

Historical Newspaper Corpora for the Deutsches Textarchiv. Ways of their Curation, Harmonization, and Provision to the Community

Susanne Haaf

Deutsches Textarchiv & CLARIN-D

@Berlin-Brandenburg Academy of Sciences
and Humanities



Introduction

Arnim, Achim von; Brentano, Clemens: Des Knaben Wunderhorn. Bd. 1. Heidelberg, 1806.



Des Knaben Wunderhorn

[Abbildung]
Alte deutsche Lieder

L. Achim v. Arnim. Clemens Brentano.

Heidelberg, bey Mohr u. Zimmer.
Frankfurt bey J. C. B. Mohr
1806.

Suche im Werk

Informationen zum Werk

- Titeldaten
- Verfügbarkeit Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung); CC BY-NC 3.0.
- Nutzungsbedingungen

Download dieses Werks

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

Metadaten zum Werk

TEI-Header - CMDI - Dublin Core

Ansichten dieser Seite

- gerendertes HTML (normierte Zeichen)
- gerendertes HTML (Originalzeichen)
- zugrunde liegendes XML/TEI
- Texttext
- normalisierte Orthographie
- Faksimile in hoher Auflösung

Feedback

Sie haben einen Fehler gefunden? Dann können Sie diesen über unsere Qualitätssicherungsplattform DTAQ melden.

Kommentar zur DTA-Ausgabe

Dieses Werk wurde gemäß den DTA-Transkriptionrichtlinien im Double-Keying-Verfahren von Muttersprachlern erfasst und in XML/TEI P5 nach DTA-Basisformat kodiert.

- ❖ Historical works (~1600-1900) of diverse text types
- ❖ 3,036 works (> 670,000 pages, > 162 million tokens)
- ❖ Manual transcription & TEI annotation (double-keying)
- ❖ High text quality

- ❖ Automatic linguistic analysis: tokens, lemmas, POS, orthography
- ❖ Access & Corpus Queries via DTA-Platform
- ❖ Download (CC licenses)
- ❖ CLARIN

Neue Suche · Ganze Sätze [Hilfe](#)

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

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

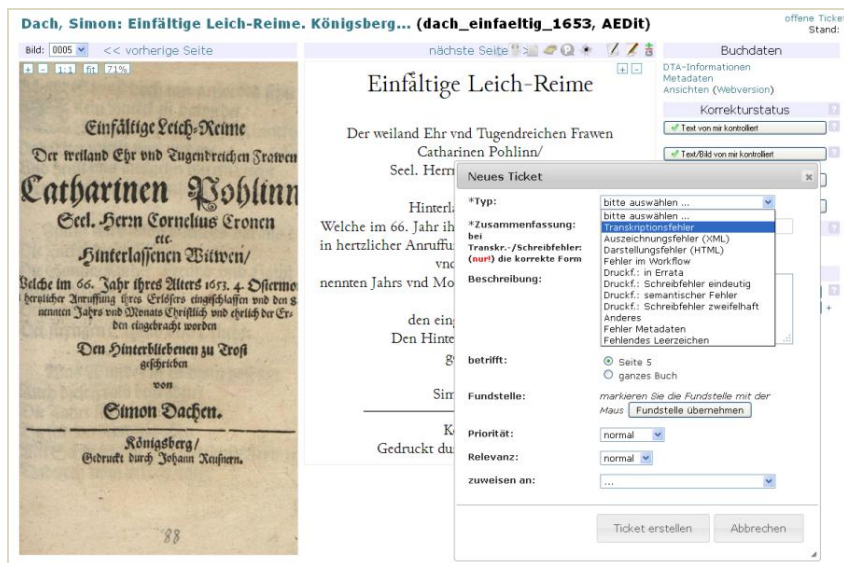
1: [achenwall_staatswissenschaft_1749:46]	... zu dem allgemeinen Natur- Völker- Staatsgeistlichen und bürgerlichen Rechte den	trefflichsten	Stoff giebet, und die Politick mit einer ...
2: [achenwall_staatswissenschaft_1749:56]	Es hat Ueberfluß an der	besten	Wolle , an Seyde, Wein, Salz ...
3: [achenwall_staatswissenschaft_1749:97]	... Handel nach Macao ist nicht nur wegen der	Chinesischen	Seyde , und allerhand Manufacturen, Thee, ...
4: [achenwall_staatswissenschaft_1749:109]	... Römische Recht ist hier nebst den Glossen in	völligem	Flor.
5: [achenwall_staatswissenschaft_1749:136]	... Gvinea, wo sie Gold, Helfenbein,	Neares	Leder , Wachs, Gummi hohlen;

```
<TEI>
  <text>
    <front>
      <pb facs="#f0007" n="[III]"/>
      <titlePage type="main">
        <docTitle>
          <titlePart type="main"> <hi
rendition="#b">Alexander von Humboldts</hi>
Rei&#x017F;e in die Aequinoktial-Gegenden</hi>
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/>
Einzig 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/>
```



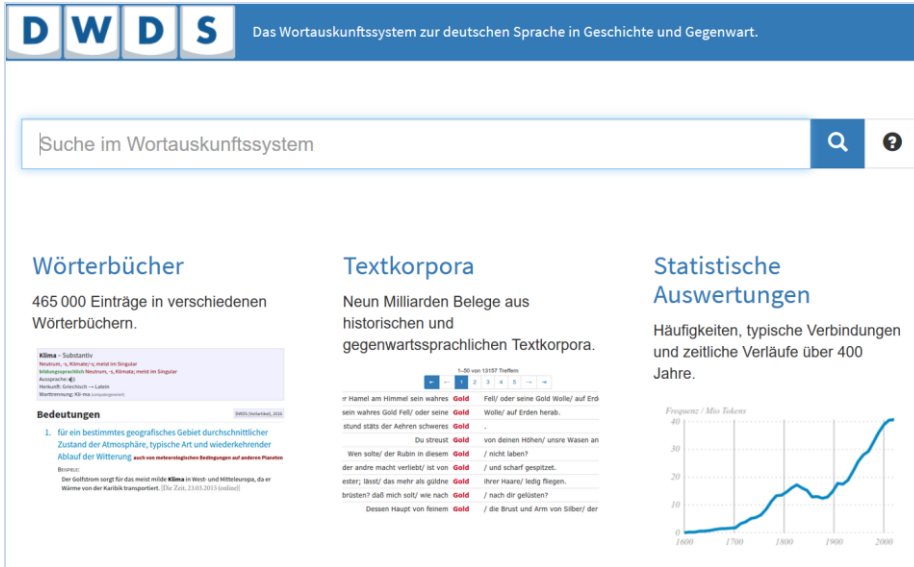
- ❖ TEI format for the annotation of historical corpora
- ❖ Schema & documentation
- ❖ Goal: Interoperability (avoid ambiguity)

cf. Haaf et al. 2014/15



- ❖ Quality Assurance Platform DTAQ
- ❖ Correct texts
- ❖ Report errors
- ❖ Add comments

<http://www.deutschestextarchiv.de/dtaq>



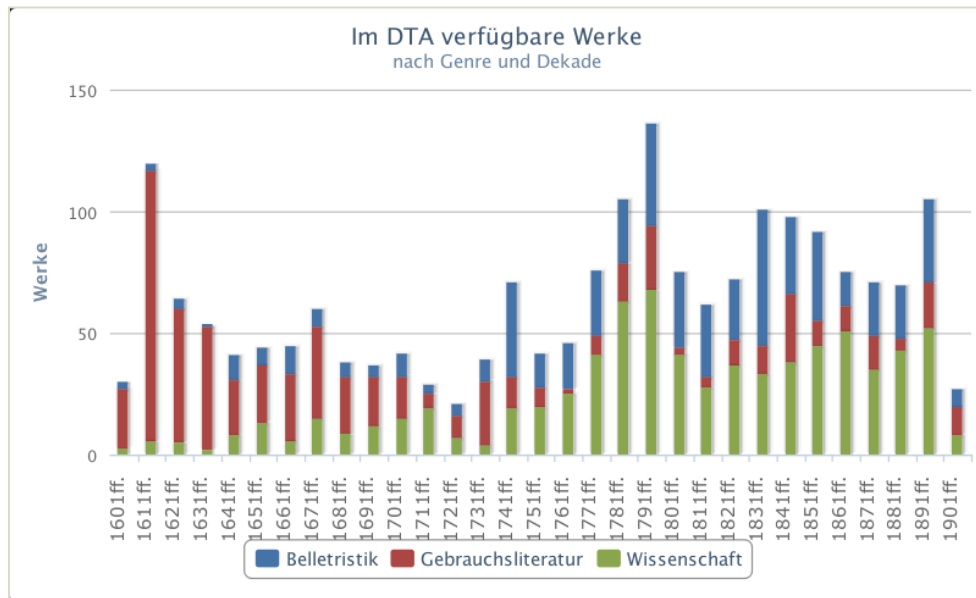
The screenshot shows the DWDS website interface. At the top, there is a search bar with the text 'Suche im Wortauskunftssystem'. Below the search bar, there are three main sections: 'Wörterbücher' (465,000 entries), 'Textkorpora' (9 billion attestations), and 'Statistische Auswertungen' (frequencies and trends over 400 years). The 'Wörterbücher' section shows the entry for 'Klima'. The 'Textkorpora' section shows a search for 'Gold' with various example sentences. The 'Statistische Auswertungen' section shows a line graph of the frequency of 'Gold' from 1600 to 2000.

- ❖ Digital Dictionary of the German Language (DWDS)
- ❖ BBAW Long-term project
- ❖ Corpus-based dictionary
- ❖ www.dwds.de

Newspaper Corpora:

- ❖ Die ZEIT (1946–current, 528m tokens)
- ❖ Tagesspiegel (1996–2005, 156m tokens)
- ❖ Berliner Zeitung (1994–2005, 237m tokens)

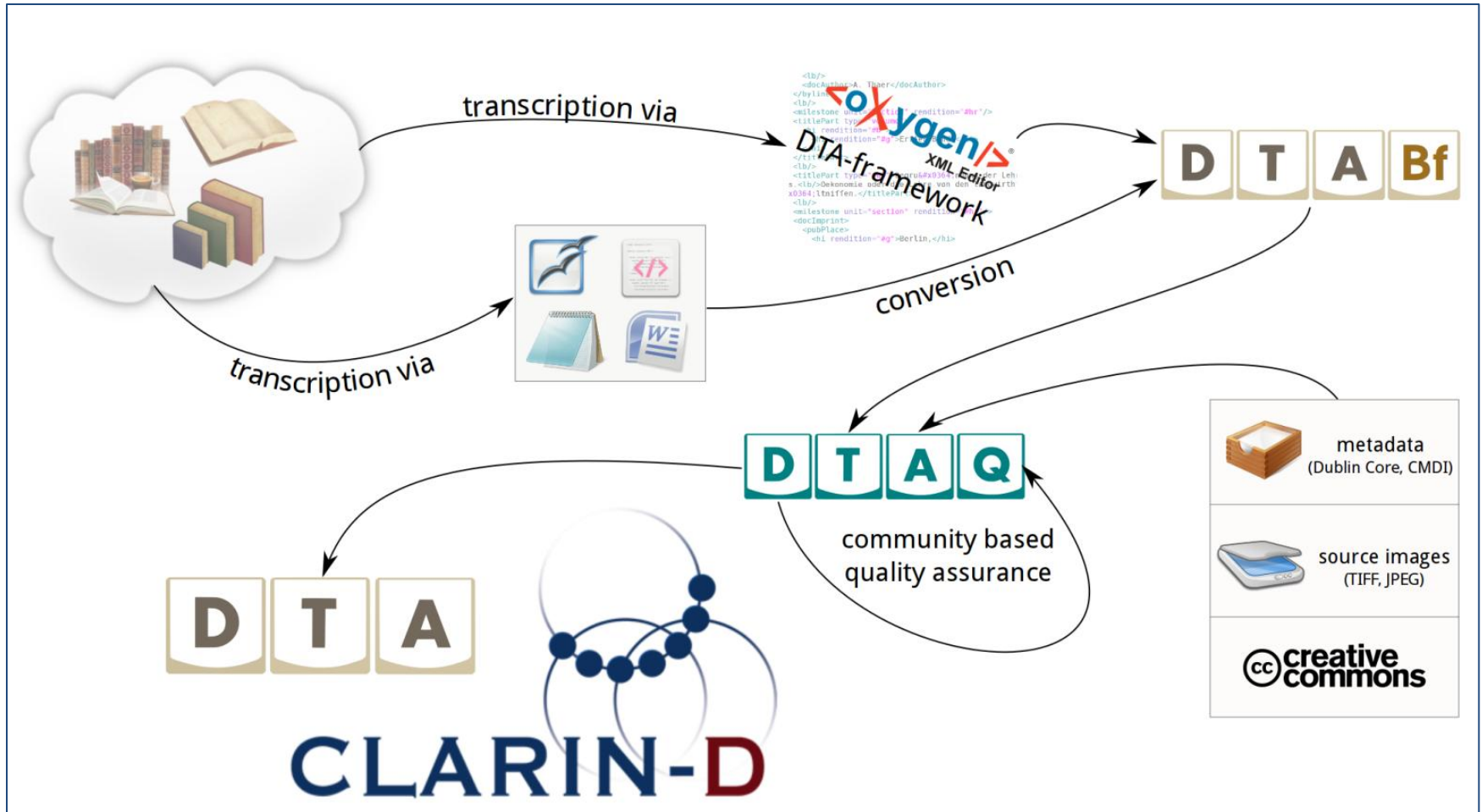




DTA Core Corpus:

- ❖ Fixed bibliography
- ❖ Various text types (Scientific, Functional, Fiction)
- ❖ No periodicals (journals or newspapers)

Current scenario: Curation of digitized newspaper texts from external projects



Everybody wins:

- ❖ DTA: Newspapers for the corpus
- ❖ Projects: Publication & quality assurance platform, connection to other corpora, repository
- ❖ Community: Joining of scattered resources, interoperability, comparison of corpora and subcorpora

Curation of Periodicals for the DTA

1. Newspapers
2. Journals

Mannheim Corpus of Historical Newspapers (MKHZ) I & II

- ❖ Partner: Prof. Dr. Stefan Engelberg, Prof. Dr. Oliver Pfefferkorn, Dr. Peter Fankhauser (IDS Mannheim)
- ❖ Project: Digitization of a Corpus of Historical Newspapers



Hamburgischer Correspondent

- ❖ Partner: Team of Prof. Dr. Britt-Marie Schuster, University of Paderborn
- ❖ Project: Digitization of the Hamburg Correspondent





Neue Rheinische Zeitung

- ❖ Partner: Dr. Jürgen Herres (BBAW)
- ❖ Project: Critical Edition of the Complete Works of Karl Marx and Friedrich Engels (MEGA)

Others

- ❖ Augsburger Allgemeine Zeitung (BBAW)
- ❖ Selected newspaper issues (Team of Prof. Dr. Thomas Gloning, Gießen University)



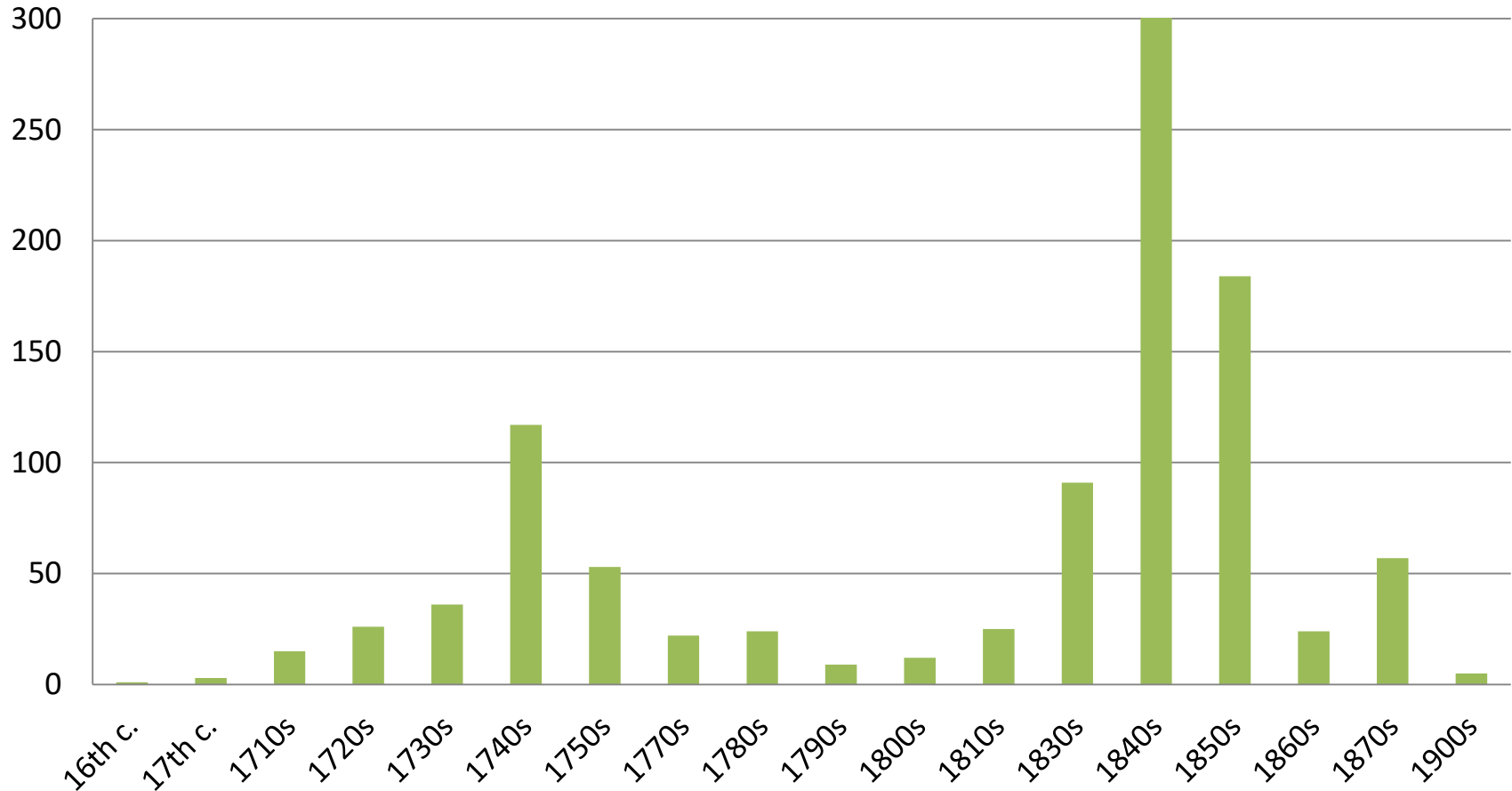
Paper	Extent	Dates	Processing for DTA
MKHZ* I	652 issues from 21 newspapers ~5000 pages	1737-1905 (mainly 19th cent.)	TUSTEP > DTABf Article structure Article types (partially)
MKHZ* II	236 issues from 6 np ~ 1850 pages		DTA workflow Article types
Hamburgischer Correspondent	204 issues 1168 pages	1712-1848	DTA workflow Article types
Neue Rheinische Zeitung	301 issues 1724 pages	1848/49	TUSTEP > DTABf Article References
Augsburger Allgemeine Zeitung	127 issues 2725 pages	1840	XML > DTABf
Others	5 np & issues	1597-1813	Other formats > DTABf

* Mannheimer Korpus Historischer Zeitungen (MKHZ)

Newspaper issues within the DTA (online [soon])



CLARIN-D



- 1720s, ~30s, ~50s, 1810s, ~60s: *less than 5 nps*
- 1740s, 1830s, ~40s, ~50s, ~70s: *5 or more nps*
- Other: *1 np*

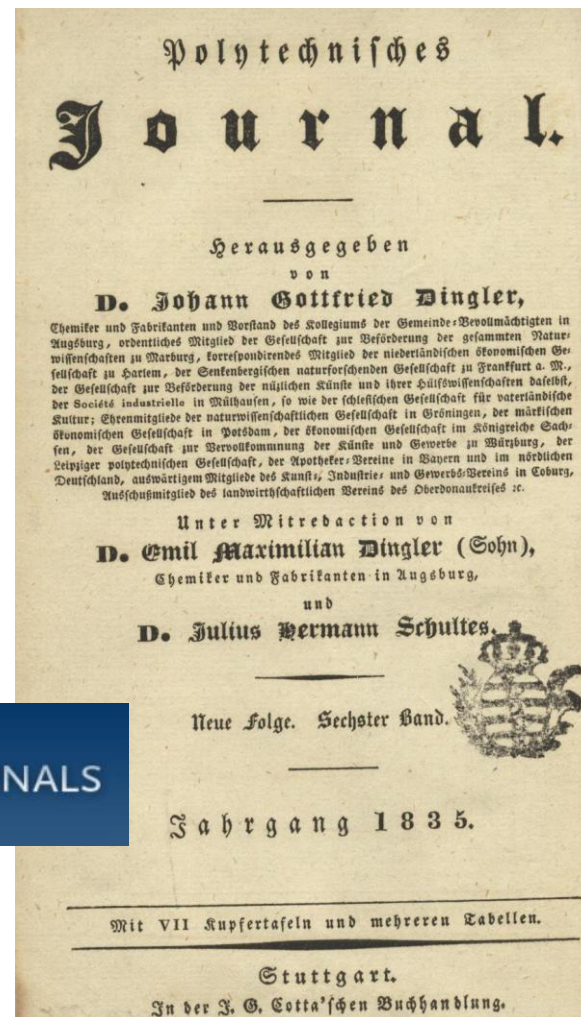
- ❖ 1840s: 440 issues
- ❖ without MKHZ II

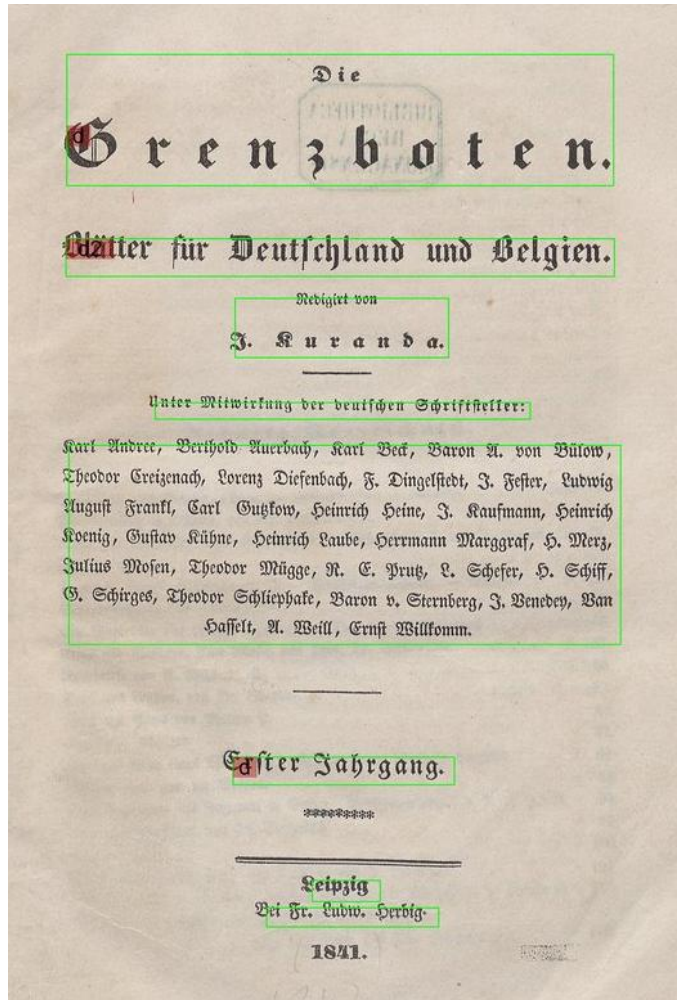
Paper	Access via DTA
MKHZ* I	Online (DTAQ, 146 issues; others in preparation, due 2016) http://www.deutschestextarchiv.de/dtaq/book
MKHZ* II	No access yet (currently digitized)
Hamburgischer Correspondent	Online (DTA): http://www.deutschestextarchiv.de/search/metadata?corpus=correspondent
Neue Rheinische Zeitung	Online (for internal use; open access by the end of 2016)
Augsburger Allgemeine Zeitung	No access yet (ready for publication; due 2016)
Others (from different sources)	Online (DTAQ) http://www.deutschestextarchiv.de/dtaq/book

* Mannheimer Korpus Historischer Zeitungen (MKHZ)

Dingler's Polytechnisches Journal

- ❖ 1820–1931
- ❖ 77m tokens
- ❖ Partner: Team of Prof. Dr. Christian Kassung, HU Berlin
- ❖ Access within DTA : Corpus queries





Die Grenzboten

- ❖ 1841–1922
- ❖ 270 vols, ~180,000 pages
- ❖ Partner: SUB Bremen
- ❖ Processing: OCR (and postcorrection) > DTABf (semi-automatically)
- ❖ Access: currently prepared; selected issues via DTAQ



Format and Processing



- Newspaper-specific additions
 - to the DTABf schema
 - to the DTABf documentation
 - Additional tagging solutions
 - for types of newspaper articles
`<div type="jPoliticalNews | jFeuilleton | ...">`
 - for introductory parts
`<div type="jEditorialStaff | jExpedition | ...">`
- etc.

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

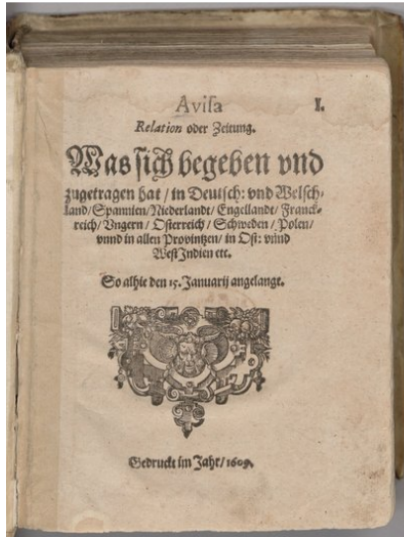
Access and Analysis

– Examples –



[Hilfe](#)

in den Titeldaten im Korpus in der Dokumentation



Text-Bild-Ansicht öffnen ...

Aviso. Relation oder Zeitung. Wolfenbüttel, 1609.

BIBLIOGRAPHISCHE ANGABEN

URN: urn:nbn:de:kobv:b4-30366-9
 Titel: Aviso
 Untertitel: Relation oder Zeitung
 Erscheinungsjahr: 1609
 Herausgeber: Schöne, Walter
 Verlag: Julius Adolph von Söhne
 Ort: Wolfenbüttel
 Auflage: 1. Auflage

INFORMATIONEN ZUM WERK

Publikationstyp: Monographie
 Verfügbarkeit: Text (TEI-XML-, HTML-, TCF-, E-Book-Fassung):
 Distributed under the Creative Commons Attribution-NonCommercial 3.0 Unported (German) License.
 Weitere Informationen: Nutzungsbedingungen.

Schriftart: Fraktur
 Genre: Zeitung
 im DTA seit: 2013-03-18 14:08:31
 Korpus: DTAE (DTA-Erweiterungen)

GRUNDLAGE DIESES DIGITALISATS

Dieses Werk ist Bestandteil der DTA-Erweiterungen. Weitere Informationen ...

Thomas Gloning: Bereitstellung der Texttranskription (2013-03-18T14:17:31Z)
 Bitte beachten Sie, dass die aktuelle Transkription (und Textauszeichnung)

Suche im Werk

[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: 413
 Zeichen: ca. 694 168
 Tokens: ca. 101 909
 Oberflächentypes: ca. 13 183

Wortwolken

- Lemmata
- Lemmata (nur Nomen)
- Types
- Types (nur Nomen)

http://www.deutschestextarchiv.de/anonym_avis_1609

nn_auswanderer01_1848 (MKHZ)

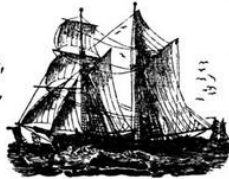
offene Tickets: 0 (0 ganzes Buch)
Stand: 2016-09-16 10:38:33

Text Text/Bild
0 - 0 - 8 0 - 0 - 8

Bild: 0001 erste Seite

nächste Seite >>

Allgemeine Auswanderungs = Zeitung.



Organ für Kunde aus deutschen Ansiedlungen, für Rath und That zu Gunsten der fortziehenden Brüder, sowie für Oeffentlichkeit in Auswanderungs-sachen überhaupt.

Wöchentliches Organ für die Rationellen Ueberlichten, Karten und Plänen, sowie mit einem Intelligenzblatte für Bekannmachungen von Behörden u. Privatn.

Montag, 3. Januar 1848.

Allgemeine Auswanderungs = Zeitung.

[Beginn Spaltensatz]

Organ für

Kunde aus deutschen Ansiedlungen,

```
<TEI>
  <text>
    <front>
      <pb facs="#f0001"/>
      <titlePart xml:id="tpm1" type="main" next="#tpm2">
        <hi rendition="#c #fr">Allgemeine Auswanderungs =
        Zeitung.</hi> </titlePart><lb/>
      <cb type="start"/>
      <titlePart type="sub"> <hi rendition="#c"> <hi
        rendition="#fr">organ</hi><lb/>
        für<lb/> <hi rendition="#fr">Kunde aus deutschen
        Ansiedlungen,</hi><lb/>
        für Rath und That<lb/>
        zu Gunsten der fortziehenden Brüder,<lb/>
        sowie für<lb/>
        Oeffentlichkeit in Auswanderungs-<lb/>
```

Notiz.

Die seit im Wochen begriffenen Auswanderungen aus Deutschland haben zunächst die Frage nach den Ursachen und Wirkungen derselben hervorgeufen; eine Frage, welche in No. 36 u. f. unseres Blattes eine umfassende und gründliche Beantwortung fand. Eine zweite, mit dieser eng zusammenhängende Frage ist die, ob der Auswanderung aus Deutschland Einhalt gethan werden könne und müsse? — Das Resultat der genaueren Forschungen mehrerer von uns getriebenen Artikel und in unserer Zeitung besprochenen Werke war übereinstimmend, daß der Auswanderung kein Einhalt gethan werden könne, so lange das Ausland dem Vaterlande gesünder, gewinniger und materielle Vorteile biete. Würde dem Auswanderungslustigen in der Heimath Dasjenige geboten, was seiner in der Fremde bairt, so würde er bald aufhören ein Auswanderungslustiger zu sein. Obst man ihm das nicht, kann man ihm das nicht geben, so ehne man ihm den Weg in die neue Heimath dadurch, daß man ihm durch Rath und That auf seiner Reise und bei seiner Niederlassung beisteht. Die Beförderung des Auswanderens in die Colonien, seine Einschiffung, Ueberfahrt, Landung und Niederlassung sind alle Punkte, welche nicht reichlich genug erwogen werden können, und denen daher auch vorzuziehen, von großer Sachkenntnis und reger Theilnahme an dem Wohl und Wehe der Schwabenzeitung die deutsche Auswanderer in No. 10, 13 und 21). Nun aber fragt es sich, ist es besser einzeln oder was bessere ist in zwanglosen Gesellschaften, oder in Vereinen auszuwandern, deren Mitglieder gewissen Gesellschaftsstatuten unterworfen sind? So ist diese Frage auch schon auf

geworfen und besprochen wurde, doch waren es nur Theoretiker, welche sich für Auswanderungs-Vereine oder Colonien erklärten, während praktische, erfahrene Männer aber jedes derartige Unternehmen den Stab brechen, weil der Zug der Auswanderung sich fast ausschließlich nach solchen Ländern richtet, deren Gesetze und Institutionen Freiheit atmen, weil daher der Junge, welcher den Kolonien durch ihre Gesellschaften anvertraut ist, dort doppelt lässig fällt und bald genug unerschrocken wird. Die Erfahrung hat auch durch eine Menge von Beispielen gezeigt, daß Colonien in freien Ländern kein Versehen haben. So wie man die allgemeinen, die Auswanderung betreffenden Fragen gründlichen, lebenshaltigen Forschungen in den Spalten unserer Zeitung unterworfen wurden und ferner noch unterliegen werden, so gingen wir auch auf das Specielle ein, welches, je nach dem Lande, welchem sich Auswanderer gewendet hatten, oder wemden wollten, einer ganz besondern Prüfung bedurfte. Um unsern Lesern einen gedrängten Ueberblick über das hierin Oeffentliche zu geben, wollen wir jeden Artikel einzeln betrachten.

Europa.

Im Jahre 1845 ließen sich viele Süddeutsche und noch mehr Schlesier durch lokale, von der russischen Regierung ausgehende Anforderungen verleiten, sich nach Polen zu wenden, um dort das verheißene Eldorado zu finden. Beamten, Militär, Säpfe und Geldern brachten die Vertrieben bald an den Vetterstab, an welchem die Weihen in ihr Vaterland juristischem. Derselben, welche dort blieben, leben in einer der reichsten Gesellschaften mündigen Knedtschaft und können mit unermüdeltem Fleiße kaum was erlangen, noch Regieren und Umwandern von ihnen fordern. Im Herbst Zeit der Polenbürger, und ganz besonders derjenige Theil derselben, welcher, der Abkennung seiner Bewohner wegen, das Sachseland genannt wird, eine nicht geringe Anzahl Deutscher, vorzüglich Württemberger, an sich, von

*) welche auch in Vorstehende erschienen und zu haben ist.

4: [hc_1080505_1848:3]	Außerordentliche	Sitzungen	können durch den Kaiser zu jeder Zeit berufen ...
5: [hc_1080505_1848:3]	... Gestalt zusammen, wenn die Zeit der ordentlichen	Sitzung	nicht früher fällt.
6: [hc_1080505_1848:3]	Die	Sitzungen	beider Häuser sind öffentlich.
7: [hc_1080505_1848:3]	Seine	Sitzungen	sind öffentlich.
8: [hc_1080505_1848:3]	... deutschen demokratischen Gesellschaft und ist in ihrer gestrigen	Sitzung	öffentlich desavouirt worden.
9: [hc_1091107_1731:1]	... zu Tuburn hingerichtet, welche in der letztern	Sitzung	zu Old- Baily zum Tode verdammet worden sind ...
10: [hc_1121507_1789:3]	... daß ihre fernere Untersuchung bis auf die nächste	Sitzung	möchte verschoben werden, weil das Parlament seinem ...
11: [hc_1182507_1789:3]	... muß ich jetzt noch etwas von den verschiedenen	Sitzungen	der Nationalversammlung sagen.
12: [hc_1182507_1789:3]	Jn der	Sitzung	vom 14ten wurden diese Königl. Antworten in Ueberlegung ...
13: [hc_1182507_1789:5]	..., wie viel das Hastingsche Verhör während dieser	Sitzung	gekostet habe.
14: [hc_1300206_1832:1]	... lassen; er hält aber den- noch seine	Sitzungen	in Vereinen von weniger als 20 Personen.
15: [hc_1300206_1832:3]	Montag und Dienstag wird das Parlament wohl keine	Sitzungen	halten, denn der erste Tag ist des ...
16: [hc_1300206_1832:3]	Jn ihrer letzten	Sitzung	am 17 Mai wurde noch eine Vorstellung um ...
17: [hc_1311808_1789:2]	Jn der	Sitzung	vom 8ten dieses kam Folgendes vor: 1 ...
18: [hc_1311808_1789:2]	Jn der außerordentlichen	Sitzung	am Sonntage, den 9ten, gieng folgendes ...
19: [hc_1330606_1832:6]	... A und B der Reform-Bill wurden in derselben	Sitzung	mit wahrhaft reißender Schnelligkeit abgemacht; an Abstimmung

DDC: Sitzung with \$con=/jPoliticalNews/ #has[flags, /correspondent/]

Complex Queries for DTAQ Texts

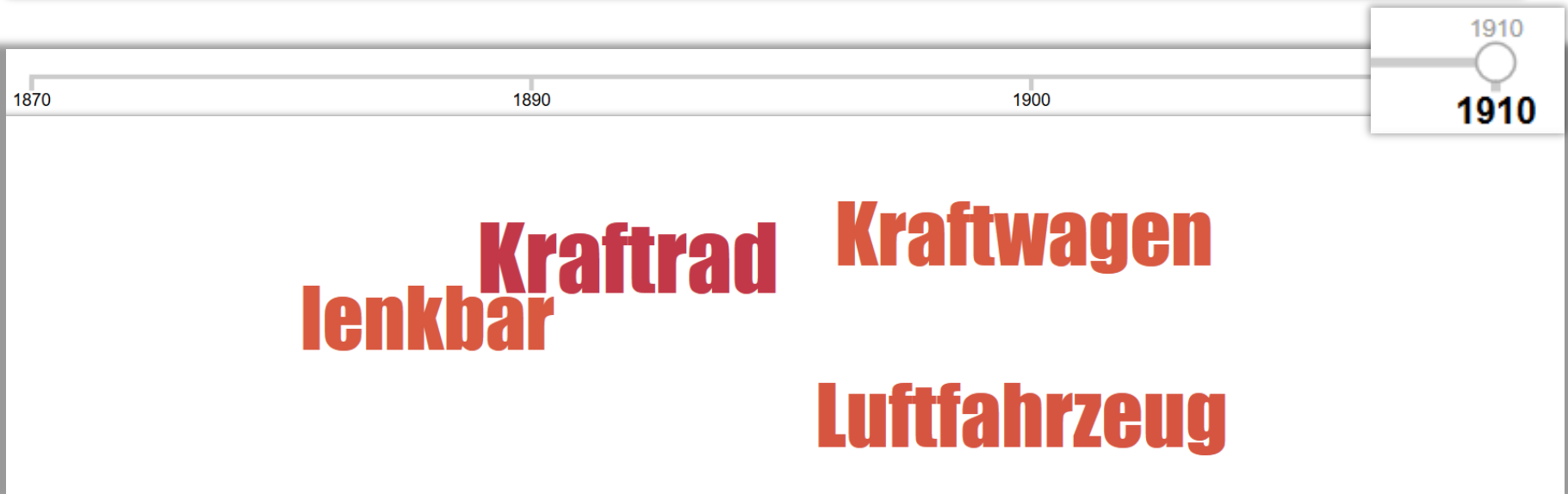
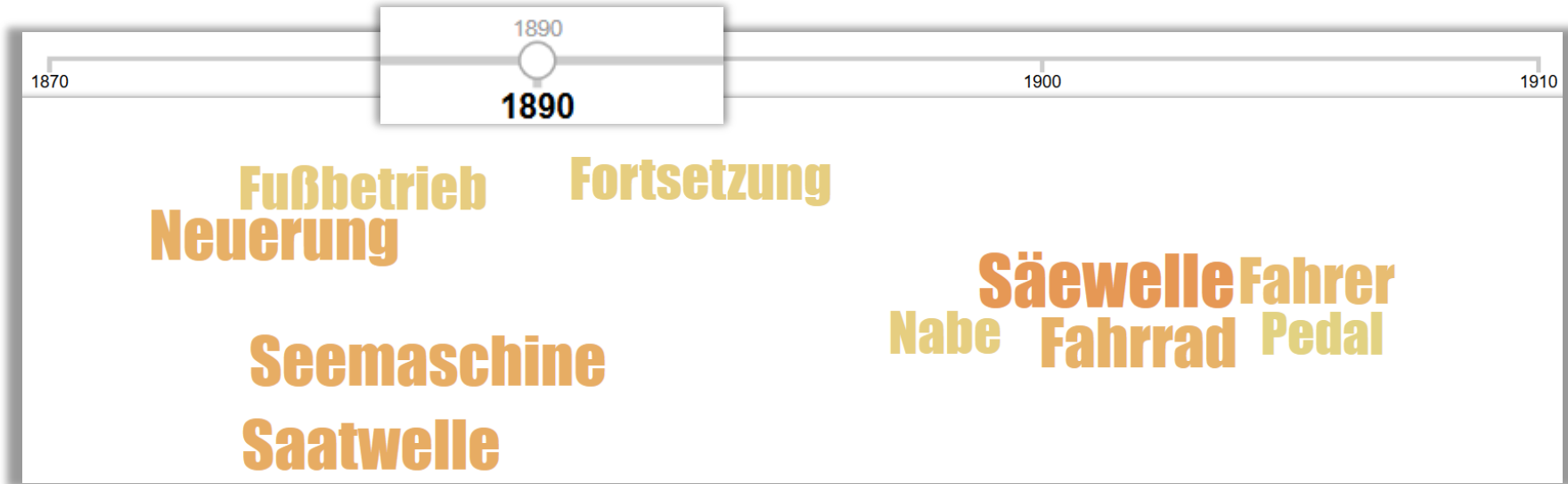
23:	[dta:hc 1390109 1789:1]	Jn der geftrigen fonntaglichen	Sitzung	ward ein Werk des Herrn Briffot de Warwille...
24:	[dta:hc 1390109 1789:2]	Jn der geftrigen fonntaglichen außerordentlichen	Sitzung	der Nationalverfammlng ward über die folgenden Artikel debattirt...
25:	[dta:hc 370209 1721:5]	... See Conpagnie verurfachte allgemeine Elend war vor eurer	Sitzung	fo groß, daß man fchwerlich zur Abhelfung...
26:	[dta:nn arbeitgeber0668 1870:4]	..., welches Koalitionsfreiheit gestattet, wurde in der	Sitzung	des östreichischen Abgeordnetenhaus vom 10. Februar angenommen.
27:	[dta:nn arbeitgeber0698 1870:6]	Jn der letzten	Sitzung	des gesetzgebenden Körpers in Paris, in der...
28:	[dta:nn arbeitgeber1068 1877:3]	... begründeten Verbandes mittelrheinischer Bildungsvereine hat in seiner gestrigen	Sitzung	den geschäftsführenden Vorstand gewählt.
29:	[dta:nn augsburg61 1871:6]	... der staatsmännischen Klugheit mit der Thiers in der	Sitzung	des 17 Febr. die Kammer von einem...
30:	[dta:nn augsburg69 1871:5]	* Zu dem Bericht über die	Sitzung	der Nationalversammlung vom 3 d. ist folgendes nachzutragen...
31:	[dta:nn augsburg69 1871:5]	Die	Sitzung	vom 4. wurde um halb 3 Uhr Nachmittags...
32:	[dta:nn augsburg69 1871:6]	Die	Sitzung	wird alsdann aufgehoben.
33:	[dta:nn augsburg69 1871:6]	Nächste	Sitzung	am Montag.
34:	[dta:nn augsburg69 1871:8]	Außerdem hat sie eine	Sitzung	der Besprechung der von allen Seiten für

DDC: Sitzung #has[textClassDWDS, /Zeitung/]

Connotations over Time with DiaCollo: Example "Fahrrad" (Bike) within Dingler



CLARIN-D



VLO / Faceted search / Search results

Search



Showing 101 to 200 of 359 results within selection for zeitung Deutsches Textarchiv (1600-1900)

Results per page: 100

Use the categories below to limit the search results to those matching the selected value(s).

Language

Collection

Deutsches Textarchiv (1600-1900)

Format

Genre

zeitung

Country

Type to search for more

Germany (359)

<< < 1 2 3 4 > >>

Staats- und Gelehrte Zeitung des hamburgischen unpartheyischen Correspondenten. Nr. 162, Hamburg, 9. Oktober 1812.



Staats/ und Gelehrte Zeitung des Hollsteinischen Correspondenten. Nr. 31, Hamburg, 12. August 1721.



Staats und Gelehrte Zeitung des Hamburgischen unpartheiischen Correspondenten. Nr. 134, Hamburg, 7. Juni 1832.



- Curation of newspapers from different sources/project partners
- Opportunistic collection
- DTA Newspaper corpus is still work in progress
- Another resource: Text&Berg-Korpus (1864–1900, 36 vols)
- Here: Mostly manual transcription and tagging
- Perspective of OCR:
 - What do researchers need?
 - Big data/good data?
 - Performance of OLR for newspapers?

Thank you!

Contact:

haaf@bbaw.de

Links:

- DTA: www.deutschestextarchiv.de
- DTA-Team: www.deutschestextarchiv.de/doku/team
- DTA and CLARIN: <http://clarin.bbaw.de/en/>
- DTABf: www.deutschestextarchiv.de/doku/basisformat_en