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

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

- 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 ...)

2. The idea of a living archive

- 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 sub*corpora*

3. Research questions & 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

4. use cases (in preparation)

- 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

a) an agreed standard for corpus encoding

- 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)

a) an agreed standard for corpus encoding

- DTABf (www.deutschestextarchiv.de/doku/basisformat)
  - 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-Ric
Encoding Initiative (TEI). Da diese Richtlinien jedoch 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 Ele
verzichtet.

Das DTA-Basisformat soll im Rahmen der DTA-Richtlinien, die daneben auch
Leitlinien des DTA sowie die Transkriptionsrichtlinien umfassen, eine uneing
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 e
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
- Formale Erschließung der Texte
- Inhaltliche Erschließung der Texte
- Übersicht über alle Basisformat-Elemente im <text>-Bereich

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

- 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
- Grundstruktur jedes TEI-Dokuments
- Strukturierung der Metadaten
- Formale Erschließung der Texte
- Inhaltliche Erschließung der Texte
- Übersicht über alle Basisformat-Elemente im <text>-Bereich

Inhaltliche Erschließung des Volltextes

Inhaltsverzeichnis dieses Dokuments

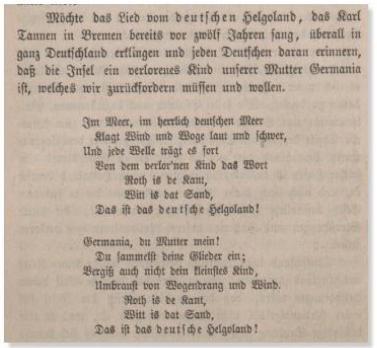
4.11 Gedichte und gebundene Sprache

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,
Umbrauft 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/>
<lg type="poem">
  <lg n="1">
    <l>Im Meer, im herrlich deutſchen Meer</l>
    <lb/>
    <l>Klagt Wind und Woge laut und ſchwer,</l>
    <lb/>
    <l>Und jede Welle trägt es fort</l>
    <lb/>
    <l>Von dem verlor'nen Kind das Wort</l>
    <lb/>
    <l>Roth is de Kant,</l>
    <lb/>
```

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.

Das Seegespenst

Informationen holen

Metadaten

Dateien – Download

Bilddateien · HTML-Dateien · generiertes TEI

Vorgesehener DTA-Verzeichnisname:

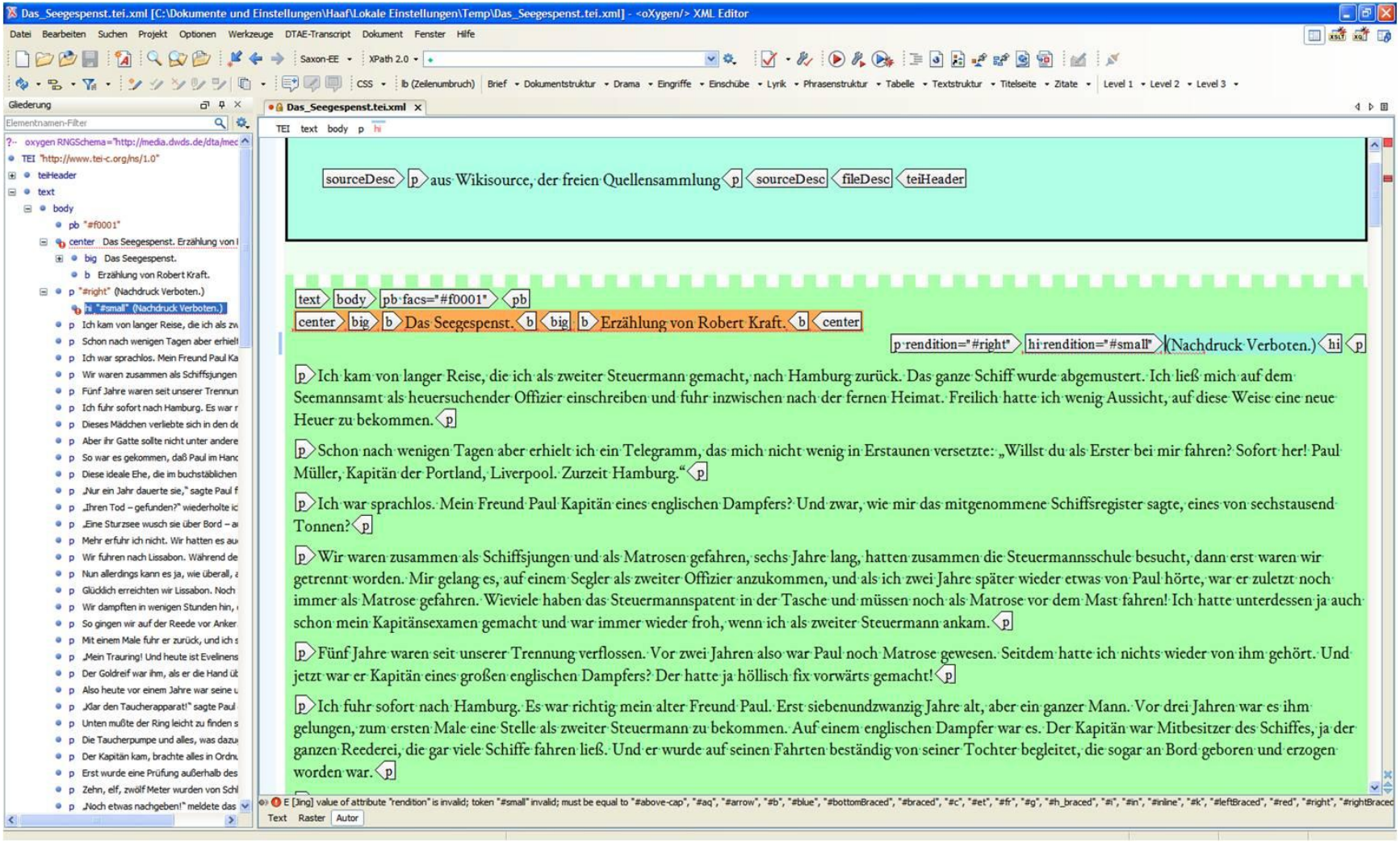
kraft_seegespenst_1910

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 ○ Antiqua ◉ keine Angabe

Überprüfen, korrigieren und ergänzen Sie ggf. die Metadaten.

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



b) Integration of texts (step 3 CLARIN-D)

- 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)

c) flexible subcorpus selection

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

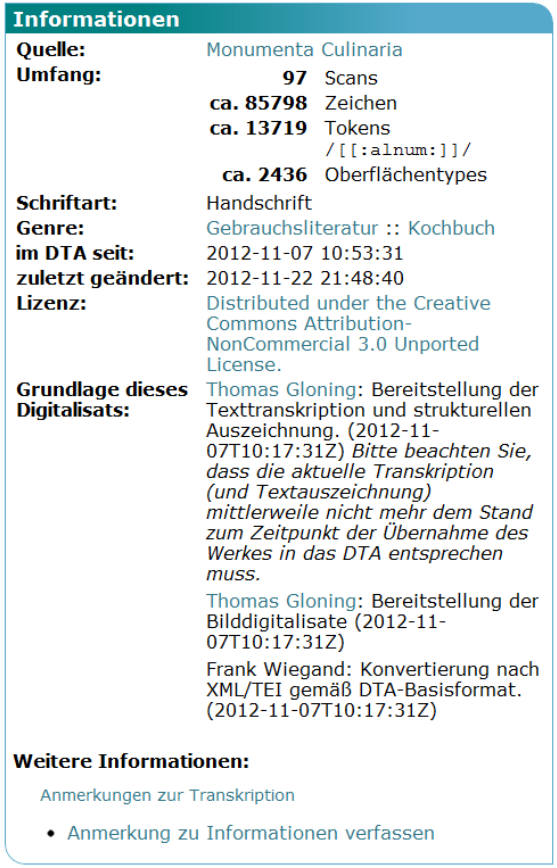
  Word #has[textclass,/\bkochliteratur\b/]

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-NonCommercial 3.0 Unported License.
Grundlage dieses Digitalisats:	Thomas Gloning: Bereitstellung der Texttranskription und strukturellen Auszeichnung. (2012-11-07T10:17:31Z) *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.* Thomas Gloning: Bereitstellung der 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

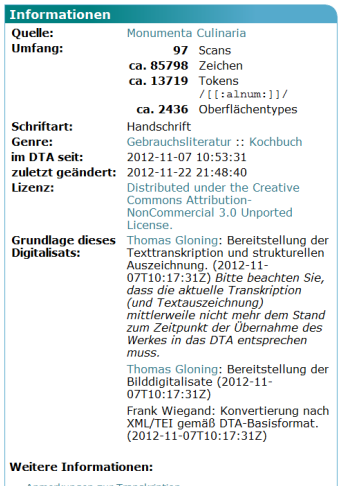
- Anmerkung zu Informationen verfassen

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 License.
Grundlage dieses Digitalisats:	Thomas Gloning: Bereitstellung der Texttranskription und strukturellen Auszeichnung. (2012-11-07T10:17:31Z) *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.* Thomas Gloning: Bereitstellung der Bilddigitalisate (2012-11-07T10:17:31Z) Frank Wiegand: Konvertierung nach XML/TEI gemäß DTA-Basisformat. (2012-11-07T10:17:31Z)

Weitere Informationen:

```
- <respStmt>
  - <persName>
      <forename>Thomas</forename>
      <surname>Gloning</surname>
    </persName>
  - <resp>
      <note type="remarkResponsibility">Bereitstellung der Texttranskription und strukturellen Auszeichnung.</note>
      <note type="remarkRevisionDTA">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.</note>
      <ref target="http://geb.uni-giessen.de/geb/volltexte/2009/7361/"/>
      <date type="importDTA">2012-11-07T10:17:31Z</date>
    </resp>
</respStmt>
```

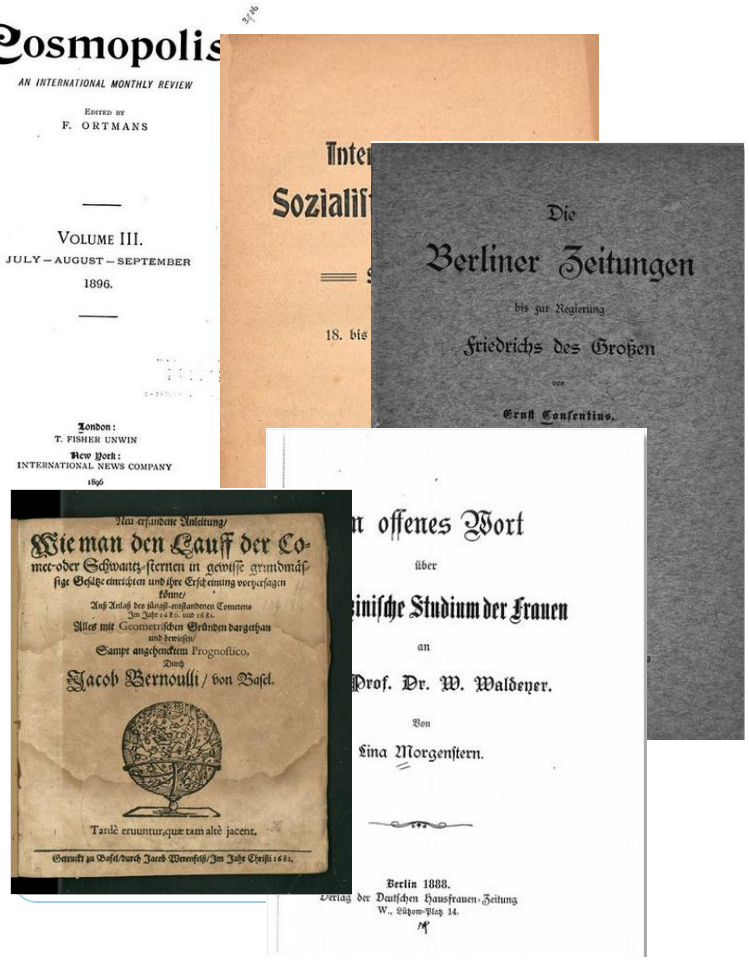
6. Organizational aspects

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

Transcription and TEI encoding

Metadata in DTA titlepage



6. Organizational aspects

- 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)

→ We need more historical corpora

7. Summary

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.

Partners in CLARIN-D

ASV Automatische Sprachverarbeitung

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

INSTITUT FÜR DEUTSCHE SPRACHE

sfs seminar für sprachwissenschaft

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN

The Language Archive

HERZOG AUGUST BIBLIOTHEK WOLFENBÜTTEL

Uni Saarland

BAS-München

Universität Stuttgart

Uni Hamburg

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

DTA

HERZOG AUGUST BIBLIOTHEK WOLFENBÜTTEL

INSTITUT FÜR DEUTSCHE SPRACHE

JUSTUS-LIEBIG-UNIVERSITÄT GIESSEN