	… Wohlgefallen siehet nun Gott auf Sie, Jhr	**Heil**	**ist theuer** vor ihm geachtet, mit aller …
12:	[buerger_gedichte_1778:231]	Das	**Brod**	**war theuer** in der Stadt; Doch theurer …
13:	[hippel_lebenslaeufe0302_1781:604]	… Nase in euren theuren Rath (denn guter	**Rath**	**ist theuer**!) steckt, schlagt dem …
14:	[sander_beschreibung01_1783:54]	Tobakrauchen ist nicht sehr üblich, der schlechte	**Tobak**	**ist theuer**, lange Pfeifen hat man nicht …
15:	[goethe_iphigenie_1787:89]	Jeder	**Augenblick**	**ist theuer**, Und unsre Rückkehr hängt an …

Deutsches Textarchiv (DTA)

Text-Bild-Ansicht öffnen …

Röntgen, Wilhelm Conrad: Ueber eine neue Art von Strahlen. Vorläufige Mittheilung. 2. Aufl. Würzburg, 1896.

Bibliographische Angaben

URN:	urn:nbn:de:kobv:b4-200905194973
Titel:	Ueber eine neue Art von Strahlen
Untertitel:	Vorläufige Mittheilung
Autor/in:	Wilhelm Conrad Röntgen (GND, Wikipedia, ADB/NDB)
Erscheinungsjahr:	1896
Verlag:	Stahel
Ort:	Würzburg
Auflage:	2. Auflage
Bibliothek:	Staatsbibliothek zu Berlin – Preußischer Kulturbesitz
Signatur:	SBB-PK, My 14999<2> R

Zugehörige Werke

- Röntgen, Wilhelm Conrad: Ueber eine neue Art von Strahlen. Fortsetzung. In: Sitzungsberichte der Würzburger Physik.-

Suche im Werk

suchen Hilfe

Ansichten für dieses Werk

- Text-Bild-Ansicht
- alle Faksimiles
- DTAQ (Qualitätssicherung)

Download

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

Metadaten

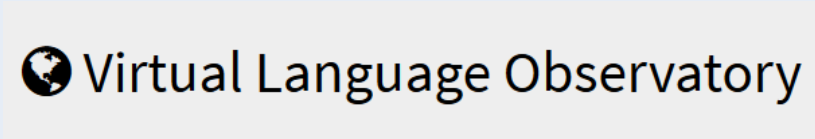
TEI-Header · CMDI · Dublin Core

Statistiken

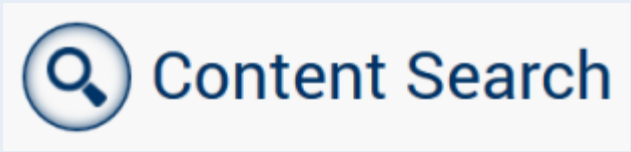
Scans:	22
Zeichen:	ca. 22 609
Tokens:	ca. 3 087
Oberflächentypes:	ca. 1 136

Anbindung CLARIN-D

Bündelung von Sprachressourcen verschiedener Anbieter
→ http://vlo.clarin.eu



Übergreifende Suche über CLARIN-D-Ressourcen
→ https://www.clarin.eu/contentsearch

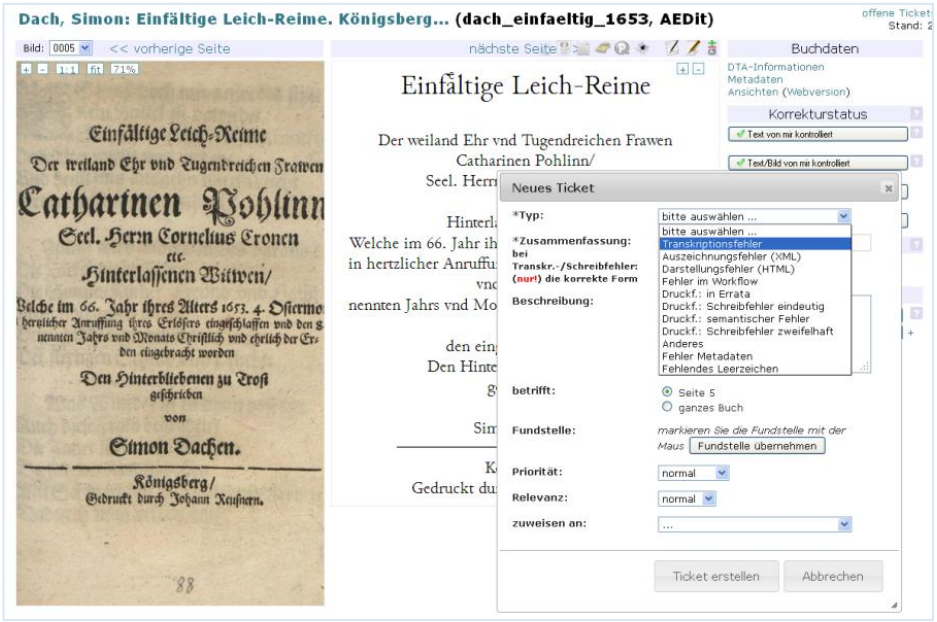


Best Practices
→ http://www.clarin-d.de/de/sprachressourcen/benutzerhandbuch.html

Langzeitverfügbarkeit von Daten in Repositories, PIDs
→ http://clarin.bbaw.de

DTA und Community

DTAQ



- Qualitätssicherung in DTAQ
- Kollaborative Plattform
- Textkorrektur/Texttranskription mit Versionierung
- Fehler melden im Ticketsystem
- Kommentarfunktion
- Zugang nach Anmeldung: 1125 angemeldete Nutzer (Stand 21.11.16)

www.deutschestextarchiv.de/dtaq

Quelle: http://www.deutschestextarchiv.de/dtaq/book/view/30951?p=16

DTAQ: Text-Editor

DTAQ | zuletzt gelesen · Hilfe · Zufallsseite | SusanneHaaf | Admin | Profil | ausloggen

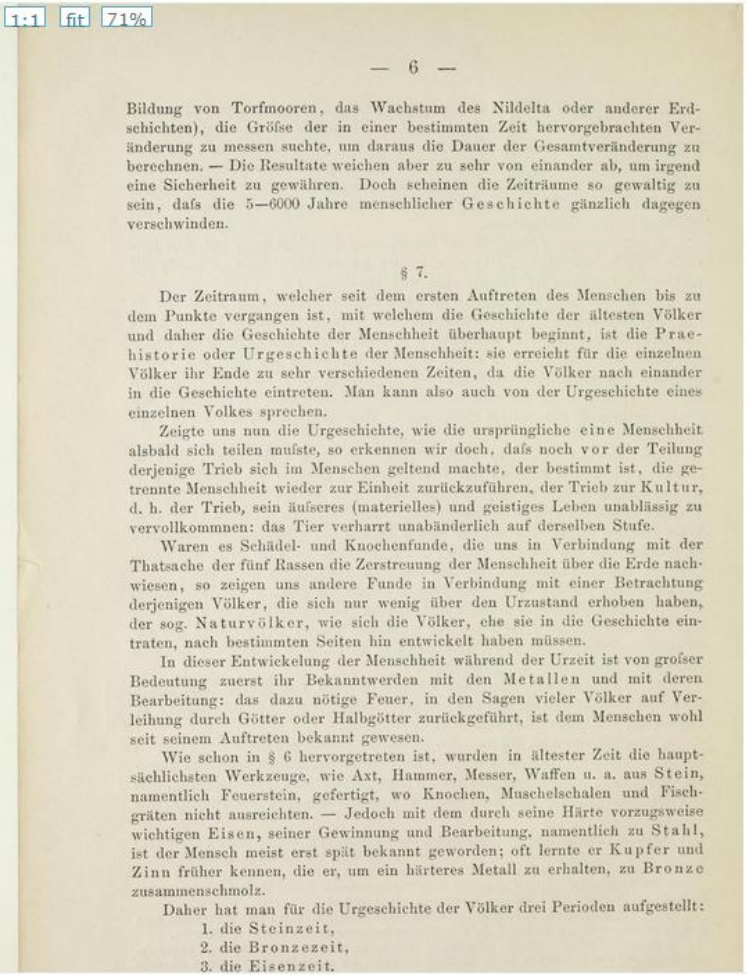
meyer_geschichte_1890 (GEI)

offene Tickets: 27 (0 *ganzes Buch*)
Stand: 2015-11-23 11:19:56

Text	Text/Bild
0 – 116 – 6	0 – 1 – 121

Bild: 0016 : — 6 — | << vorherige Seite | nächste Seite

1:1 | fit | 71%



— 6 —

Bildung von Torfmooren, das Wachstum des Nildelta oder anderer Erdschichten), die Gröfse der in einer bestimmten Zeit hervorgebrachten Veränderung zu messen suchte, um daraus die Dauer der Gesamtveränderung zu berechnen. — Die Resultate weichen aber zu sehr von einander ab, um irgend eine Sicherheit zu gewähren. Doch scheinen die Zeiträume so gewaltig zu sein, dafs die 5—6000 Jahre menschlicher Geschichte gänzlich dagegen verschwinden.

§ 7.

Der Zeitraum, welcher seit dem ersten Auftreten des Menschen bis zu dem Punkte vergangen ist, mit welchem die Geschichte der ältesten Völker und daher die Geschichte der Menschheit überhaupt beginnt, ist die Praehistorie oder Urgeschichte der Menschheit: sie erreicht für die einzelnen Völker ihr Ende zu sehr verschiedenen Zeiten, da die Völker nach einander in die Geschichte eintreten. Man kann also auch von der Urgeschichte eines einzelnen Volkes sprechen.

Zeigte uns nun die Urgeschichte, wie die ursprüngliche eine Menschheit alsbald sich teilen mufste, so erkennen wir doch, dafs noch vor der Teilung derjenige Trieb sich im Menschen geltend machte, der bestimmt ist, die getrennte Menschheit wieder zur Einheit zurückzuführen, der Trieb zur Kultur, d. h. der Trieb, sein äufseres (materielles) und geistiges Leben unablässig zu vervollkommnen: das Tier verharrt unabänderlich auf derselben Stufe.

Waren es Schädel- und Knochenfunde, die uns in Verbindung mit der Thatsache der fünf Rassen die Zerstreuung der Menschheit über die Erde nachwiesen, so zeigen uns andere Funde in Verbindung mit einer Betrachtung derjenigen Völker, die sich nur wenig über den Urzustand erhoben haben, der sog. Naturvölker, wie sich die Völker, ehe sie in die Geschichte eintraten, nach bestimmten Seiten hin entwickelt haben müssen.

In dieser Entwickelung der Menschheit während der Urzeit ist von grofser Bedeutung zuerst ihr Bekanntwerden mit den Metallen und mit deren Bearbeitung: das dazu nötige Feuer, in den Sagen vieler Völker auf Verleihung durch Götter oder Halbgötter zurückgeführt, ist dem Menschen wohl seit seinem Auftreten bekannt gewesen.

Wie schon in § 6 hervorgetreten ist, wurden in ältester Zeit die hauptsächlichsten Werkzeuge, wie Axt, Hammer, Messer, Waffen u. a. aus Stein, namentlich Feuerstein, gefertigt, wo Knochen, Muschelschalen und Fischgräten nicht ausreichten. — Jedoch mit dem durch seine Härte vorzugsweise wichtigen Eisen, seiner Gewinnung und Bearbeitung, namentlich zu Stahl, ist der Mensch meist erst spät bekannt geworden; oft lernte er Kupfer und Zinn früher kennen, die er, um ein härteres Metall zu erhalten, zu Bronze zusammenschmolz.

Daher hat man für die Urgeschichte der Völker drei Perioden aufgestellt:

1. die Steinzeit,
2. die Bronzezeit,
3. die Eisenzeit.

Bildung von Torfmooren, das Wachstum des Nildelta oder anderer Erdschichten), die Gröfse der in einer bestimmten Zeit hervorgebrachten Veränderung zu messen suchte, um daraus die Dauer der Gesamtveränderung zu berechnen. — Die Resultate weichen aber zu sehr von einander ab, um irgend eine Sicherheit zu gewähren. Doch scheinen die Zeiträume so gewaltig zu sein, dafs die 5—6000 Jahre menschlicher Geschichte gänzlich dagegen verschwinden.

§ 7.

Der Zeitraum, welcher seit dem ersten Auftreten des Menschen bis zu dem Punkte vergangen ist, mit welchem die Geschichte der ältesten Völker und daher die Geschichte der Menschheit überhaupt beginnt, ist die Praehistorie oder Urgeschichte der Menschheit: sie erreicht für die einzelnen Völker ihr Ende zu sehr verschiedenen Zeiten, da die Völker nach einander in die Geschichte eintreten. Man kann also auch von der Urgeschichte eines einzelnen Volkes sprechen.

Zeigte uns nun die Urgeschichte, wie die ursprüngliche eine Menschheit alsbald sich teilen mufste, so erkennen wir doch, dafs noch vor der Teilung derjenige Trieb sich im Menschen geltend machte, der bestimmt ist, die getrennte Menschheit wieder zur Einheit zurückzuführen, der Trieb zur Kultur, d. h. der Trieb, sein äufseres (materielles) und geistiges Leben unablässig zu vervollkommnen: das Tier verharrt unabänderlich auf derselben Stufe.

Waren es Schädel- und Knochenfunde, die uns in Verbindung mit der Thatsache der fünf Rassen die Zerstreuung der Menschheit über die Erde nachwiesen, so zeigen uns andere Funde in Verbindung mit einer Betrachtung derjenigen Völker, die sich nur wenig über den Urzustand erhoben haben, der sog. Naturvölker, wie sich die Völker, ehe sie in die Geschichte eintraten, nach bestimmten Seiten hin entwickelt haben müssen.

In dieser Entwickelung der Menschheit während der Urzeit ist von grofser Bedeutung zuerst ihr Bekanntwerden mit den Metallen und mit deren Bearbeitung: das dazu nötige Feuer, in den Sagen vieler Völker auf Verleihung durch Götter oder Halbgötter zurückgeführt, ist dem Menschen wohl seit seinem Auftreten bekannt gewesen.

Wie schon in § 6 hervorgetreten ist, wurden in ältester Zeit die haupt

Quelle: http://www.deutschestextarchiv.de/dtaq/book/view/30951?p=16

Kooperationen im Rahmen von DTAE

- Alexander-von-Humboldt Forschungsstelle (BBAW): Digitalisierung unselbstständig erschienener Schriften Alexander von Humboldts
- Forschungsstelle für Personalschriften (AdW Mainz): Digitalisierung von Leichenpredigten aus dem Bestand der ehemaligen Stadtbibliothek Breslau
- Herzog August Bibliothek Wolfenbüttel, Wolfenbütteler Digitale Bibliothek (WDB), Projekt AEDit Frühe Neuzeit: Integration von Volltexten aus verschiedenen Projekten der HAB/WDB
- Johann Friedrich Blumenbach – online (AdW zu Göttingen): Integration von deutschsprachigen Texten Blumenbachs in das DTA und Bereitstellung einer Rechercheschnittstelle
- Marx-Engels-Gesamtausgabe (BBAW): Aufbereitung und Publikation der digitalisierten „Neuen Rheinischen Zeitung"
- Prof. Dr. Thomas Gloning (Universität Gießen, Institut für Germanistik): Aufbereitung und Publikation digitalisierter, diskurslinguistisch relevanter Texte
- Projekt „Digitalisierung des Polytechnischen Journals" (Projektleitung: Prof. Dr. Christian Kassung, Humboldt-Universität Berlin: Institut für Kulturwissenschaft): Integration des Polytechnischen Journals als Spezialkorpus in das DTA und Bereitstellung einer Rechercheschnittstelle
- Max-Planck-Institut für Bildungsforschung: Integration christlicher Ratgeber in das DTA und Bereitstellung einer Rechercheschnittstelle
- Projekt „Sandrart.net": Integration von Sandrarts „Teutscher Akademie" in das DTA-Kernkorpus und Bereitstellung einer Rechercheschnittstelle
- Projekt „Von Leipzig in die Romania" (Projektleitung: Prof. Dr. Elisabeth Burr, Universität Leipzig: Institut für Romanistik): Korrektur und Annotation von OCR-Rohtexten des DTAs durch Studierende
- Staats- und Universitätsbibliothek (SuUB) Bremen, Digitale Sammlung "Die Grenzboten": DFG-Projekt zur Nachbearbeitung des OCR-Volltextes der Zeitschrift "Die Grenzboten" (1841–1922; 270 Bände, ca. 180000 Einzelseiten); TEI-XML-Strukturierung mit semiautomatischen Methoden und Werkzeugen des DTA, Bereinigung von des Volltextes von OCR-Fehlern auf der Grundlage des DTA als Referenzkorpus

http://www.deutschestextarchiv.de/dtae

DTAE & CLARIN

DTA: „Datenkultur der Offenheit und des Teilens“

- Community: Zusammenfügen verstreuter Ressourcen, Interoperabilität, Vergleich von Korpora und Subkorpora
- Projekte: Publikations- und Qualitätssicherungsplattform, Anbindung an andere Korpora, Längerfristige Verfügbarkeit

DTA/TEI und Interoperabilität

TEI und Interoperabilität

- Wenn jeder nach den TEI/P5-Richtlinien auszeichnet: Erhalten wir interoperable Texte?

  Oder:

- Sind TEI-Texte per se interoperabel?

  Cf. Unsworth: *Computational Work with Very Large Text Collections*, 2011.

- Interoperabilität:
  - Möglichkeit, Texte mit denselben Methoden und Tools auszuwerten mit verlässlichen Resultaten

TEI und Interoperabilität

- Empfehlungen des Rats für Infrastrukturen für eine NFDI
  - Spannungsfeld Vereinheitlichung – Vielfalt (S. 34)
  - Bedingungen für Interoperabilität schaffen (S. 35)
  - Übergreifende Standards notwendig (S. 46)
  - Qualitätssicherung durch „gute Dokumentation“ (S. 47)
  - Interoperabilität ist auch für Qualitätsstandards ein wichtiges Ziel (S. 53)

Quelle: http://www.rfii.de/ (Positionspapier von 05/2016)

Das DTA-Basisformat

DTA-Basisformat (DTABf)

```
<TEI>
  <text>
    <front>
      <pb facs="#f0007" n="[III]"/>
      <titlePage type="main">
        <docTitle>
          <titlePart type="main"> <hi
rendition="#b">Alexander von Humboldts<lb/>
Rei&#x017F;e in die Aequinoktial-Gegenden<lb/>
des neuen Kontinents.</hi> </titlePart>
        </docTitle><lb/>
        <byline><hi rendition="#g">In
deut&#x017F;cher Bearbeitung</hi><lb/>
von<lb/><docAuthor><hi rendition="#b">Hermann
Hauff.</hi></docAuthor><lb/>
Nach der Anordnung und unter Mitwirkung des
Verfa&#x017F;&#x017F;ers.<lb/>
Einzige von A. von Humboldt anerkannte Ausgabe
in deut&#x017F;cher Sprache.</byline><lb/>
        <milestone rendition="#hr"
unit="section"/>
        <titlePart type="volume"> <hi
rendition="#b">Er&#x017F;ter Band.</hi>
</titlePart><lb/>
        <figure/><lb/>
        <docImprint><pubPlace><hi
rendition="#b"><hi rendition="#g">Stuttgart</hi>.
```

DTABf

- TEI-Format für die Auszeichnung von historischen Texten in Korpora
- Schema & Dokumentation
- Ziel: Interoperabilität gewährleisten (Ambiguität vermeiden)

DTABf: Was wird ausgezeichnet

- Formale (Layout-) Textauszeichnung
- Inhaltliche (Logische, Konzeptuelle) Textauszeichnung
- Auszeichnung von Metadaten
- Richtlinien zur Transkription

DTABf: Tagset

Tagset:	Echtes Subset der TEI
TEI:	**552** Elemente + Attribute
DTABf:	**129** Elemente + Attribute und Werte **83** im Textbereich (<text>) (davon 7 nur in Manuskripten, 76 in Drucken) **46** im Headerbereich (<teiHeader>)

DTABf: Elemente

Ziel: Vermeidung von Ambiguitäten
→ Interoperabilität

TEI

```
<rs type="propNounPersName">Sokrates</rs>,
<name type="person">Platon</name> und
<persName>Aristoteles</persName> waren Philosophen.
```

DTABf

```
<persName>Sokrates</persName>,
<persName>Platon</persName> und
<persName>Aristoteles</persName> waren Philosophen.
```

DTABf: Attribute

- Festlegungen im DTABf nicht nur für Elemente, sondern auch für Attribute und deren Werte
- Beispiel *Anmerkungen*: `<note>`

1. „lokale“ Attribute (= TEI P5):

   `@place, @type, @n`

2. globale Attribute (= TEI P5):

   `@xml:id, @sameAs, @corresp, @prev, @next, @rendition, @rend`

3. nicht gültig (↔ TEI P5):

   `@cert, @resp, @subtype, @evaluate, @facs, @ana, @targetLang, @target, ...`

DTABf: Attribut-Werte

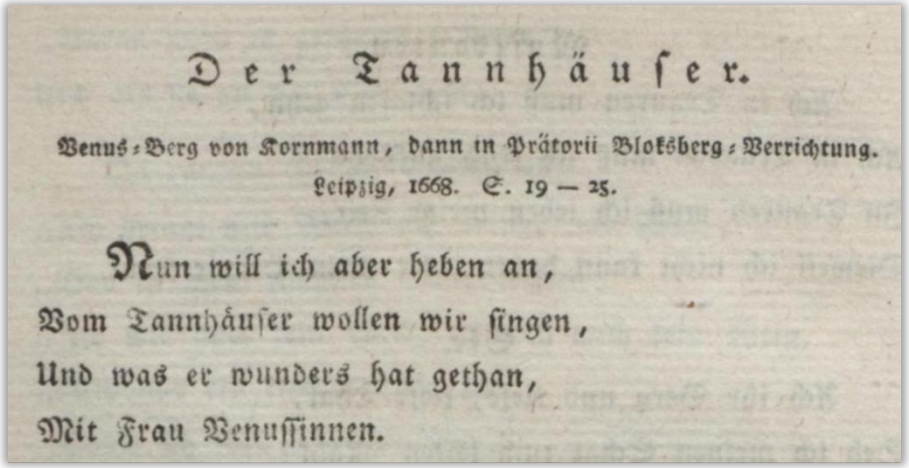
1. Wert-Auswahl für ein bestimmtes Attribut festgelegt (je Klasse oder Element)

```
<note place="left">
    zu München
</note>
```

Sandrart, Joachim von: L'Academia Todesca. Bd. 1,1. Nürnberg, 1675.
<http://www.deutschestextarchiv.de/sandrart_academie0101_1675>

DTABf: Attribut-Werte

2. Datenstruktur eines Wertes festgelegt (je Element oder Klasse)



Der Tannhäuser.

Venus-Berg von Kornmann, dann in Prätorii Bloksberg-Verrichtung. Leipzig, 1668. S. 19 — 25.

Nun will ich aber heben an,
Vom Tannhäuser wollen wir singen,
Und was er wunders hat gethan,
Mit Frau Venussinnen.

```
<lg type="poem">
  <lg n="1">
    <l>Nun will ich aber
       heben an,</l><lb/>
    <l>Vom Tannhäuser wollen wir singen,</l><lb/>
    <l>Und was er wunders hat gethan,</l><lb/>
    <l>Mit Frau Venussinnen.</l>
  </lg><lb/>
  <lg n="2">...</lg>
<lg><lb/>
```

(Arnim/Brentano: Des Knaben Wunderhorn. Bd. 1. Heidelberg, 1806, S. 86, <http://www.deutschestextarchiv.de/arnim_wunderhorn01_1806/95>)

DTABf: Werte

1. Wert-Auswahl je Attribut festgelegt (je Element oder Klasse)

```
<note>    @place="left|right|end|foot"
          @type="editorial"
<lg>      @type="poem"
```

2. Attribut festgelegt, Datenstruktur des Wertes festgelegt

```
<lg>      @n="[data.count]"
```

3. keine Festlegung beim Wert (wenn unabdingbar)

```
<note>    @n (für das Fußnotenzeichen)
```

DTA-Basisformat: Komponenten

Komponenten:

a. formale Spezifikation mittels TEI-ODDs & Schemas
b. zusätzliche Constraints mittels Schematron
c. ausführliche Dokumentation
d. Korpus

basisformat.sch

```
<pattern id="allElements">
    <rule context="tei:*">
        <report test="@rendition and @rend" role="WARNING">
            [W0001] The usage of @rend or @rendition should be exclusionary.
        </report>
    </rule>
</pattern>
<pattern id="teiHeaderElements">
    <rule context="tei:addName | tei:address | tei:addrLine | tei:email | tei:biblFull | tei:country |
        tei:forename | tei:genName | tei:measure | tei:msDesc | tei:nameLink | tei:publicationStmt |
        tei:resp | tei:respStmt | tei:roleName | tei:surname | tei:titleStmt | tei:title">
        <report test="ancestor::tei:text" role="ERROR">
            [E0001] Element "<name/>" not allowed anywhere within element "text".
        </report>
    </rule>
</pattern>
```

Verfügbarkeit: http://www.deutschestextarchiv.de/doku/basisformat

DTABf: Dokumentation

Das DTA-Basisformat DTA

Inhaltsverzeichnis dieses Dokumen

1 Einführung
2 DTABf Dokumentation
3 DTABf Schema
4 Hilfreiche Tools und Anwendungen
4.1 Webformular zur DTABf-konforr
4.2 oXygen-Framework zur Texterfa

1 Einführung

Auf den folgenden Seiten wird das TEI/X
Grundlage für die Annotation sämtlicher
soll im Rahmen der DTA-Richtlinien, die
sowie die Transkriptionsrichtlinien umfas
ermöglichen und dabei gleichzeitig Variat
dass die Kohärenz der DTA-Texte unterei
folgt dabei den P5-Richtlinien der Text Er
historischer gedruckter Werke in einem K

2 DTABf Dokumentation

- Einführung und erste Schritte
- Strukturierung der Metadaten
- Formale Erschließung der Texte
- Inhaltliche Erschließung der Texte
- Auszeichnung von Zeitungen
- Auszeichnung von Manuskripten

Inhaltliche Erschließung des Volltextes

Inhaltsverzeichnis dieses Dokuments

1 Allgemeines
2 Texteinteilung
3 Einleitende Informationen zum Buch
3.1 Allgemeines
3.2 Titelblätter
3.2.1 Haupttitelseite
3.2.2 Titelseite einer Reihe und zugehörige Haupttite
3.3 Widmungen
3.4 Epigraphe
3.5 Inhaltsverzeichnis
4 Inhaltliche Kodierung des Textkörpers
4.1 Allgemeines
4.2 Unterbrechungen zusammenhängender Textbestan
4.2.1 Eine Textpassage wird durch einen Einschub ur
4.2.2 Eigenständige Textpassagen korrespondieren ir
fremdsprachliche Texte und deren Übersetzung), steh
4.3 Fußnoten
4.3.1 Auf eine Seite begrenzte Fußnoten
4.3.2 Fortlaufende Fußnoten
4.4 Endnoten
4.4.1 Verweise auf Endnoten im Text
4.4.2 Wiedergabe der Endnoten
4.5 Randbemerkungen (Marginalien)
4.6 Fremdsprachliches Material
4.7 Inhaltszusammenfassung
4.8 Bibliographie
4.9 Nachsatz
4.10 Zitate und Epigraphe
4.10.1 Allgemeines
4.10.2 Auszeichnung von Zitaten
4.10.3 Auszeichnung von Epigraphen
4.10.4 Zitate/Epigraphe als Versgruppen
4.11 Gedichte und gebundene Sprache
4.12 Dramen
4.12.1 Die Figurenaufstellung
4.12.2 Die Strukturierung in Akte/Aufzüge und Szenen
4.12.3 Bühnenanweisungen

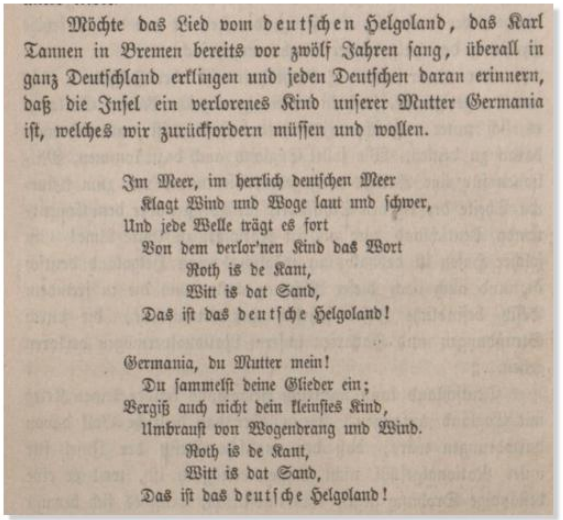
4.11 Gedichte und gebundene Sprache

Gedichte in Prosawerken und Gedichtbänden mit einfacher Struktur (d. h. nur Gedichttitel und Strophen, keine zusätzlichen Textbestandteile) werden mittels `<lg>`s strukturiert. Dabei umschließt das Element `<lg>` zum einen das gesamte Gedicht, wobei diese Verwendung durch das Attribut-Wert-Paar `@type="poem"` angezeigt wird. Das `<lg>`-Element umschließt weiterhin jede einzelne Strophe. In diesem Fall steht kein `@type`-Attribut. Bei mehrstrophigen Gedichten wird die jeweilige Strophennummer im `@n`-Attribut des `<lg>`-Elements angegeben. Der Gedichttitel steht im `<head>`-Element der äußeren `<lg>`. Verse werden innerhalb der `<lg>` mittels `<l>[...]</l>` ausgezeichnet. Ein Vers kann dabei über einen Zeilenumbruch hinausreichen.

Gedichte in Prosawerken:

```
<lg type="poem">
  <head>[Titel]</head> <!-- sofern vorhanden -->
  <lg n="[Strophennummer]"> <!-- sofern kein einstrophiges Gedicht -->
    <l>[Vers]</l>
    <l>[Vers]</l>
  </lg>
</lg>
```

Beispiel



Möchte das Lied vom deutſchen Helgoland, das Karl Tannen in Bremen bereits vor zwölf Jahren ſang, überall in ganz Deutſchland erklingen und jeden Deutſchen daran erinnern, daß die Inſel ein verlorenes Kind unſerer Mutter Germania iſt, welches wir zurückfordern müſſen und wollen.

Im Meer, im herrlich deutſchen Meer
Klagt Wind und Woge laut und ſchwer,
Und jede Welle trägt es fort
Von dem verlor'nen Kind das Wort
Roth is de Kant,
Witt is dat Sand,
Das iſt das deutſche Helgoland!

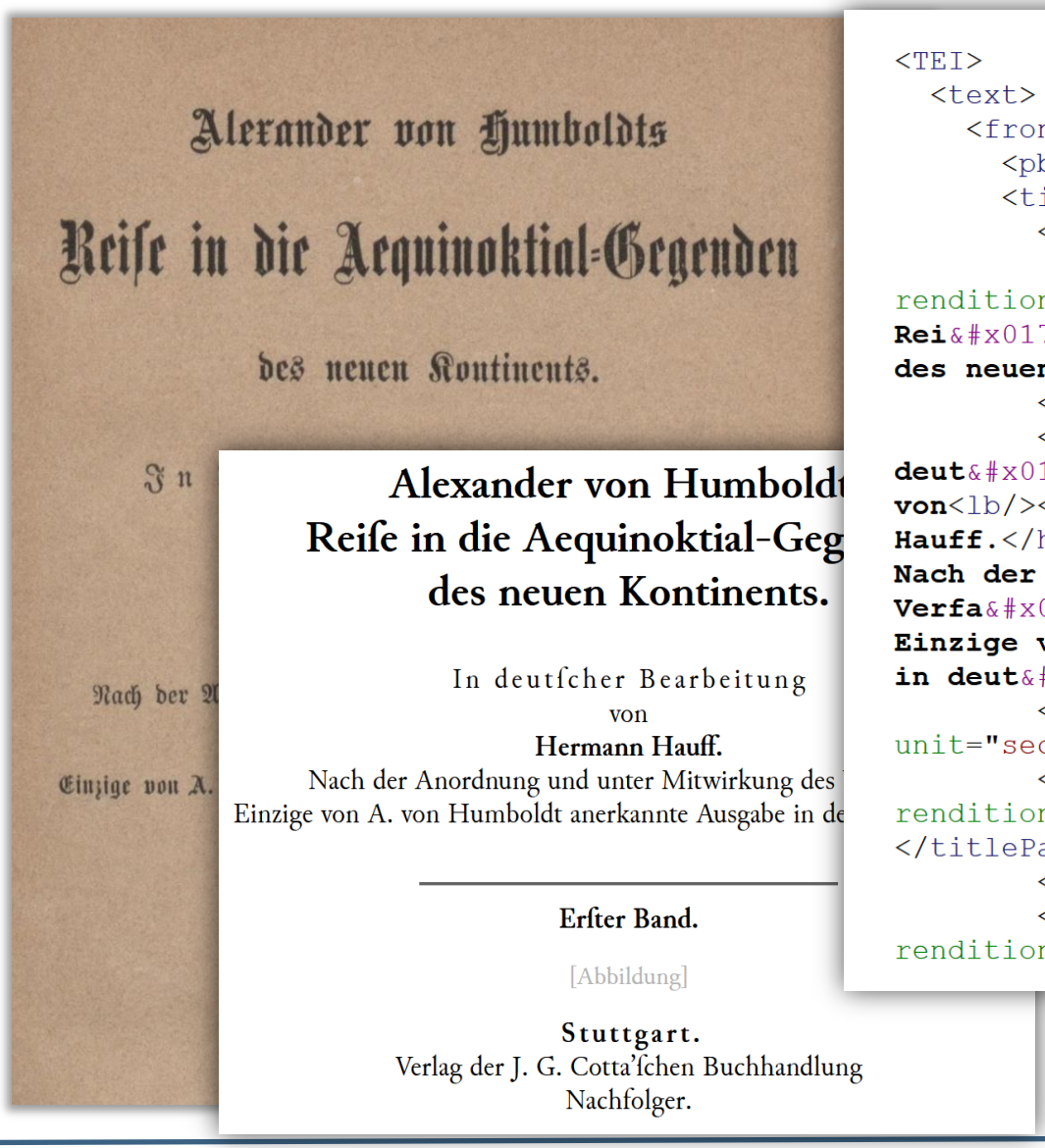
Germania, du Mutter mein!
Du ſammelſt deine Glieder ein;
Vergiß auch nicht dein kleinſtes Kind,
Umbrauſt von Wogendrang und Wind.
Roth is de Kant,
Witt is dat Sand,
Das iſt das deutſche Helgoland!

Werner, Reinhold: Erinnerungen und Bilder aus dem Seeleben. Berlin, 1880. [Faksimile 214]

```
<p>Möchte das Lied vom <hi rendition="#g">deutſchen</hi> Helgoland, das
Karl<lb/>Tannen in Bremen bereits vor zwölf Jahren ſang, überall in<lb/>ganz
Deutſchland erklingen und jeden Deutſchen daran erinnern,<lb/>daß die Inſel ein
verlorenes Kind unſerer Mutter Germania<lb/>iſt, welches wir zurückfordern müſſen
und wollen.</p>
<lb/>
```

http://www.deutschestextarchiv.de/doku/basisformat

DTABf: Textinstanzen



Alexander von Humboldts

Reiſe in die Aequinoktial-Gegenden

des neuen Kontinents.

Alexander von Humboldt
Reiſe in die Aequinoktial-Geg
des neuen Kontinents.

In deutſcher Bearbeitung
von
Hermann Hauff.
Nach der Anordnung und unter Mitwirkung des
Einzige von A. von Humboldt anerkannte Ausgabe in de

Erſter Band.

[Abbildung]

Stuttgart.
Verlag der J. G. Cotta'ſchen Buchhandlung
Nachfolger.

```
<TEI>
  <text>
    <front>
      <pb facs="#f0007" n="[III]"/>
      <titlePage type="main">
        <docTitle>
          <titlePart type="main"> <hi
rendition="#b">Alexander von Humboldts<lb/>
Rei&#x017F;e in die Aequinoktial-Gegenden<lb/>
des neuen Kontinents.</hi> </titlePart>
        </docTitle><lb/>
        <byline><hi rendition="#g">In
deut&#x017F;cher Bearbeitung</hi><lb/>
von<lb/><docAuthor><hi rendition="#b">Hermann
Hauff.</hi></docAuthor><lb/>
Nach der Anordnung und unter Mitwirkung des
Verfa&#x017F;&#x017F;ers.<lb/>
Einzige von A. von Humboldt anerkannte Ausgabe
in deut&#x017F;cher Sprache.</byline><lb/>
        <milestone rendition="#hr"
unit="section"/>
        <titlePart type="volume"> <hi
rendition="#b">Er&#x017F;ter Band.</hi>
</titlePart><lb/>
        <figure/><lb/>
        <docImprint><pubPlace><hi
rendition="#b"><hi rendition="#a">Stuttgart</hi>.
```

DTABf: Levels

Level 1: notwendig: `<cb/>, <list>, ≶, <note>`
Level 2: empfohlen: `<choice>, <fw>, <lb/>`
Level 3: fakultativ: `<foreign>, <persName>`
Level 4: unzulässig: `<ab>, <div1>, <g>`

- unterschiedliche editorische Erschließungstiefen
  - DTA-Kernkorpus: bis Level 2
- Alternativen für unzulässige Elemente

DTABf: Leitgedanken

- Identisches Tagging für semantisch ähnliche Phänomene

  → Homogene Textstrukturierung des gesamten Korpus
- Möglichst hohe Abdeckung von Phänomenen in geschriebenen historischen Texten
- Vermeidung von Ambiguitäten
- Dokumentation von Entscheidungen

Modularisierung des DTABf

DTABf: Modularisierung

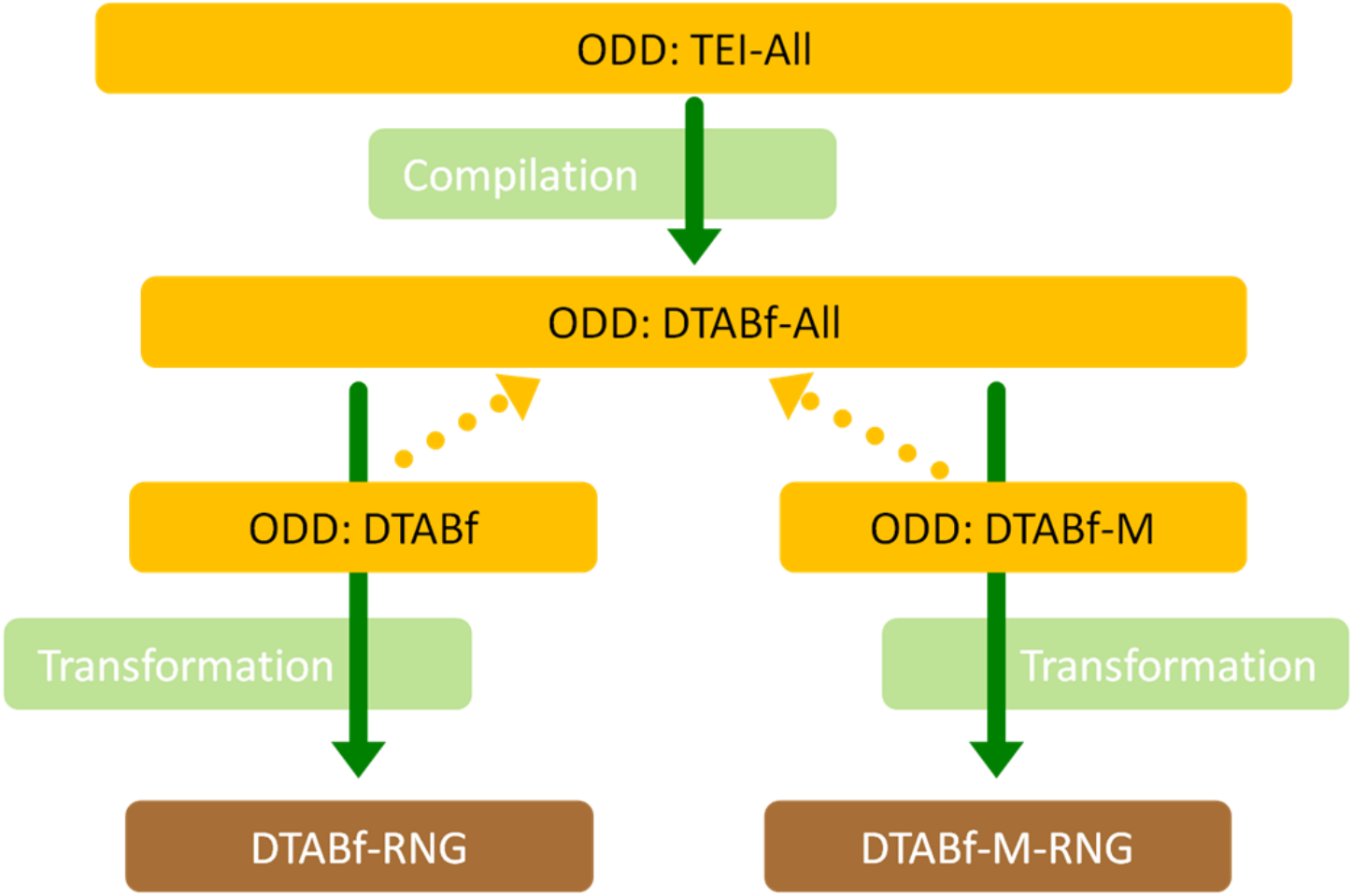
- DTA-Korpora wachsen stetig
  - ➢ speziell Textkuration (in CLARIN-D)
- Anpassungen für neue Texttypen
  - ➢ Manuskripte, Zeitungen, Funeralschriften
- Anpassungen an bislang unbekannte Phänomene
- Konsequenz für das DTABf: Modularisierung!

  a. des Schemas

  b. der Dokumentation

DTABf: Modularisierung des Schemas

- Modularisierung durch ODD-Chaining



DTABf: Chaining ODDs -- Beispiel

ODD für DTABf-All:

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

ODD für DTABf:

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

ODD für DTABf-M:

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

Cf. Haaf/Thomas, jTEI 10, im Erscheinen

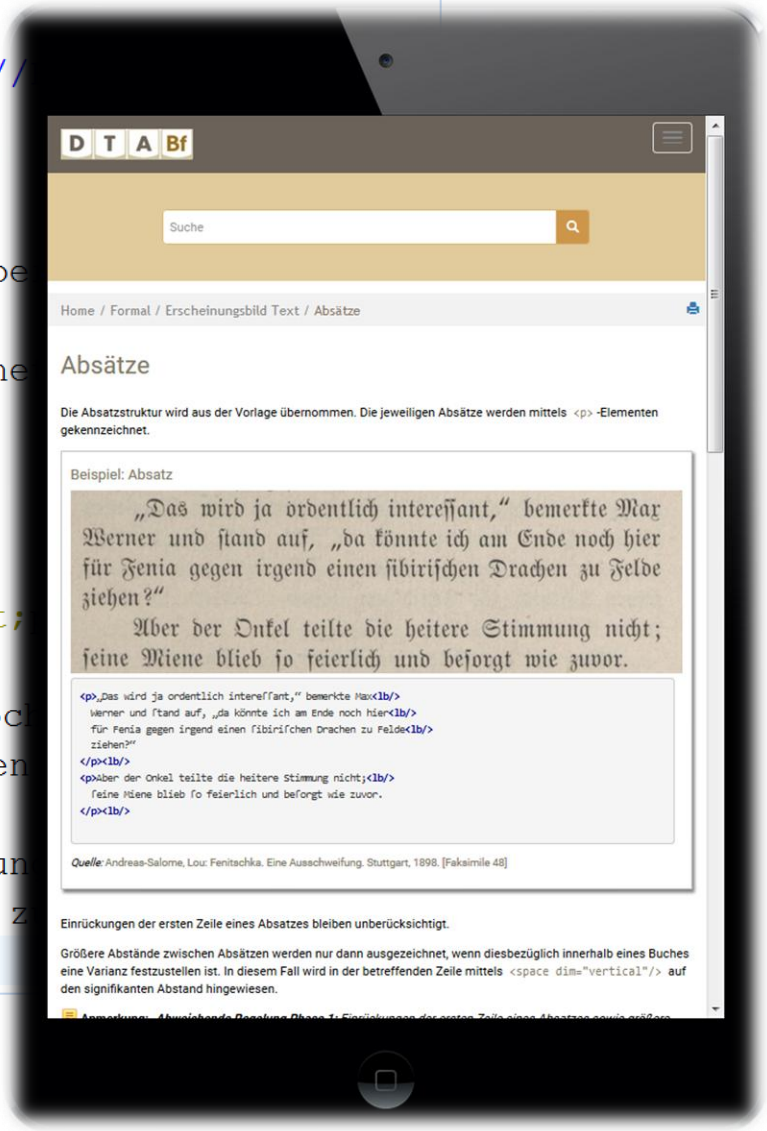
DTABf: Modularisierung der Dokumentation

- Modularisierung mithilfe von DITA
- XML-basiertes Dokumentieren

  → verschiedene Ausgabeformate (mobile Website, Webhelp, PDF, ...)
- individuelle Zusammenstellung der Dokumentationsthemen möglich

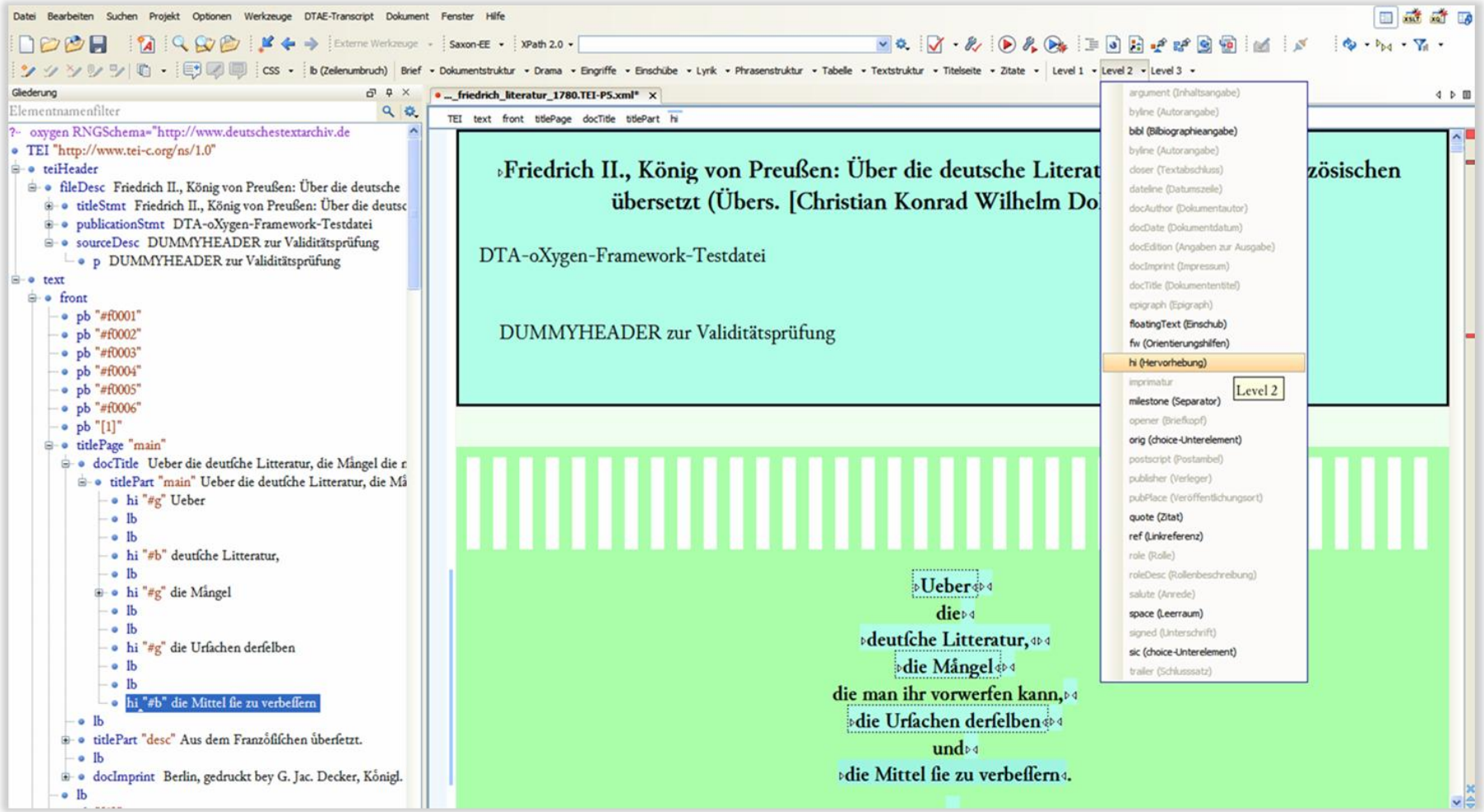
DTABf: Modularisierung der Dokumentation

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE topic PUBLIC "-//OASIS//DTD DITA Topic//
<topic id="topic_ry3_ptj_qs">
  <title>Absätze</title>
  <body>
    <p>Die Absatzstruktur wird aus der Vorlage übe
jeweiligen Absätze werden mittels
<codeph>&lt;p&gt;</codeph>-Elementen gekennzeichne
    <example>
      <title>Absatz</title>
      <fig>
        <image href="img/XHqxBBsaJ4.png"/>
        <codeblock outputclass="language-xml">&lt;
ordentlich intereſſant,“ bemerkte Max&lt;lb/&gt;
  Werner und ſtand auf, „da könnte ich am Ende noc
  für Fenia gegen irgend einen ſibiriſchen Drachen
Felde&lt;lb/&gt; ziehen?“&lt;/p&gt;&lt;lb/&gt;
&lt;p&gt;Aber der Onkel teilte die heitere Stimmun
  ſeine Miene blieb ſo feierlich und beſorgt wie z
&lt;/p&gt;&lt;lb/&gt;
```

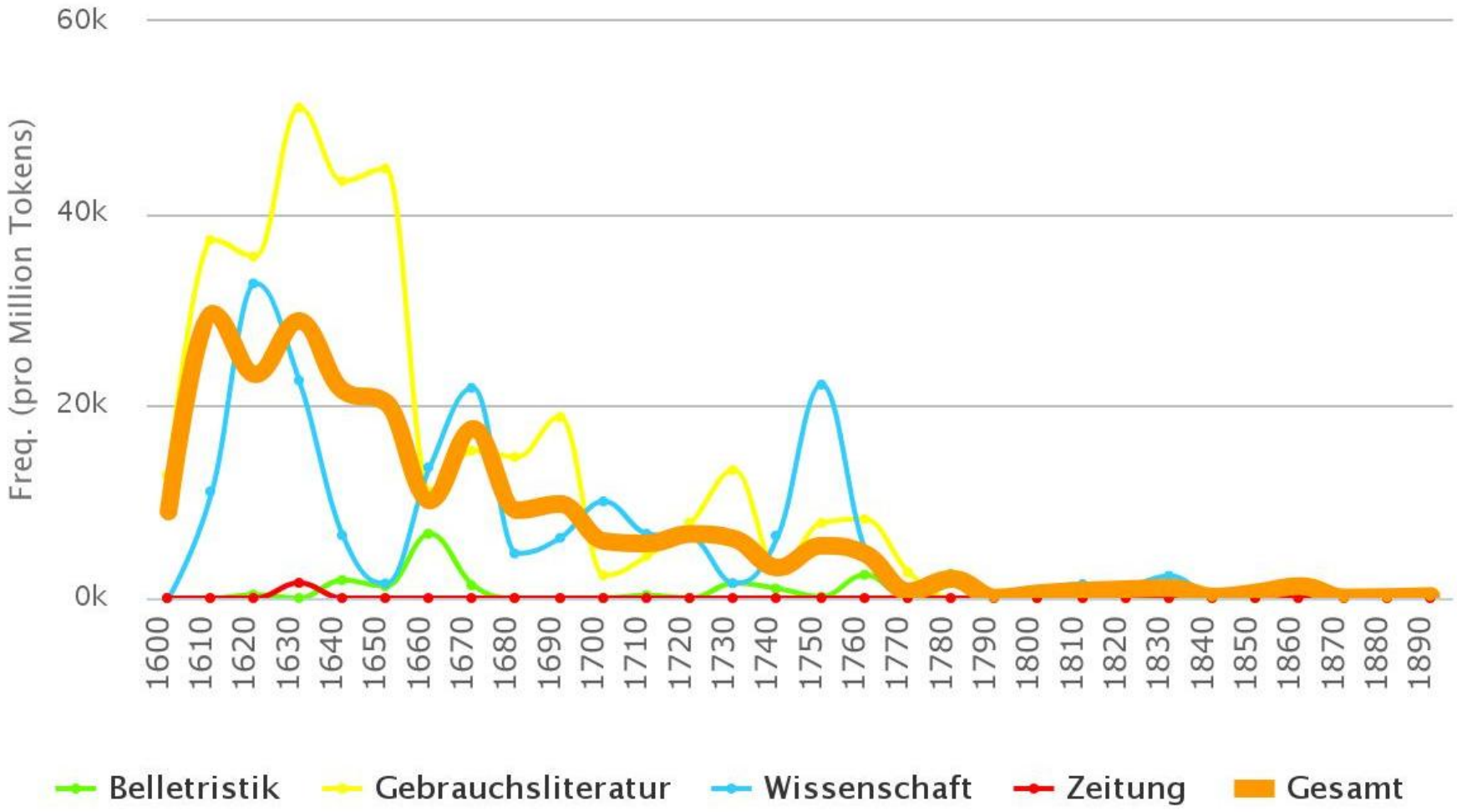


DTABf: Nutzung und Nutzen

DTABf: Hilfsmittel (Beispiel: DTAoX)



DTABf: Nutzen (Beispiel Marginalien)



TEI: <note place="left | right">

DDC: $con=/left/ || $con=/right/

DTABf: Nutzen (Beispiel: „frohlocken”)

- Duden (1999): 3 Bedeutungen
  1. *„Schadenfreude empfinden”*
  2. *„vor Freude jubeln; jauchzen“*
  3. *„lobsingen (veraltet)”*
- Korpusanalyse zur Verteilung von "frohlocken"
  - DDC: frohlocken with $xpath=/\b**p**\b/ (in Prosa-Paragraphen)
  - DDC: frohlocken with $xpath=/\b**l**\b/ (in Lyrik-Versen)
- Resultat:
  - übliche Verwendung: frohlocken (+ PP, z.B. "über")
  - *nur in der Lyrik*: frohlocken + Dativobj. ("jem.m frohlocken")
  - entspricht Bedeutung 3

DTABf: Nutzen (Beispiel: „frohlocken”)

„frohlocken” mit Dativ in der Lyrik:

Frolocket eurem Held/ nach Streit hat er jetzt Ruh.

(Mühlpfort 1686: 48)

Mir hüpfen die Berg’, und die Hügel! Mir frohlocket der Wald!

(Klopstock 1773: 147)

Frohlockt dir eine halbe Welt.

(Zachariae 1764: 254)

Wiehert ihm Rosse! frohlockt ihm, Adler!

(Kosegarten 1798: 41)

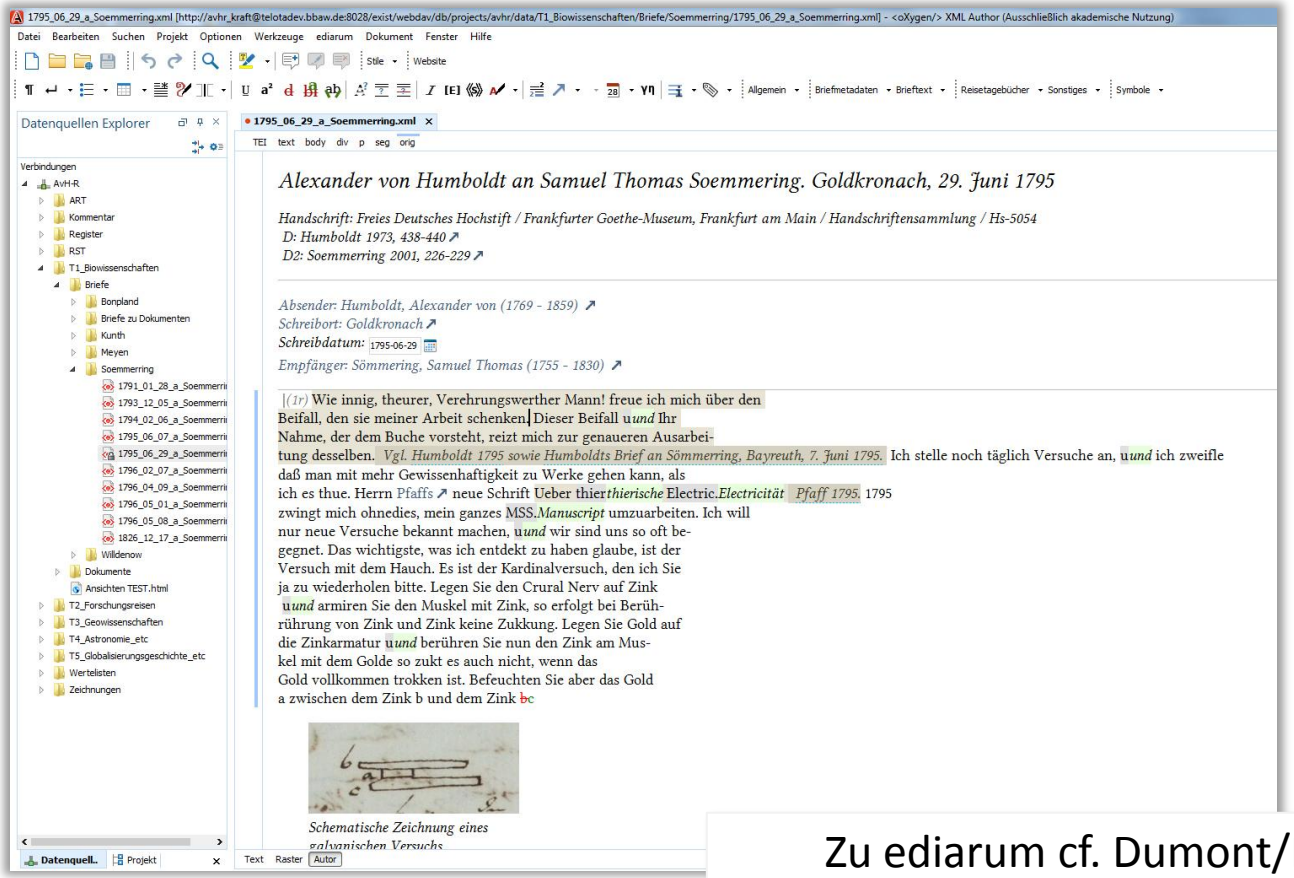
Frohlockt ihm, alle Frommen!

(Candidus 1854: 21)

Für weitere Beispiele cf. Haaf @LREC 2016

DTABf in der Community

- DTABf-Nutzung:
  - DTA-Korpora + DTA-Erweiterungen
  - Externe Projekte unabhängig vom DTA
  - beginnend: Projekte der BBAW → DTABf-Anpassung für *ediarum*



Zu ediarum cf. Dumont/Fechner, jTEI 8

DTABf in der Community

- DTABf ist eines der Formate, die *TEI Simple* zugrunde gelegt wurden
- Empfohlenes Format (unter anderen):
  - für die Auszeichnung historischer Korpora (CLARIN 2012)
  - als Archivformat für literaturwissenschaftliche Editionen (DFG 2016)
  - als Annotationsformat für linguistische Korpora (DFG 2014)

DTABf: Ausblick

kurzfristig (bis Anfang 2017):

- Publikation der neuen Fassung ("Chained" Schema und DITA-Dokumentation)
- "Umzug" aller Quelldateien auf github (Kommentierung durch die Community, Versionierung)

längerfristig (Beginn der Arbeiten 2017):

- Überarbeitung des DTA-oXygen-Frameworks
- DTABfversusTEI-Schema
- Dokumentation in Englisch

PARTHENOS

Pooling Activities, Resources and Tools for Heritage E-research Networking, Optimization and Synergies

Literatur

Zum DTA-Basisformat

- Haaf, Susanne, Alexander Geyken und Frank Wiegand: *The DTA "Base Format": A TEI Subset for the Compilation of a Large Reference Corpus of Printed Text from Multiple Sources.* In: jTEI 8 (2014/15). http://jtei.revues.org/1114

Zitierte Literatur

- CLARIN-D-Benutzerhandbuch. 2012. http://de.clarin.eu/de/hilfe/benutzerhandbuch
- DFG: Förderkriterien für wissenschaftliche Editionen in der Literaturwissenschaft. 2016. http://www.dfg.de/download/pdf/foerderung/antragstellung/forschungsdaten/foerderkriterien_editionen_literaturwissenschaft.pdf
- DFG: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. 2014. http://www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf
- Dumont, Stefan und Martin Fechner: *Bridging the Gap: Greater Usability for TEI encoding*. In: jTEI 8 (2015). https://jtei.revues.org/1242.
- Haaf, Susanne: *Corpus Analysis based on Structural Phenomena in Texts: Exploiting TEI Encoding for Linguistic Research.* In: Proceedings of the LREC 2016. http://www.lrec-conf.org/proceedings/lrec2016/pdf/1154_Paper.pdf.
- Haaf, Susanne und Christian Thomas: *Introducing the DTABf-M: A Manuscript-specific Extension to the DTA ›Base Format‹ (DTABf).* In: jTEI 10 (accepted; in publication)