

A living archive of 15th to 19th c. German: corpus strategies, technology, organization

Alexander Geyken (BBAW) Thomas Gloning (JLU Gießen)

Historical Corpora 2012 – Frankfurt, 8-12-2012









Overview

Part I:

- corpus situation 15th to 19th c. German
- idea of a living archive
- research questions and use cases

Part II:

- technical aspects
- organizational aspects









CLA

1. Corpus situation 15th-19th c. German

- CLARIN-D
- no balanced corpus of sufficient size for Early New High German and older NHG
- no integrated platform (data format, metadata, tools)
- no established culture of sharing and making textual data publicly available



wealth of diverse textual data (collections, research projects, individuals ...)











- a collaborative, sustainable and interoperable platform
- contains: textual data, metadata including aspects of linguistic variation, expert technology
- promote a culture of sharing
- develop integration scenarios for textual data coming from research projects, individuals
- the living archive as an opportunistic repository
 - producing specific textual subsets using metadata
 - acquisition strategies: producing systematic subcorpora











- design of subcorpora for research questions requires
 - metadata (date, location, text type, subject field, author)
 - knowledge about the history of the German Language, its text types, changing parameters of linguistic evolution
- three examples
 - word usage and semantic evolution across text types
 - linguistic profiles of historical text types
 - development of constructions/grammatical patterns











- a culinary corpus from the 15th-19th century
 - linguistic: evolution of thematic and functional vocabulary
 - corpus technology: How will corpus tools work with older material?
- putting the Dingler Corpus (1820-1931) to use
 - linguistic: evolution of technical terminology and textual organization in the seminal phase of technical science
 - corpus technology: handling large data sets; sharing/integration technology (Berlin, Leipzig)









- 5. technical aspects
- a) an agreed standard of corpus encoding (TEI-P5)
- b) infrastructure for the lifecycle of digital texts
- c) flexible subcorpus selection
- 6. organizational aspects

promote culture of sharing with appropriate metadata











- TEI-P5 as a starting point
- DTA-base format: a suitable subset of TEI-P5 for structuring of historical printed text
 - basic idea: meet criteria of interoperability (Unsworth 2011)
 - focus on non-controversial aspects of the text, thus providing unambiguous solutions for text annotation
 - establish high quality transcription of the text
 - provide high quality metadata
- DTABf adopted by CLARIN-D as best practice for historical corpora (cf. CLARIN-D user guide)











- DTABf (<u>www.deutschestextarchiv.de/doku/basisformat</u>)
 - provide as much expressiveness as possible by being as precise as possible
 - E.g.: provide closed set of values for (most) attributes, be restrictive wrt. the selection of elements
 - be flexible wrt. the integration of other formats
 - annotation levels
 - level 1: required (<pb/>, <list>, <lg>, <note>, ...)
 - level 2: recommended (<choice>, <fw>, <lb/>, ...)
 - level 3: optional (<foreign>, <persName>, ...)
 - level 4: proscribed (<ab>, <div1>, <g>, ...)









DTA-base format: Documentation

www.deutschestextarchiv.de/doku/basisformat

DTA-Basisformat – Einführung

Die folgende Darstellung dokumentiert das XML-Basisformat des DTA, welch für die Annotation der DTA-Volltexte bildet. Das Basisformat folgt den P5-Rid Encoding Initiative (TEI). Da diese Richtlinien iedoch Lösungen für sämtliche der Textaufbereitung anbieten sollen und daher entsprechend vielfältig und bedürfen sie im konkreten Einzelfall einer näheren Spezifikation. Daher wur P5-Richtlinien für die Textstrukturierung im DTA-Korpus eine Tag-Auswahl ge die das DTA-Basisformat bildet. Dieses Tagset ist mit den P5-Richtlinien der konform: auf Erweiterungen (tei.extensions) durch davon abweichende Eler verzichtet.

Das DTA-Basisformat soll im Rahmen der DTA-Richtlinien, die daneben auch (Leitlinien des DTA sowie die Transkriptionsrichtlinien umfassen, eine uneinge Textaufbereitung ermöglichen und dabei gleichzeitig Variationsspielräume b so einschränken, dass die Kohärenz der DTA-Texte untereinander gewährle dieses Ziel stellt die Ausrichtung des DTA-Korpus in der Diachronie einerseit Textsortenvielfalt andererseits eine große Herausforderung dar, resultiert si einer strukturellen Variabilität der Vorlagen, der mit dem zur Verfügung steh Genüge getan werden muss.

Mit der Ausarbeitung des DTA-Basisformats wollen wir einen Vorschlag für ei Volltext-Aufbereitung historischer Texte unterbreiten. Damit sollen zum eine dem Basisformat kompatibel sind, in das DTA einfließen können, zum andere Verwendung von DTA-Texten in anderen Volltextarchiven erleichtert werden

Zu den Dateien der Basisformat-Dokumentation

- Erläuterungen zur Basisformat-Dokumentation
- Grundstruktur jedes TEI-Dokuments
- Strukturierung der Metadaten
- male Erschließung der Text
- Bersicht über alle Basisformat-F te im <text>-Bereich

Das DTA-Basisformat liegt im ODD- und RNG-Format vor

berlin-brandenburgische

AKADEMIE DER WISSENSCHAFTEN

- DTA-Basisformat als ODD-File
- DTA-Basisformat als RNG-File

Stand dieser Seite: We

Dokumente zum DTA-Basisformat:

- DTA-Basisformat Einführung
- Erläuterungen zur Basisformat-Dokumentation

- Inhaltliche Erschließung der Texte
- Übersicht über alle Basisformat-Elemente im <text>-Bereich

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 Haupttitels
- 3.3 Widmungen
- 3.4 Epigraphe
- 3.5 Inhaltsverzeichnis
- 4 Inhaltliche Kodierung des Textkörpers
- 4.1 Allgemeines
- 4.2 Unterbrechungen zusammenhängender Tex thestandteile
- 4.2.1 Eine Textpassage wird durch einen Einschub unterbrochen
- 4.2.2 Eigenständige Textpassagen korrespondieren inhaltlich miteinander (z.B. fremdsprachliche Texte und deren Übersetzung), stehen jedoch nicht in linearer Abfolae
- 4.3 Fußnoten
- 4.3.1 Auf eine Seite begrenzte Fußnaten
- 4.3.2 Fortlaufende Fußnoten
- 4.4 Endnoten 4.4.1 Verweise auf Endnoten im Text
- 4.4.2 Wiedergabe der Endnot
- 4.8 Bibliographie
- 4.9 Nachsatz
- 4.10.2 Auszeichny
- 4 10 3 n Epigraphen
- 4.10.4 Zitate/Epigraphe als Versg
- 12 Dramen 4.12.1 Die Figurenaufstellung

4.11 Gedichte und gebundene Sprache

e in Prosawerken und Gedichtbänden mit einfacher Struktur (d. h. nur Gedichttitel und Geochte in Prosawerken und Gedichtbänden mit einfacher Struktur (d. h. nur Gedichttite Strophen, keine zusätzlichen Textbestandteile) werden mittels <1g>s strukturiert. Dabei nschließt das Element <1g> zum einen das gesamte Gedicht, wobei diese Verwendung durch das Attribut-Wert-Paar @type="poem" angezeigt wird. Das <1g>-Element umschließt weiterhin jede einzelne Strophe. In diesem Fall steht kein @type-Attribut. Bei mehrstrophigen Gedichten wird die jeweilige Strophennummer im @n-Attribut des <1g>-Elements angegeben. Der Gedichttitel steht im <head>-Element der äußeren <1g>. Verse werden innerhalb der <1g> mittels <1>[...]</1> ausgezeichnet. Ein Vers kann dabei über einen Zeilenumbruch hinausreichen.

CLARIN

Gedichte in Prosawerken:

<lg type="poem">

<head>[Titel]</head> </-- sofern vorhanden --> <lp><lg n="[Strophennummer]"> </-- sofern kein einstrophiges Gedicht --> <1>[Vers]</1> <1>[Vers]</1>

</lg>

</1g>



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

Shochte das Lied vom <hi rendition="#g">deut[chen</hi> Helgoland, das Karl<lb/>Tannen in Bremen bereits vor zwölf Jahren fang, überall in<lb/>b/>panz Deutschland erklingen und jeden Deutschen daran erinnern, <lb/>daß die Infel ein verlorenes Kind unferer Mutter Germania<lb/>ift, welches wir zurückfordern müffen und wollen.

```
<1b/>>
<lg type="poem">
```

<ld n="1"> <l>Im Meer, im herrlich deutschen Meer</l> <1b/>> <1>Klagt Wind und Woge laut und fchwer,</1> <1b/>> <l>Und jede Welle trägt es fort</l> <1b/>> <1>Von dem verlor'nen Kind das Wort</1> <1b/>> <1>Roth is de Kant.</1> <1h/>>









4.5 Randbemerkungen (Marginalien) 4.6 Fremdsprachliches Mater 4.7 Inhaltszusammenfassu

- 4.10 Zitate und Epigra
- 4.10.1 Allgemeines

ng von Zitaten

4.11 Gedichte und gebundene Sprach

4.12.2 Die Strukturierung des Dramas

Grundstruktur iedes TEI-Dokuments

Strukturierung der Metadaten

• Formale Erschließung der Texte

b) Infrastructure and lifecycle of digital texts



- software to provide easy integration of legacy data into CLARIN-D infrastructure:
 - generic web-based framework (TEI-Integrator, Uni Leipzig), or
 - DTA workflow: extract metadata via DTA Web-form; convert text via OxGarage into TEI-P5 first, and then to DTA-BF via DTA-oXygen-Framework; integrate into CLARIN-D infrastructure
 - transcription and TEI-encoding control via DTAQ-web platform (www.deutschestextarchiv.de/dtaq)
 - linguistic annotation platform provided by CLARIN-D F-AG7 (working group computational linguistics)









b) Integration of texts (step 1 metadata)

DTAE – Import aus Wikisource

Schritt 1: Datenimport vorbereiten

Bitte geben Sie den Seitentitel des Textes bei Wikisource ein. Bitte beachten Sie, dass das ausgewählte Werk keine Teilseite eines übergeordneten Werkes ist.

Informationen h	olen
Linear Control of Cont	<text><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></text>

Metadaten

Vorgesehener DTA-Verzeichnisname: kraft_seegespenst_1910						
					Textvorlage	
					Autor:	Robert Kraft
Titel:	Das Seegespenst					
Untertitel:						
aus:	Das Buch für Alle, Illustrierte Familienzeitung, Jahrgang 1910, Siebtze					
Herausgeber:						
Ausgabe:						
Verlag:	Union Deutsche Verlagsgesellschaft					
Ort:	Stuttgart					
Jahr:	1910					
vorw. Schrifttype:	© Fraktur © Antigua ◉ keine Angabe					

Überprüfen, korrigieren und ergänzen Sie ggf. die Metadaten.









CLAR

b) Integration of texts (step 2: TEI-P5 DTABf)



tei Beacheiten Suchen Projekt Ontionen Warke	une DTAF-Transmit Delement Fenster Hilfe	
		XSLT XQ
	■ → Saxon-EE • 1998b 2.0 • • · · · · · · · · · · · · · · · · ·	
· B· F. 2 / 2 / 9 / 0	• 📑 🖉 💷 CSS • b (Zelenumbruch) Brief • Dokumentstruktur • Drama • Eingriffe • Einschübe • Lyrk • Phrasenstruktur • Tabele • Textstruktur • Titelseite • Zitate • Level 1 • Level 2 • Level 3 •	
erung di 4 ×	e A Das Seegespenst.teixml x	4
namen-Filter Q		
voen RNGSchema="http://media.dwds.de/dta/mec	ici text booy p	
i "http://www.tei-c.org/ns/1.0"		
telHeader	sourceDesc n aus Willicourse der freier Quellensammlung n sourceDesc fileDesc failleader	
text	sources (p) aus wikisource, det neten Quenchisammung (g) (sources est) (neteres) (neteres)	
body		
pb *#f0001*		
🖻 🍗 center Das Seegespenst. Erzählung von I		
b Erzählung von Robert Kraft.		
p *#right* (Nachdruck Verboten.)	text body ph-facs="#f0001" ob	
🍗 hi "#small" (Nachdruck Verboten.)		
p Ich kam von langer Reise, die ich als zw	center big b Das Seegespenst. b big b Erzahlung von Kobert Kraft. b center	
p Schon nach wenigen Tagen aber erhielt	p rendition="#right" hi rendition="#small" (Nachdruck Verboten.)	hi 🗸
p Ich war sprachlos. Mein Freund Paul Ka		
p Wir waren zusammen als Schiffsjungen	[p] Ich kam von langer Reise, die ich als zweiter Steuermann gemacht, nach Hamburg zurück. Das ganze Schiff wurde abgemustert. Ich liels mich auf dem	
p Fünf Jahre waren seit unserer Trennun	Seemannsamt als heuersuchender Offizier einschreiben und fuhr inzwischen nach der fernen Heimat. Freilich hatte ich wenig Aussicht, auf diese Weise eine neue	e
p Ich fuhr sofort nach Hamburg. Es war r	Heyer zu bekommen	
p Dieses Madchen verliebte sich in den de		
p Aber inr Gatte solte nicht unter andere	P>Schon nach wenigen Tagen aber erhielt ich ein Telegramm, das mich nicht wenig in Erstaunen versetzte: "Willst du als Erster bei mir fahren? Sofort her! Par	ul
 p So war es gekommen, dau Paul im Hank a Diere ideale Ebe, de im hudstäblichen 	Müller, Kapitän der Portland I iverpool Zurzeit Hamburg "n	
p Diese ideale crie, de in buchstabilitien	Multi, Kapitan dei Fortiand, Elverpool. Zaizer Hamourg. (p)	
 p Jvur ein Jahr Gauerte sie, sagte Paul r p Ibran Tod – oef inden 2ª wiederholte id 	Delth war sprachlos. Mein Freund Paul Kapitän eines englischen Dampfers? Und zwar, wie mit das mitgenommene Schiffsregister sagte, eines von sechstausene	d-
a Fine Shirzee witch de liber Bord - a		
 p Achte Statistee Hostif ale doel bord – di n Mehr erfuhr ich nicht. Wir hatten es au 	Tonnen: \2	
p Wir fuhren nach Lissabon, Während de	D Wir waren zusammen als Schiffsjungen und als Matrosen gefahren sechs Jahre lang, hatten zusammen die Steuermannsschule besucht, dann erst waren wir	
p Nun allerdings kann es ta, wie überall, a	(p) with a dealined as of minipargent and as a matterial general part angle nation and the second management between the second management of the second sec	
p Glücklich erreichten wir Lissabon, Noch	getrennt worden. Mir gelang es, auf einem Segler als zweiter Omzier anzukommen, und als ich zwei Jahre spater wieder etwas von Paul norte, war er zuletzt noch	N
P Wir dampften in wenigen Stunden hin,	immer als Matrose gefahren. Wieviele haben das Steuermannspatent in der Tasche und müssen noch als Matrose vor dem Mast fahren! Ich hatte unterdessen ja	auc
p So gingen wir auf der Reede vor Anker	schon mein Kapitänsexamen gemacht und war immer wieder froh, wenn ich als zweiter Steuermann ankam.	
p Mit einem Male fuhr er zurück, und ich s	Depressing in the American and Physical Provide State and the State in the State in the State Stat State State Stat State State S	
p "Mein Trauring! Und heute ist Evelinens	[p] Fünf Jahre waren seit unserer Trennung verflossen. Vor zwei Jahren also war Paul noch Matrose gewesen. Seitdem hatte ich nichts wieder von ihm gehört. U	Ind
p Der Goldreif war ihm, als er die Hand üt	jetzt war er Kapitän eines großen englischen Dampfers? Der hatte ja höllisch fix vorwärts gemacht!	
p Also heute vor einem Jahre war seine u		
p "Klar den Taucherapparat!" sagte Paul	p/Ich fuhr sofort nach Hamburg. Es war richtig mein alter Freund Paul. Erst siebenundzwanzig Jahre alt, aber ein ganzer Mann. Vor drei Jahren war es ihm	
p Unten mußte der Ring leicht zu finden s	gelungen, zum ersten Male eine Stelle als zweiter Steuermann zu bekommen. Auf einem englischen Dampfer war es. Der Kapitän war Mithesitzer des Schiffes, ia	der
p Die Taucherpumpe und alles, was dazu	any and a second s	-
p Der Kapitän kam, brachte alles in Ordni.	ganzan recedere, die ganziele Seinne rainen nets. Ond er wurde auf seinen Painten bestandig von seiner Fochter begienet, die sogar an Bord geboren und erzogen	
p Erst wurde eine Prüfung außerhalb des	worden war. (p)	
 p Zehn, elf, zwölf Meter wurden von Schl 		
p "Noch etwas nachgeben!" meldete das ⊻	w we tang vaue or aurouve renoison is invalid; token #smail invalid; must be equal to "#above-cap", #arow, "#b", "#blue", "#bottomBraced", "#rc", "#t", "#t", "#t", "#in", "#n", "#nin", "#k", "#leftBraced", "#red", "#right	, #righ
2	IEXT RASEE AUTO	











- convert TEI-Header into CMDI (harvest via OAI-PMH)
- upload TEI-P5/DTABf into CLARIN-D service center repository
- provide different "incarnations" of the file (xml, tei, text, tcf, image) via CMDI resource proxy description
- build full text index and provide access via CLARIN federated content search
- enable lemma based and spelling tolerant search (cf. talk by Jurish)









depends on metadata annotation and flexible filters of search engine :

- example 1: display only "first editions"/ word #has[edition,/first/]
- example 2: exclude non primary works (such as dictionaries or encyclopedias) word #has[textclass,/^(?!dict[encyclop).*/]
- example 3: flexible filter expansions in order to select text classes, e.g. culinary texts Word #has[textclass,/\bkochliteratur\b/]









6. Organizational aspects



system of reputation

- system of reputation for making resources available to the community
- generated from metadata in TEI-Header

Metadata in DTA titlepage

Informationen		
Quelle:	Monumenta Culinaria	
Umfang:	97 Scans	
	ca. 85798 Zeichen	
	ca. 13719 Tokens	
	/[[:alnum:]]/	
	ca. 2436 Oberflächentypes	
Schriftart:	Handschrift	
Genre:	Gebrauchsliteratur :: Kochbuch	
im DTA seit:	2012-11-07 10:53:31	
zuletzt geändert:	2012-11-22 21:48:40	
Lizenz:	Distributed under the Creative	
	Commons Attribution-	
	License.	
Grundlage dieses Digitalisats:	Thomas Gloning: Bereitstellung der Texttranskription und strukturellen Auszeichnung. (2012-11- 07T10:17:312) Bitte beachten Sie, dass die aktuelle Transkription (und Textauszeichnung) mittlerweile nicht mehr dem Stand zum Zeitpunkt der Übernahme des Werkes in das DTA entsprechen muss.	
	Bilddigitalisate (2012-11- 07T10:17:31Z)	
	Frank Wiegand: Konvertierung nach XML/TEI gemäß DTA-Basisformat. (2012-11-07T10:17:31Z)	
Weitere Informationen:		
Anmerkungen zur Transkription		

Anmerkung zu Informationen verfassen









6. Organizational aspects



system of reputation

- System of reputation for making resources available to the community
- generated from metadata











Metadata in DTA titlepage

Informationen		
Quelle:	Monumenta Culinaria	
Umfang:	97 Scans	
	ca. 85798 Zeichen	
	ca. 13719 Tokens	
	/[[:alnum:]]/	
	ca. 2436 Oberflächentypes	
Schriftart:	Handschrift	
Genre:	Gebrauchsliteratur :: Kochbuch	
im DTA seit:	2012-11-07 10:53:31	
zuletzt geändert:	2012-11-22 21:48:40	
Lizenz:	Distributed under the Creative	
	Commons Attribution-	
	NonCommercial 3.0 Unported	
Crundlago diococ	Themps Cleaning: Respitctellung der	
Digitalisats:	Texttranskription und strukturellen	
bigitalibator	Auszeichnung. (2012-11-	
	07T10:17:31Z) Bitte beachten Sie,	
	dass die aktuelle Transkription	
	(und Textauszeicnnung) mittlerweile nicht mehr dem Stand	
	zum Zeitpunkt der Übernahme des	
	Werkes in das DTA entsprechen	
	muss.	
	Thomas Gloning: Bereitstellung der	
	Bilddigitalisate (2012-11-	
	Grank Wiegands Kanuartierung nach	
	XMI /TEL gemäß DTA-Basisformat	
	(2012-11-07T10:17:31Z)	
	. ,	
Weitere Informationen:		

6. Organizational aspects



system of reputation

- publish "digital portfolio" of a scholar via reporting system
- e.g.: Thomas Gloning's digital portfolio in DTA:
 - Lindnerin,
 - Helene Lange,
 - Clara Zetkin,
 - Berliner Wochenzeitung,
 - Lina Morgenstern,
 - Jakob Bernoulli

berlin-brandenburgische

AKADEMIE DER WISSENSCHAFTEN

Transcription and TEI encoding









Metadata in DTA titlepage





- Promote culture of sharing: win-win situation
- current situation for historical texts
 - DTA: 280,000 pages of highly accurate text
 - CLARIN-D curation: 35,000 pages
 - DTAE: 220,000 pages
 - Zeno/DirectMedia: ~ 600,000 pages
 - vs. (modern text corpora (20th/21st century):
 - IDS (BBAW): 5 (2.5) billion tokens (~ 20 (10) million pages)
- \rightarrow We need more historical corpora











curation and integration of dispersed historical text resources of the 15th-19th century into the CLARIN-D infrastructure

Aims

- curate, standardize dispersed data
- quality assurance
- sustainable infrastructure
- distributed center structure
- free access to researchers via cc-licence
- improve current status of corpora for historical German from 15th-19th c.



