



Aufwertung digitaler Texte und ihre Integration in die CLARIN-D-Infrastruktur – Ein CLARIN-D-Use-Case –

Ein gescannter, OCR-erkannter und frei lizenziert vorliegender Text soll so weit aufgewertet werden, dass er

- mit digitalen Methoden untersucht werden kann,
- mit anderen Quellen kompatibel ist und
- in einem CLARIN-D-Repository dauerhaft gespeichert werden kann.

Interessant für


Forschende aus den Geschichtswissenschaften, Politikwissenschaften, Philologien

Ausgangslage:

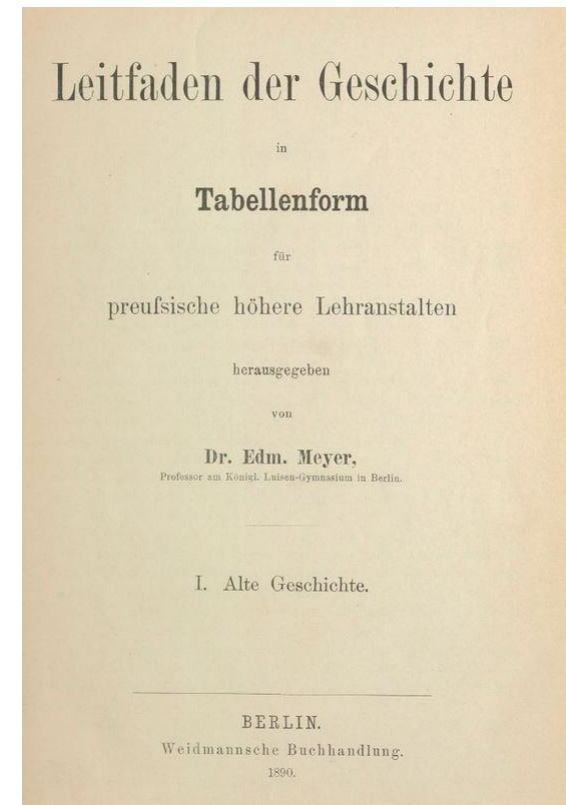
Deutschsprachiger Text liegt frei lizenziert vor

Ziel:

Integration von Texten in die CLARIN-D-Infrastruktur



Die folgenden Screenshots und Anmerkungen basieren auf den Erfahrungen des CLARIN-D-Kurationsprojektes „[Quellen des Neuen](#)“ (2015/16) der Facharbeitsgruppe „Neuere Geschichte“. Dabei wurde ein digitales Korpus für die Erforschung des Zusammenhangs von schulischer Lehre und universitärer Wissensproduktion und -vermittlung zwischen Aufklärung und Moderne erstellt.



Agenda

1. Suchen und Finden eines CLARIN-D-Zentrumspartners
2. Analyse der Ausgangslage, Festlegung der Ziele
3. Festlegen von Arbeitsablauf und Zuständigkeiten
4. Datenkuration

Digitalisierung, OCR, Datentransfer, Teilautomatische Layouterkennung, Strukturierung mit dem „ZoningTool“ des DTA, Konvertierung ins DTA-Basisformat, OCR-Korrektur und Annotationen in DTAQ, Metadaten, PIDs

5. Resultate

Verbesserte Volltextsuche, Linguistische Analyse und Suche, Diachroner und synchroner Häufigkeitsvergleich, Keywords in Context, Nutzung von Vergleichskorpora, Kollokationsanalyse

1. Schritt: Suchen und Finden eines CLARIN-D-Zentrumspartners

<http://clarin-d.de/de/aufbereiten/clarin-zentrum-finden>

Welcher Art sind die Daten?

- ✓ Geschriebene Sprache
- ✓ Deutsche Sprache
- ✓ Daten frei lizenziert
- ✓ Optische Zeichenerkennung (OCR)
- und/oder
- ✓ Digitale Editionen

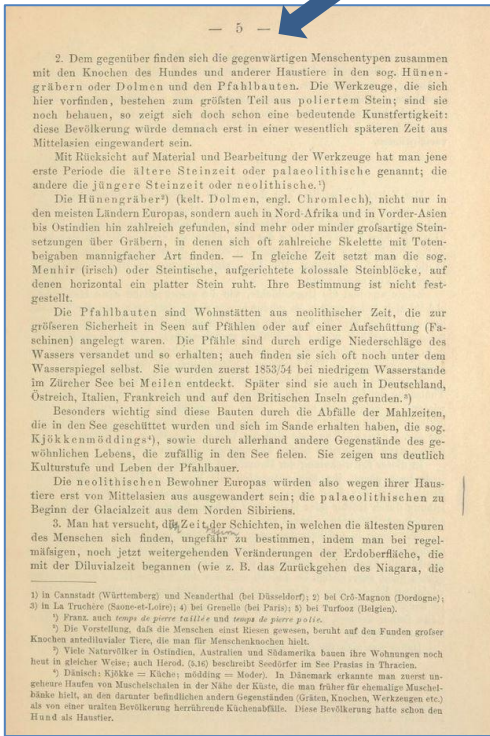


<http://clarin.bbaw.de/de/>

The screenshot shows the CLARIN-D website interface. At the top, there is a navigation bar with the CLARIN-D logo and several icons for search, evaluation, preparation, and help. Below the navigation bar, the main content area is titled "Finden Sie ein Archiv für Ihre Daten". Underneath, it lists various language centers and their specialties. The first entry, "BBAW Berlin: Deutsche Sprache, Lexika, diachrone Korpora (vor 1900), digitale Editionen, Texterfassungsmethoden (OCR)", is circled in green. Other entries include Eberhard Karls Universität Tübingen, Hamburger Zentrum für Sprachkorpora, Institut für deutsche Sprache, Ludwig-Maximilians-Universität München, MPI für Psycholinguistik, Universität Leipzig, and Universität des Saarlandes. At the bottom, there is a form with a checkbox "Sie haben deutsch-sprachige Daten oder sind in Deutschland ansässig?" and a text input field for describing the data. Below the input field, there are two radio buttons for "Gesprochene Sprache" and "Geschriebene Sprache".

2. Schritt: Analyse der Ausgangslage, Festlegung der Ziele

Ausgangspunkt ist in diesem Beispiel sind die von der digitalen Schulbuchbibliothek „GEI-Digital“ im Internet bereitgestellten digitalen Abbilder von Schulbuch-Seiten und die Ergebnisse der automatischen Texterkennung (OCR).



gei.digital
Die digitale Schulbuch-Bibliothek

Startseite Suchen Stöbern Aktuelles Über das Projekt Partner Nutzungsbedingungen ... Schnittstelle DE / EN

Bibliographische Daten

URN: urn:nbn:de:0220-gd-4103207

Titel: Alte Geschichte
Personen: Meyer, Edmund
PURL: http://gei-digital.gei.de/viewer/?resolver?urn=urn:nbn:de:0220-gd-4103207

Anzahl der digitalisierten Seiten: 23

Inhaltsverzeichnis

- Leitfaden der Geschichte in Tabellenform
- Alte Geschichte
 - Einband
 - Titelseite
 - Vorbemerkung
 - Einleitung
 - I. Hamitische Periode: Ägypter
 - II. Semitische Periode
 - III. Indogermanische Periode
 - Anhang. Zur römischen Literaturgeschichte
 - Werbung
 - Einband

Beispielseite: <http://gei-digital.gei.de/viewer/!fulltext/PPN648845621/15/>

SOLL
Fehlerfreier (texttreuer) Volltext
Strukturauszeichnung
Annotationen
Metadaten im CLARIN-D [CMDI-Format](#)

IST-Zustand des Volltextes
Viele OCR-Fehler
Kaum Strukturauszeichnung (nur Kapitel, keine Absätze u.ä.)
Keine inhaltliche Auszeichnung
Bibliographische Daten in den Formaten METS, DC, MARCXML

3. Schritt: Festlegen von Arbeitsablauf und Zuständigkeiten

Exemplarisch vorgestellt wird der Arbeitsablauf im Kurationsprojekt „Quellen des Neuen“. Die folgende Übersicht des Workflows zeigt in *Schwarz* die Aufgaben der HistorikerInnen im Projekt, in *Blau* diejenigen der MitarbeiterInnen des [CLARIN-D-Zentrums an der BBAW](#).

- I. Datenerhebung/Digitalisierung (ggfs. mit Dienstleistern)
- II. Datentransfer, *Software-Bereitstellung*
- III. *Teilautomatische Layouterkennung*
- IV. Strukturauszeichnung mit dem Zoning Tool des DTA (*Support*)
- V. *Konvertierung in das DTA-Basisformat; Metadaten-Aufnahme; Integration in die Korrekturumgebung*
- VI. *Kontrolle und Korrektur der OCR-Textdaten* und der XML-Strukturierung in DTAQ (*Support*)
- VII. *PIDs*

Es empfiehlt sich die Erstellung eines individuellen Datenmanagementplans. Ein [Template zur Erstellung](#) bieten die Webseiten von CLARIN-D.

4. Schritt: Datenkuration

I. Datenerhebung/Digitalisierung

Sind im Beispielfall bereits erledigt.

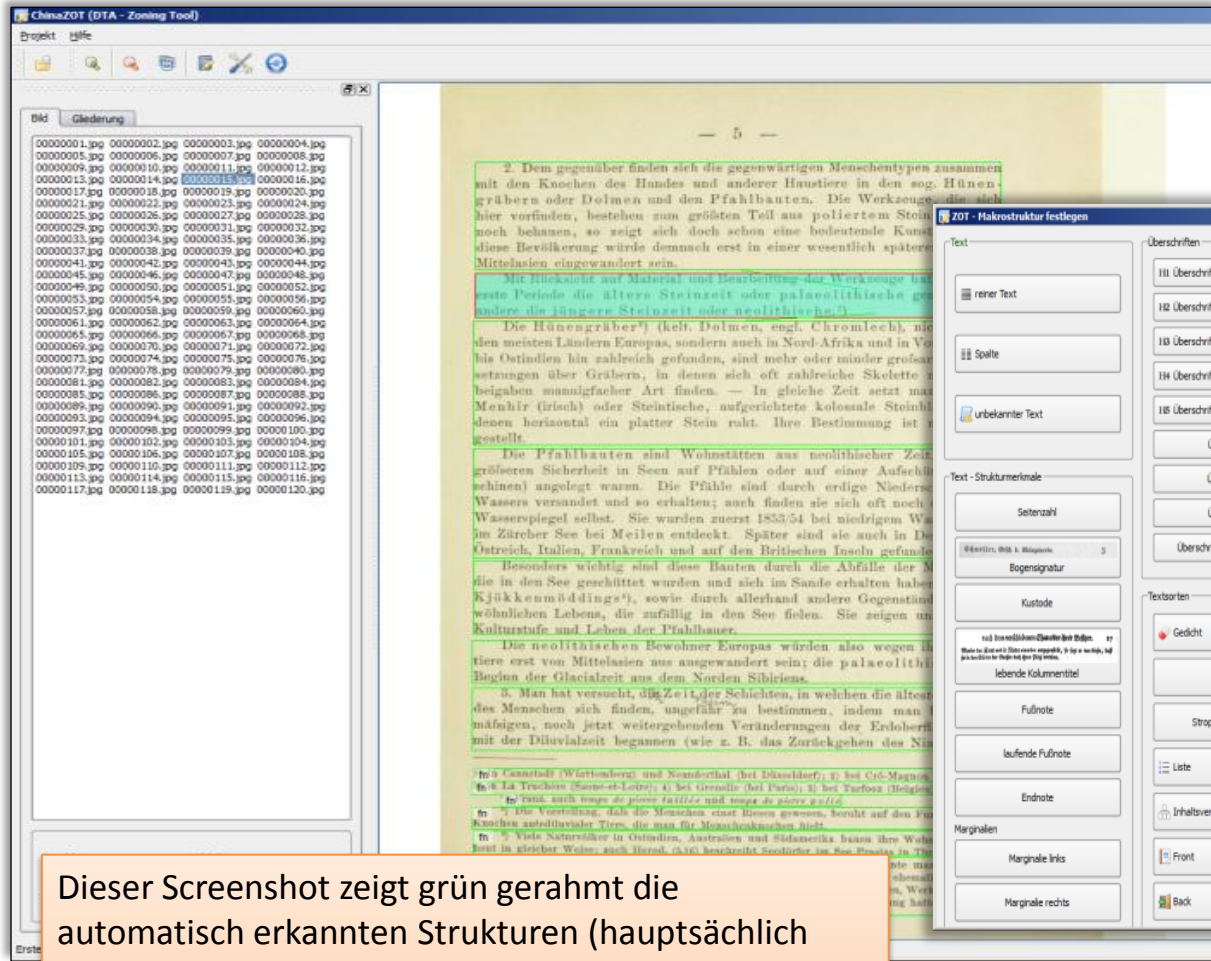
II. Datentransfer, Software-Bereitstellung

Die Bilder und OCR-erfassten Texte (JPGs und ABBYY-XML) der zur Kuration ausgewählten Ressourcen werden dem CLARIN-D-Zentrumspartner auf einem Speichermedium zugestellt oder zugänglich gemacht. Die ProjektmitarbeiterInnen erhalten Zugriff auf das „[ZoningTool](#)“ des DTA (Desktop-Version). Die Daten werden außerdem auf einem Server des CLARIN-Zentrums bereitgestellt, der den BearbeiterInnen per SSH zugänglich ist. So ist kollaboratives Arbeiten mit dem „ZoningTool“ möglich (siehe Punkt IV.).

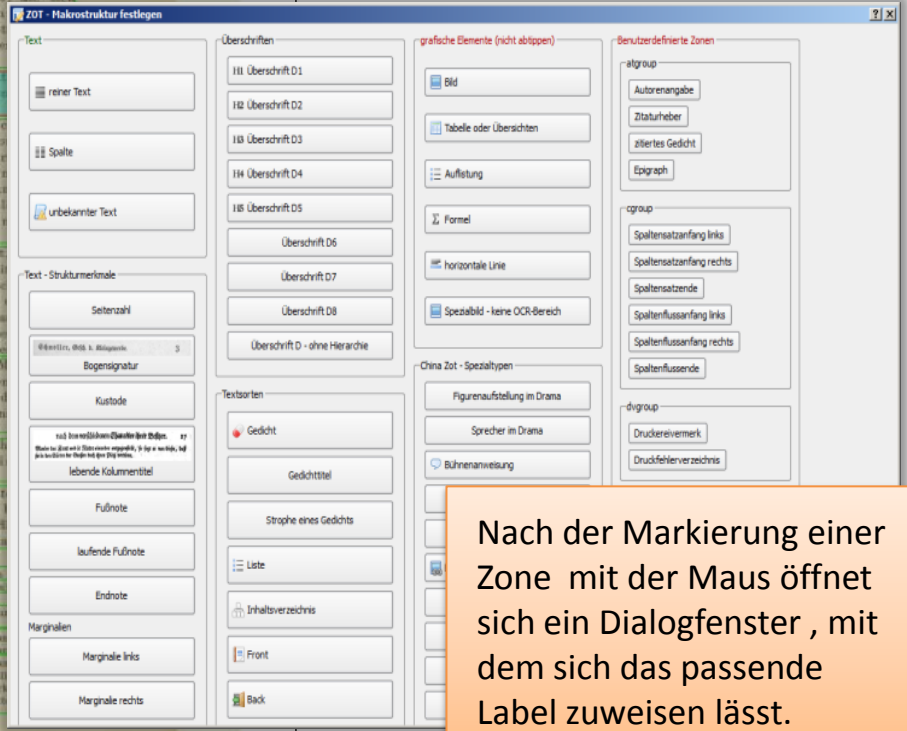
III. Automatische Layouterkennung

Zunächst erfolgt eine automatische Layouterkennung auf Grundlage der durch die ABBYY-OCR-Software vorgenommenen *Page Segmentation*. Die erkannten Textzonen werden sodann klassifiziert (*Segment Classification*). Dabei werden Faktoren wie die Position einer erkannten Textzone auf der Seite, ihre Größe, die Größe der enthaltenen Zeichen etc. einbezogen. Das Ergebnis dieses Schritts sind die Textzonen mit Koordinaten im Bild sowie die Benennungen (Labels) der dadurch repräsentierten Strukturen.

IV. Strukturauszeichnung mit dem „Zoning Tool“



Dieser Screenshot zeigt grün gerahmt die automatisch erkannten Strukturen (hauptsächlich Absätze und Fußnoten), denen bereits bestimmte Labels zugewiesen wurden (z.B. „fn“ für „Fußnote“). Diese Zonen werden nun kontrolliert und nötigenfalls berichtigt und ergänzt. Auf diesem Bild fehlt z.B. noch eine Zone+Label für die Seitenangabe („-5-“) oben.



Nach der Markierung einer Zone mit der Maus öffnet sich ein Dialogfenster, mit dem sich das passende Label zuweisen lässt.

Die Strukturauszeichnung nach TEI könnte auch direkt in den (ABBY-)XML-Dateien vorgenommen werden. Durch die Nachnutzung der ABBYY-Strukturdaten und die Verwendung des „ZoningTools“ wird sie jedoch vereinfacht und teilweise automatisiert. Die resultierenden XML-Daten können dann im XML-Editor nachkorrigiert werden.

V. Konvertierung in das DTA-Basisformat; Metadatenaufnahme; Integration in die Korrekturumgebung

Das fertige Zoning wird am CLARIN-D-Zentrum kontrolliert und in das TEI/XML-Basisformat des [Deutschen Textarchivs \(DTABf\)](#) konvertiert. Dabei werden die OCR-Textdaten mit den Zoning-Daten vereinigt, so dass ein XML-Text im TEI-basierten DTA-Basisformat (DTABf) vorliegt.

Die Metadaten werden aus vorhandenen Metadatenätzen übernommen und manuell ergänzt. Zur Erfassung einzelner Ressourcen steht auch ein [Metadatenformular](#) zur Verfügung, mit dem ein DTABf-konformer TEI-Header erzeugt wird. Dieser wird dann automatisch nach CMDI (und/oder DC) konvertiert.

Die Component Metadata Infrastructure (CMDI) wurde von CLARIN entwickelt, um die Vielzahl der existierenden Metadaten-Formate für Sprachdaten und Tools kompatibel zu machen. Einführende und weiterführende Informationen zu CMDI gibt es hier: <http://www.clarin.eu/content/component-metadata>

www.deutschestextarchiv.de/dtae/submit/clarin

Dieses Formular unterstützt die Verzeichnung von Metadaten zu einzelnen Textressourcen. Nachdem alle notwendigen Daten in die entsprechenden Felder eingetragen wurden, können Sie anhand derer einen DTABf-kompatiblen TEI-Header generieren lassen. Dieses Formular wird Ihnen vom Deutschen Textarchiv zur Verfügung gestellt.

CLARIN-D Metadatenformular zur Aufnahme einzelner Ressourcen

-Hinweise-

Felder, in denen mehrere Angaben sinnvoll erscheinen (z. B. *Verlag* oder *Schlagwörter*), können mit eben solchen versehen werden, wenn diese durch ein Semikolon getrennt werden.

Einordnung der Vorlage

<input type="checkbox"/> Eintellige Monographie (M)	<input type="checkbox"/> Unselbständiger Teil einer Monographie (z. B. Beitrag in Sammelband, Backkapitel) (DM)
<input type="checkbox"/> Teil einer mehrbändigen Monographie (MM)	<input type="checkbox"/> Unselbständige Schrift in einem Band, der Teil einer Reihe ist (DS)
<input type="checkbox"/> Selbstständiger Band einer Reihe (MS)	<input type="checkbox"/> Artikel einer Zeitschrift/Zeitung (ZA)
<input type="checkbox"/> Teil einer mehrbändigen Monographie, die Unselbständiger Teil einer Reihe ist (MMST)	<input type="checkbox"/> Zeitschriften/Zeitungsausgabe (Z)

Titel:

Haupttitel:

Untertitel (1):

Kurztitel:

Personalia:

Rolle (1):

Vorname (1):

Nachname (1):

GND-Nummer (1):

weiteres Feld zur Person:

Angaben zur Publikation der Vorlage:

Druckort:

Erscheinungsjahr:

Der Text wird in die Korrekturumgebung des Deutschen Textarchivs ([DTAQ](#)) integriert. Dort steht er BenutzerInnen nach Anmeldung zur Verfügung und kann weiter kontrolliert und bearbeitet werden.

VI. Kontrolle und Korrektur von OCR und XML in DTAQ

(1)

Die Abbildung unten zeigt den mit dem „ZoningTool“ des DTA „gezonten“ Text, wie er sich nach seiner Konvertierung in das TEI-basierte DTA-Basisformat in der Qualitätssicherungsumgebung des Deutschen Textarchivs (DTAQ) darstellt.

The screenshot shows the DTAQ web interface. At the top, there is a navigation bar with the DTAQ logo, a search bar, and user information (MKeller | Profil | ausloggen). Below this, the document title is 'meyer_geschichte_1890 (DTAE)'. The main content area displays a page with a yellow highlight over a paragraph. The right-hand sidebar contains several sections: 'Buchdaten' with DTA-Informationen, Metadaten, and Ansichten; 'Korrekturstatus' with two green checkmarks indicating that the text and image have been controlled; 'Tickets für diese Seite' with a 'neues Ticket' button; and 'Suche' with search filters for DDC and grep. The highlighted text in the main area discusses the discovery of human remains and tools in the Stone Age, mentioning the Hünengräber (dolmens) and Pfahlbauten (pile dwellings).

Nach Registrierung einsehbar: http://www.deutschestextarchiv.de/meyer_geschichte_1890/15

VI. Kontrolle und Korrektur von OCR und XML in DTAQ

(2)

Im DTAQ-eigenen Texteditor werden die OCR-Fehler berichtigt; mit dem ergänzenden XML-Editor können weitere Auszeichnungen vorgenommen werden. Diese Änderungen werden mithilfe des GIT-Versionsverwaltungssystems dokumentiert. Falls ein Phänomen auch nach Konsultation der [Richtlinien](#) nicht bearbeitet werden kann, ermöglicht ein Ticketsystem den Verweis an ExpertInnen des DTA.

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

offene Tickets: 27 (0 ganzes Buch) | Text | Text/Bild | Stand: 2015-11-23 11:19:56 | 0 - 116 - 6 0 - 1 - 121

meyer_geschichte_1890 (GEI)

Bild: 0016 : - 6 - << vorherige Seite

XML-Editor

Instant-Editor

Im XML-Editor werden Annotationen kontrolliert, korrigiert und neu hinzugefügt. Das [DTA-Basisformat](#) ermöglicht z.B. die Auszeichnung von fremdsprachlichen Elementen, Grafiken, Hervorhebungen, Named Entities usf.

Mittels GIT-Versionierung wird der Änderungsverlauf im Detail gespeichert.

Im Texteditor werden OCR-Fehler korrigiert.

Bildung von Torfmooren, das Wachstum des Nildelta oder anderer Erdschichten), die Größe 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, daß 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 Præhistorie 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 mußte, so erkennen wir doch, daß noch vor der Teilung derjenige Trieb sich im Menschen geltend machte, der bestimmt ist, die gesamte Menschheit wieder zur Einheit zurückzuführen, der Trieb zur Kultur, abzässig zu tuße.

§ 7.

alsbald sich teilen mußte, so erkennen wir doch, daß noch vor der Teilung derjenige Trieb sich im Menschen geltend machte, der bestimmt ist, die gesamte Menschheit wieder zur Einheit zurückzuführen, der Trieb zur Kultur, abzässig zu tuße.

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 Præhistorie 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.

VI. Kontrolle und Korrektur von OCR und XML in DTAQ

(3)

Die unten stehende Abbildung zeigt eine Teilansicht der fertig korrigierten Seite. Rechts wurde ein Ticket für das noch ungelöste Problem angelegt.

The screenshot displays the DTAQ (Digital Text Annotation Query) interface. The main content area shows a document page titled "meyer_geschichte_1890 (DTAE)". The text on the page discusses Neolithic sites and the discovery of the first house in Europe. A correction ticket is visible on the right side of the page, indicating a problem with the XML markup. The ticket is titled "Auszeichnungsfehler (XML)" and describes an issue with the markup for the first house in Europe. The ticket is currently in the "neue" (new) state.

DTAQ

zuletzt gelesen · Hilfe · Zufallsseite

MKeller | Profil | ausloggen

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

DTA-Informationen
Metadaten
Ansichten
nächstes Ticket

Korrekturstatus

✗ Text doch nicht kontrolliert
✓ MKeller, 2015-09-25T12:19

✓ Text/Bild von mir kontrolliert

Tickets für diese Seite

neues Ticket

#71244 [2015-09-25T11:44, MKeller]
Auszeichnungsfehler (XML)
1) in Cannstadt (Württemberg) und Neanderthal (bei Düsseldorf); 2) bei Crô-Magnon (Dordogne);
Fortlaufende Fußnote

Suche

DDC
grop

vorherige Seite

nächste Seite

0015 : - 5 -

1000

griecher Art finden. — In gleiche Zeit setzt man die sog. Menhir (irisch) oder Steintische, aufgerichtete kolossale Steinblöcke, auf denen horizontal ein platter Stein ruht. Ihre Bestimmung ist nicht festgestellt.

Die Pfahlbauten sind Wohnstätten aus neolithischer Zeit, die zur größeren Sicherheit in Seen auf Pfählen oder auf einer Aufschüttung (Faschinen) angelegt waren. Die Pfähle sind durch irdige Niederschläge des Wassers versandet und so erhalten; auch finden sie sich oft noch unter dem Wasserspiegel selbst. Sie wurden zuerst 1853/54 bei niedrigem Wasserstande im Zürcher See bei Meilen entdeckt. Später sind sie auch in Deutschland, Osterreich, Italien, Frankreich und auf den Britischen Inseln gefunden.¹⁾

Besonders wichtig sind diese Bauten durch die Abfälle der Mahlzeiten, die in den See geschüttet wurden und sich im Sande erhalten haben, die sog. Kjökkennöddings²⁾, sowie durch allerhand andere Gegenstände des gewöhnlichen Lebens, die zufällig in den See fielen. Sie zeigen uns deutlich Kulturstufe und Leben der Pfahlbauer.

Die neolithischen Bewohner Europas würden also wegen ihrer Haustiere erst von Mittelasien aus ausgewandert sein; die palaeolithischen zu Beginn der Glacialzeit aus dem Norden Sibiriens.

3. Man hat versucht, die Zeit der Schichten, in welchen die ältesten Spuren des Menschen sich finden, ungefähr zu bestimmen, indem man bei regelmäßigen, noch jetzt weitergehenden Veränderungen der Erdoberfläche, die mit der Diluvialzeit begannen (wie z. B. das Zurückgehen des Niagara, die

wöhnlichen Lebens, die zufällig in den See fielen. Sie zeigen uns deutlich Kulturstufe und Leben der Pfahlbauer.

Die neolithischen Bewohner Europas würden also wegen ihrer Haustiere erst von Mittelasien aus ausgewandert sein; die palaeolithischen zu Beginn der Glacialzeit aus dem Norden Sibiriens.

3. Man hat versucht, die Zeit der Schichten, in welchen die ältesten Spuren des Menschen sich finden, ungefähr zu bestimmen, indem man bei regelmäßigen, noch jetzt weitergehenden Veränderungen der Erdoberfläche, die mit der Diluvialzeit begannen (wie z. B. das Zurückgehen des Niagara, die

1) in Cannstadt (Württemberg) und Neanderthal (bei Düsseldorf); 2) bei Crô-Magnon (Dordogne);
3) in La Truchère (Saone-et-Loire); 4) bei Grenelle (bei Paris); 5) bei Turfooz (Belgien).
2) Franz. auch *temps de pierre taillée* und *temps de pierre polie*.
3) Die Vorstellung, daß die Menschen einst Riesen gewesen, beruht auf den Funden großer Knochen antediluvialer Tiere, die man für Menschenknochen hielt.
4) Dänisch: Kjökke = Küche; mödding = Moder). In Dänemark erkannte man zuerst ungeheure Haufen von Muschelschalen in der Nähe der Küste, die man früher für ehemalige Muschelbänke hielt, an den darunter befindlichen andern Gegenständen (Gräten, Knochen, Werkzeugen etc.) als von einer uralten Bevölkerung herrührende Küchenabfälle. Diese Bevölkerung hatte schon den Hund als Haustier.

1) Französisch auch *temps de pierre taillée* und *temps de pierre polie*.
2) Die Vorstellung, daß die Menschen einst Riesen gewesen, beruht auf den Funden großer Knochen antediluvialer Tiere, die man für Menschenknochen hielt
3) Viele Naturvölker in Ostindien, Australien und Südamerika bauen ihre Wohnungen noch heute in gleicher Weise; auch Herod. (3,16) beschreibt Seedorfer im See Prasias in Thracien.
4) Dänisch: Kjökke = Küche; mödding = Moder). In Dänemark erkannte man zuerst ungeheure Haufen von Muschelschalen in der Nähe der Küste, die man früher für ehemalige Muschelbänke hielt, an den darunter befindlichen andern Gegenständen (Gräten, Knochen, Werkzeugen etc.) als von einer uralten Bevölkerung herrührende Küchenabfälle. Diese Bevölkerung hatte schon den Hund als Haustier.

none

VII. Vergabe von PIDs

Die Daten werden nun am CLARIN-D Zentrum mit persistenten Identifikatoren (persistant identifiers, PIDs) versehen, die eine Langzeitarchivierung und die dauerhafte Referenzierbarkeit gewährleisten.

5. Schritt: Resultate (1): Sichtbarkeit und Langzeitverfügbarkeit

1 Treffer

ID	Autor	Titel	Typ	Bibliographie	Datum	Herausgeber	Textso
dta:2627	Tewes, Hermann	Menschenrassen und Völkertypen – Material zu geographischen Unterredungen auf der Oberstufe mehrklassiger Volks- und Bürgerschulen. Zugleich eine Erläuterung der gleichnamigen Bilderwerke (vollständige digitalisierte Ausgabe)	Text	Tewes, Hermann: Menschenrassen und Völkertypen. Bd. 2. 2. Aufl. Leipzig, 1913.	1913-01-01	Deutsches Textarchiv (DTA-Erweiterungen)	Gebrauchslit Schulbuch

Die kuratierten Ressourcen sind jetzt über [das Repository](#) und die weiteren Angebote des betreuenden CLARIN-D-Zentrums verfügbar und auffindbar.

Virtual Language Observatory
Explore the world of language resources and technology from different perspectives

VLO > Faceted search > Search: "Tewes" > Sele 1 results

SEARCH
Tewes

SEARCH RESULTS
1 results

Tewes, Hermann: Menschenrassen und Völkertypen. Bd. 2. 2. Aufl. Leipzig, 1913. Expand - COLLECTION

Sie können über CLARINs [Virtual Language Observatory](#) (VLO) gefunden ...

Content Search Aggregator Help

Search text Vlieβhaarige

Search for Any Language

German Text Archive (DTA) – Berlin-Brandenburg Academy of Sciences and Humanities

Ein anderer Forscher teilt die Menschen nach ihrem Haar in Wollhaarige und Schlichthaarige und jene wieder in Büschel- und Vlieβhaarige, diese in Straff- und Lockenhaarige ein.

Tokio 27. Tolderinos 45. Toldo 45. Tomahawk 53. Topnaars 59. Totem 54. Tschechen 10. Tuaregs 9. Tungusen 11. Tunis 9. Turktatarn 11. V. Veldschoendragers 60. Vereinigte Staaten 6. Vlieβhaarige 7.

... und über CLARINs [Federated Content Search](#) (FCS) im Verbund mit weiteren CLARIN-Korpora durchsucht werden.

5. Schritt: Resultate (2): Volltextsuchen im korrekten Text möglich

 Zeigte uns nun die Urgeschichte, wie die ursprüngliche eine Menschheit
 alsbald sich teilen **mufste**, so erkennen **Avir doch**, **dafs** noch vor der Teilung
 derjenige Trieb sich im Menschen geltend machte, der bestimmt ist, die ge-
 trennte Menschheit wieder zur Einheit zurückzuführen, der Trieb zur Kultur,
 d. h. der Trieb, sein äußeres (materielles) und geistiges Leben unablässig zu
 vervollkommen: das Tier verharrt unabänderlich auf derselben Stufe.
 Waren es Schädel- und Knochenfunde, die uns in Verbindung mit der
 Thatsacie der fünf Rassen die Zerstreung der Menschheit über die Erde nach-
 wiesen, so zeigen uns andere Funde in Verbindung mit ei-
 igen

OCR-Fehler wurden korrigiert.
So wird sichergestellt, dass z.B.
Volltextsuchen auch alle Treffer
im Text finden können.

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 ge-
trennte Menschheit wieder zur Einheit zurückzuführen, der Trieb zur Kultur,
d. h. der Trieb, sein äußeres (materielles) und geistiges Leben unablässig zu
vervollkommen: das Tier verharrt unabänderlich auf derselben Stufe.

Waren es Schädel- und Knochenfunde, die uns in Verbindung mit der
Thatsache der fünf Rassen die Zerstreung der Menschheit über die Erde nach-
wiesen, so zeigen uns andere Funde in Verbindung mit einer Betrachtung
derjenigen Völker, die sich nur wenig über den Urzustand erhoben haben,



5. Resultate (3): Linguistische Abfragen mit DDC (Suchabfragesprache des [DTA/DWDS](#))

D T A

**Linguist.
Analyse**

Zeigte uns nun die Urgeschichte, wie die ursprüngl
alsbald sich teilen **musfte**, so erkennen **wir doch**, **dafs** no
derjenige Trieb sich im Menschen geltend machte, der b
trennte Menschheit wieder zur Einheit zurückzuführen,
d. h. der Trieb, sein äußeres (materielles) und geistiges I
vervollkommen: das Tier verharret unabänderlich auf de

Waren es Schädel- und Knochenfunde, die uns in V
Thatsache der fünf Rassen die Zerstreung der Mensch
wiesen, so zeigen uns andere Funde in Verbindung mit e
derjenigen Völker, die sich nur wenig über den Urzustan
der sog. Naturvölker, wie sich die Völker, ehe sie in die Geschichte ein-
traten, nach bestimmten Seiten hin entwickelt haben müssen.

Das DTA bietet eine automatische linguistische Analyse historischer Wortformen, die auch zur orthographischen Normierung von Texten genutzt wird. Zusammen mit der Indizierung durch die [Suchmaschine DDC](#) ermöglicht dies z.B. (a) bei der Suche nach „Tatsache“ auch Schreibvarianten wie „Thatsache“ zu finden und umgekehrt; (b) bei der Suche nach einem Lemma wie „Tatsache“ auch die zugehörigen Wortformen (z.B. „Tatsachen“) zu finden.

Waren es **Schädel-** und Kno

Tatsache
nachwies
derjenige
der sog.

NN / +exlex,-id,-xid,-msafe,+moota,-mootxy
alt:
neu:
richtig:

Waren es Schädel- u

Tatsache
nachwies

POS-Tag: NN
Lemma: Tatsache

5. Resultate (4): Frequenzanalyse und Vergleich mit Referenzkorpora möglich



1: [dta:weigeldt_erdkunde_1912:26]

Einer dieser **Schädel** ist von **Blumenbach** in seinem vortrefflichen
kranilogischen Werke...

2: [dta:fuhrrott_neanderthaler_1865:68]

die allgemeine Aehnlichkeit dieses bekannten **Blumenbach'schen**...

3: [dta:schauberg_freimaurerei02_1861:56]

... Haller, Linné, Buffon, Cuvier, **Blumenbach**, Kant, Herder, Steffens, Rudolph...

4: [dta:humboldt_natur02_1808:3]

Einer dieser **Schädel** ist von Herrn **Blumenbach** in feinem vortrefflichen
kranilogischen...

5: [dta:blumenbach_menschengeschlecht2_1...:2]

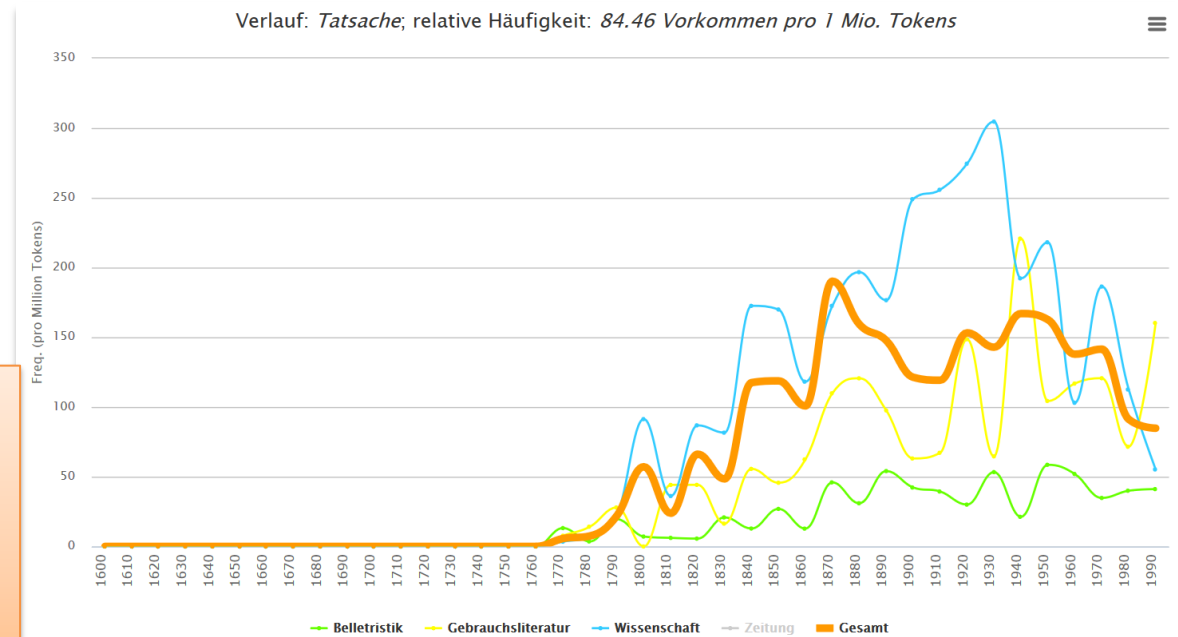
naturhistorischen Abbildungen 1. Heft...

chlecht 17...:23]

Eine Abbildung ihrer **Schädel** f. in Herrn **Blumenbach's** naturhistorischen
Abbildungen 1....

Durch die Integration in das Deutsche Textarchiv (DTA) kann man mit einer Keyword-in-Context (KWIC)-Darstellung Textstellen aus den neu integrierten und allen bereits im DTA vorhandenen Werken direkt miteinander vergleichen.

Im DTA lässt sich auch die Konjunktur eines Begriffes über die Zeit und in verschiedenen Subkorpora bestimmen.



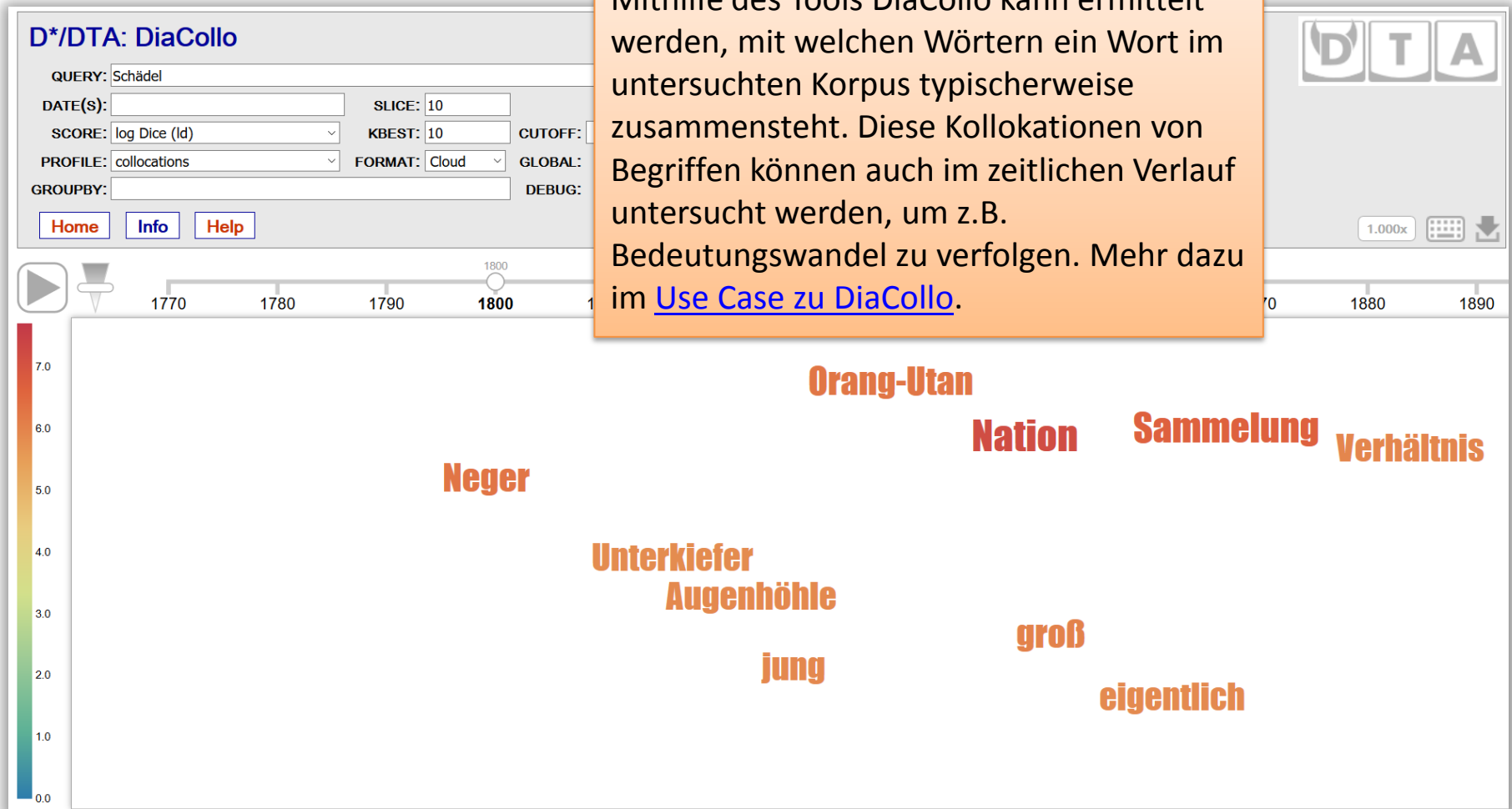
5. Resultate (5): Kollokationsanalyse möglich



Kollokationsanalyse zu "Schädel" in DiaCollo

Unter: <http://kaskade.dwds.de/dstar/dta/diacollo>

Mithilfe des Tools DiaCollo kann ermittelt werden, mit welchen Wörtern ein Wort im untersuchten Korpus typischerweise zusammensteht. Diese Kollokationen von Begriffen können auch im zeitlichen Verlauf untersucht werden, um z.B. Bedeutungswandel zu verfolgen. Mehr dazu im [Use Case zu DiaCollo](#).



Anwendungsmöglichkeiten

Vorher:

- Volltextsuche im fehlerhaften Text

Nachher:

- Volltextsuche im korrigierten Text
- Linguistische Abfragen mit DDC (Suchabfragesprache des DTA/DWDS)
- Vergleich mit Referenzkorpora
- Kollokationsanalyse mit DiaCollo (vgl. z.B. <http://clarin-d.de/de/kollokationsanalyse-in-diachroner-perspektive>)
- Statistische Analysen (Stilometrie, Wortschatz, Frequenzen, Lemmalisten, Verteilungen, ...)
- Verschiedene (Meta-)Datenformate, um die Daten nachzunutzen (TEI, TCF; CMDI, DC)
- Nutzung von CLARIN-Angeboten:
Auffinden von relevanten Ressourcen im Virtual Language Observatory (VLO)
<https://vlo.clarin.eu/> und über die Federated Content Search (FCS)
(<https://clarin.eu/contentsearch/>)
- Auswerten von Ressourcen mit verschiedenen Tools: <http://clarin-d.de/de/auswerten>
- Hosting und Bereitstellung der Daten in CLARIN-Repositories
➤ hier: <http://clarin.bbaw.de>