GEORG ECKERT INSTITUT Leibniz-Institut für internationale Schulbuchforschung

berlin-brandenburgische AKADEMIE DER WISSENSCHAFTEN

CLARIN-D

Aufwertung digitaler Texte und ihre Integration in die CLARIN-D-Infrastruktur

– Ein CLARIN-D-Use-Case –

Ein gescannter, OCR-erkannter und frei lizensiert 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-Repositorium dauerhaft gespeichert werden kann.

Interessant für

Forschende aus den Geschichtswissenschaften, Politikwissenschaften, Philologien

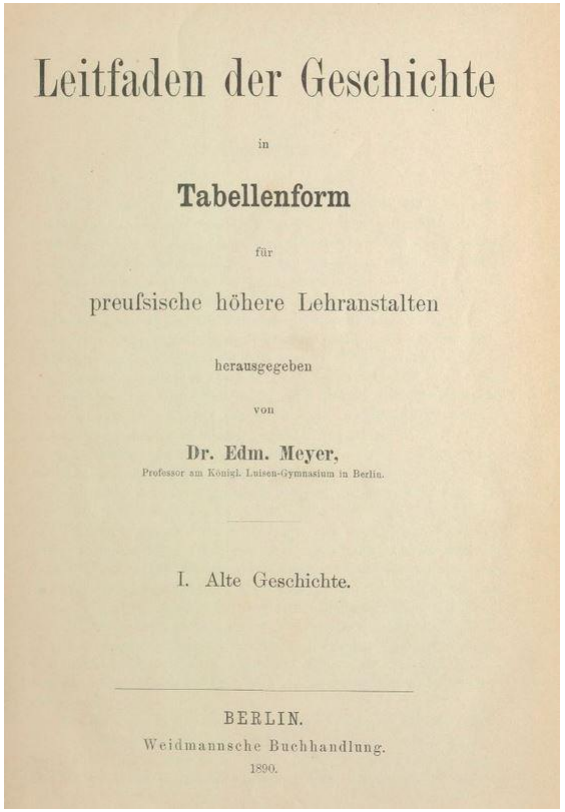
Ausgangslage:

Deutschsprachiger Text liegt frei lizensiert 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.



Leitfaden der Geschichte

in

Tabellenform

für

preufsische höhere Lehranstalten

herausgegeben

von

Dr. Edm. Meyer,

Professor am Königl. Luisen-Gymnasium in Berlin.

I. Alte Geschichte.

BERLIN.

Weidmannsche Buchhandlung.

1890.

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

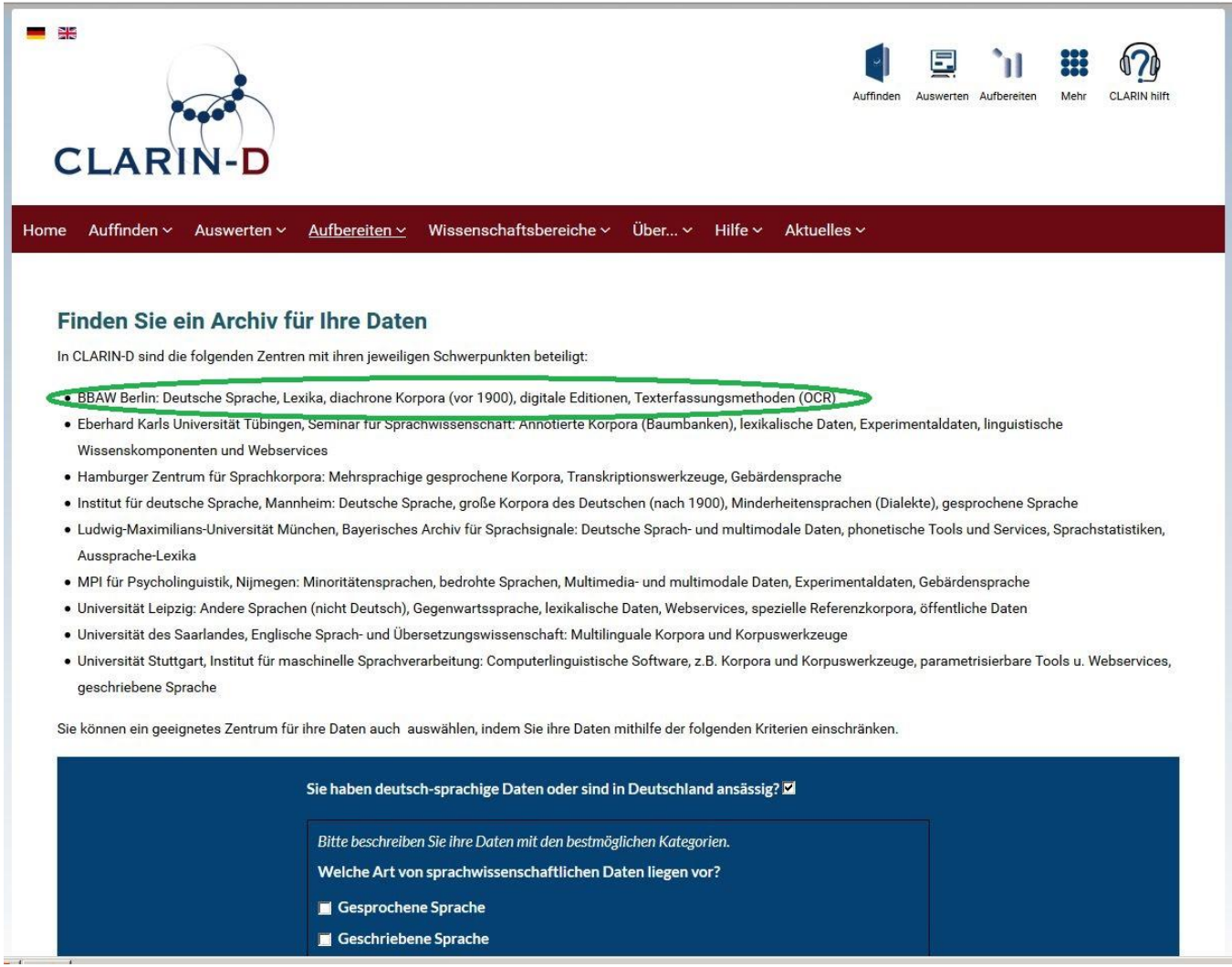
Welcher Art sind die Daten?

- ✓ Geschriebene Sprache
- ✓ Deutsche Sprache
- ✓ Daten frei lizensiert
- ✓ Optische Zeichenerkennung (OCR)

und/oder

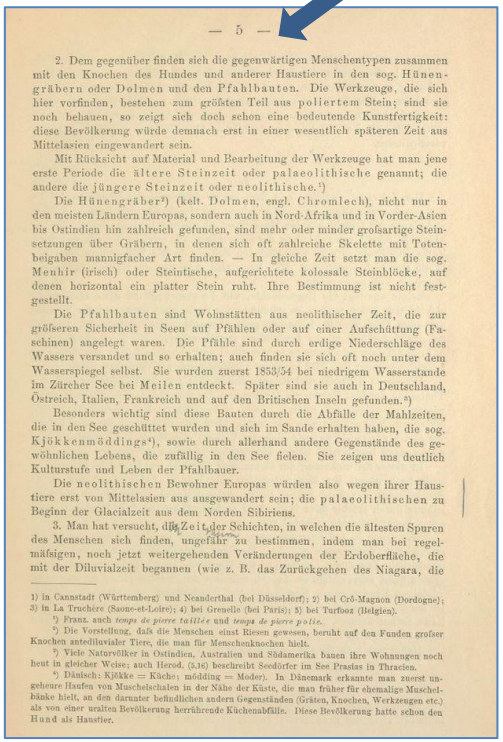
- ✓ Digitale Editionen

⇒ http://clarin.bbaw.de/de/



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).



— 5 —

2. Dem gegenüber finden sich die gegenwärtigen Menschentypen zusammen mit den Knochen des Hundes und anderer Haustiere in den sog. Hünengräbern oder Dolmen und den Pfahlbauten. Die Werkzeuge, die sich hier vorfinden, bestehen zum größten Teil aus poliertem Stein; sind sie noch behauen, so zeigt sich doch schon eine bedeutende Kunstfertigkeit: diese Bevölkerung würde demnach erst in einer wesentlich späteren Zeit aus Mittelasien eingewandert sein.

Mit Rücksicht auf Material und Bearbeitung der Werkzeuge hat man jene erste Periode die ältere Steinzeit oder palaeolithische genannt; die andere die jüngere Steinzeit oder neolithische.[1])

Die Hünengräber[2]) (kelt. Dolmen, engl. Chromlech), nicht nur in den meisten Ländern Europas, sondern auch in Nord-Afrika und in Vorder-Asien bis Ostindien hin zahlreich gefunden, sind mehr oder minder grofsartige Steinsetzungen über Gräbern, in denen sich oft zahlreiche Skelette mit Totenbeigaben mannigfacher 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 gröfseren Sicherheit in Seen auf Pfählen oder auf einer Aufschüttung (Fascbinen) angelegt waren. Die Pfähle sind durch erdige 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, Östreich, Italien, Frankreich und auf den Britischen Inseln gefunden.[3])

Besonders wichtig sind diese Bauten durch die Abfälle der Mahlzeiten, die in den See geschüttet wurden und sich im Sande erhalten haben, die sog. Kjökkenmöddings[4]), 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 regelmäfsigen, 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, dafs die Menschen einst Riesen gewesen, beruht auf den Funden grofser Knochen antediluvialer Tiere, die man für Menschenknochen hielt.

[4]) Viele Naturvölker in Ostindien, Australien und Südamerika bauen ihre Wohnungen noch heut in gleicher Weise; auch Herod. (5.16) beschreibt Seedörfer im See Prasias in Thracien.

[4]) Dänisch: Kjökke = Küche; möding = 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.

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.

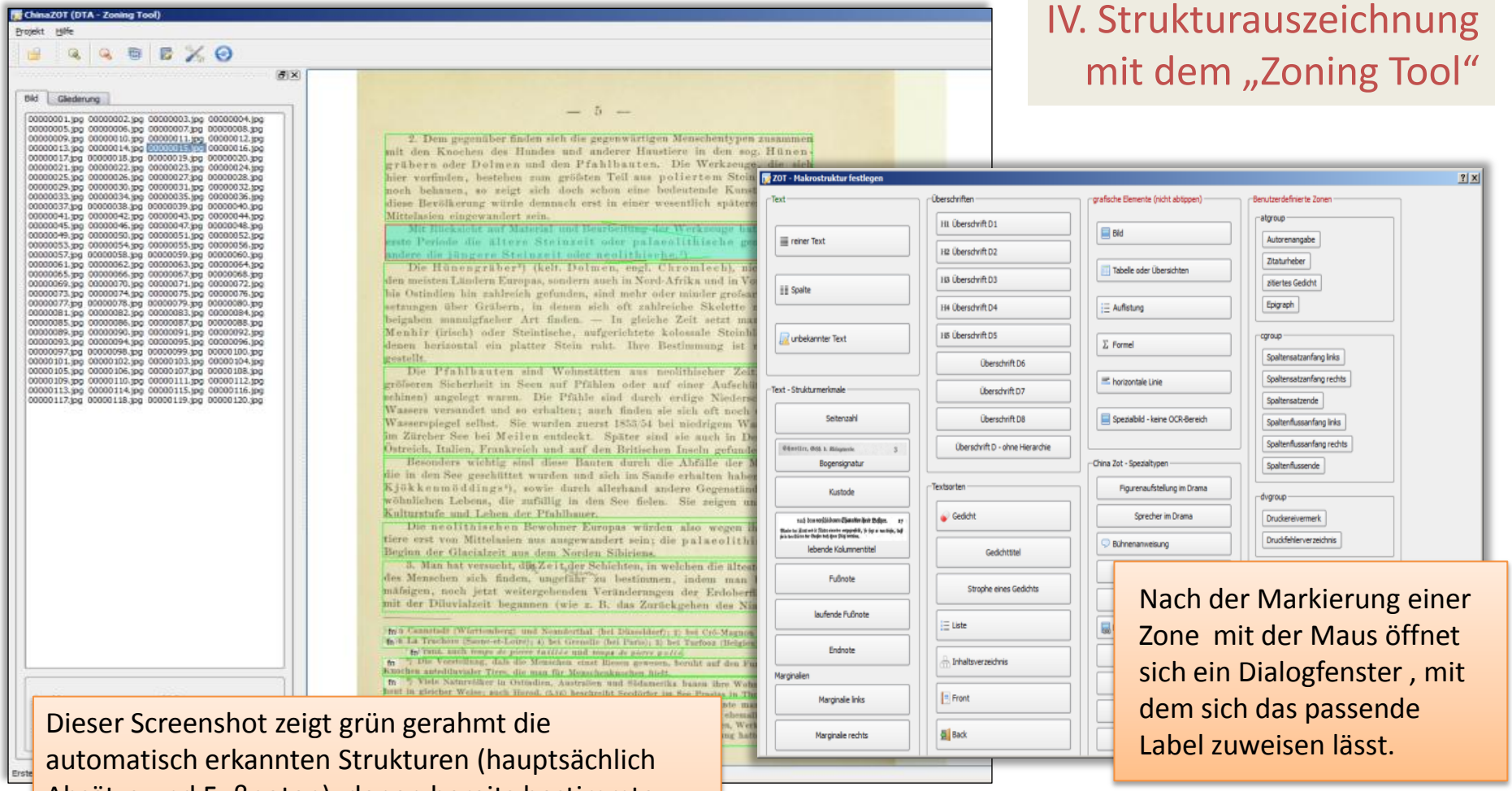
II. Datentransfer, Software-Bereitstellung

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.

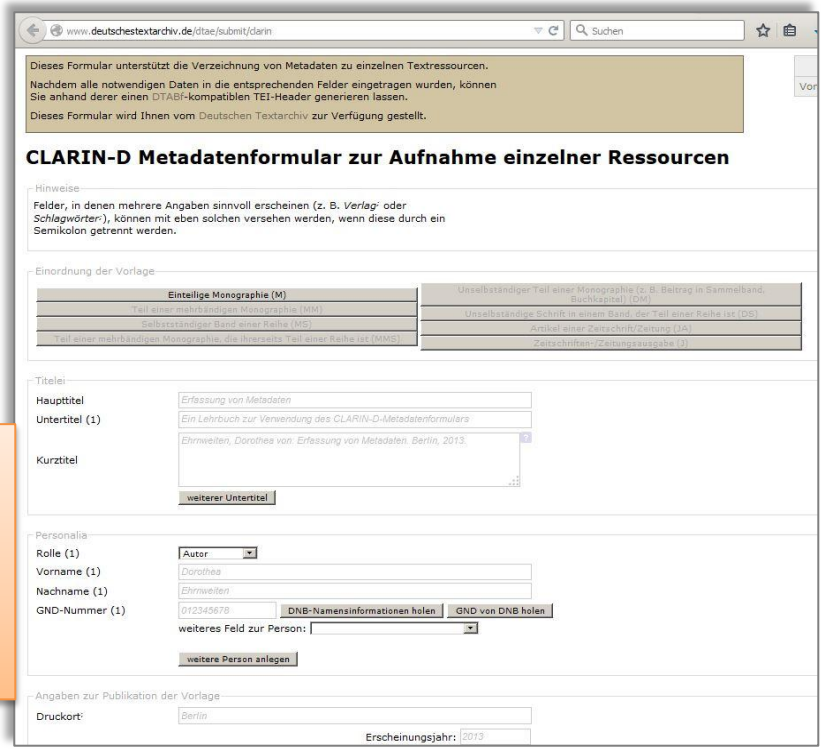
Nach der Markierung einer Zone mit der Maus öffnet sich ein Dialogfenster , mit dem sich das passende Label zuweisen lässt.

Die Strukturauszeichung nach TEI könnte auch direkt in den (ABBYY-)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 Metadatensätzen 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.



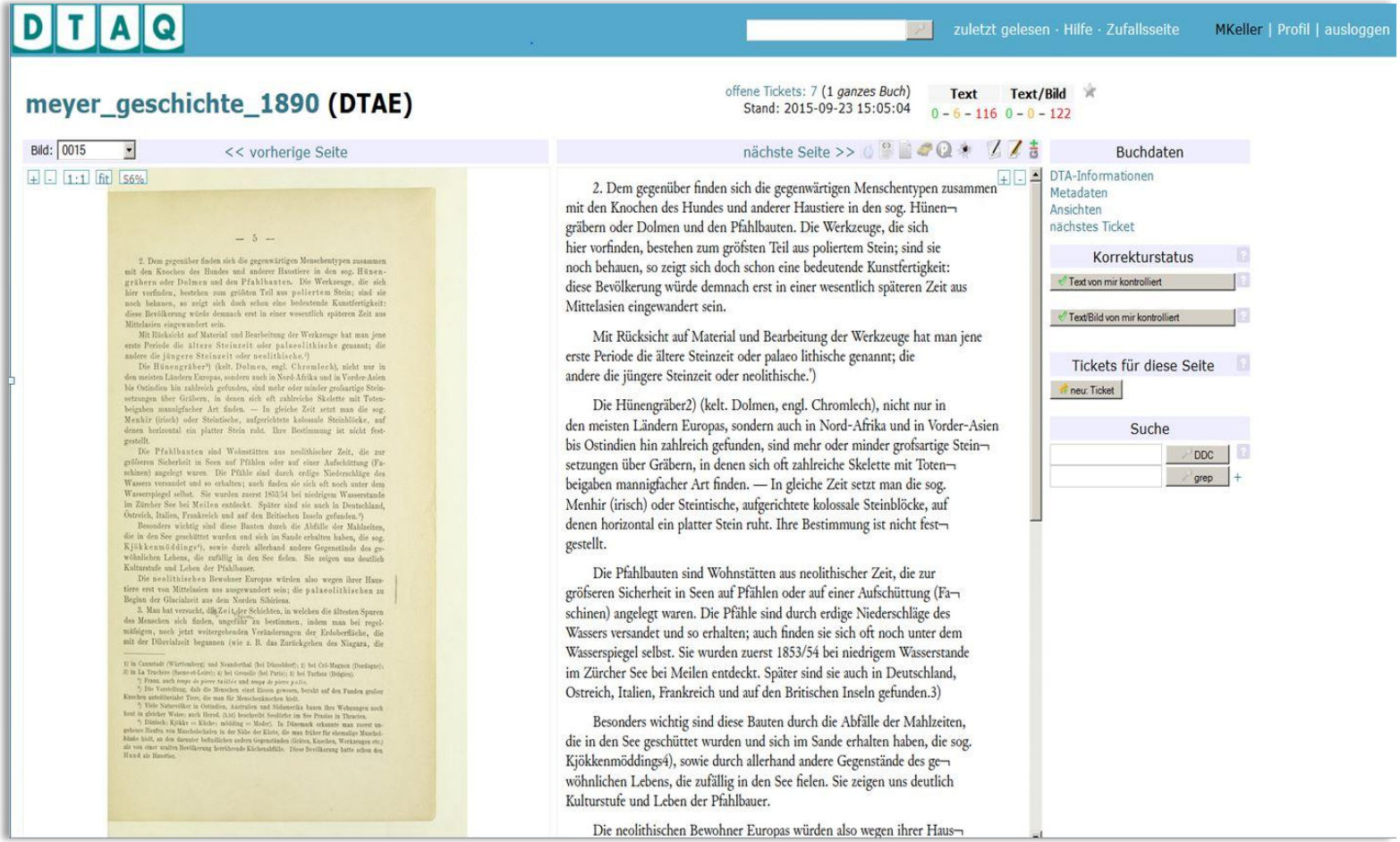
Die Component Metadata Infrastructure (CMDI) wurde von CLARIN entwickelt, um die Vielzahl der existierenden Metadaten-Formate für Sprachdaten und Tools kompatibel zu machen.

Einführende und weiterführende Informationen zu CMDI gibt es hier: http://www.clarin.eu/content/component-metadata

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)

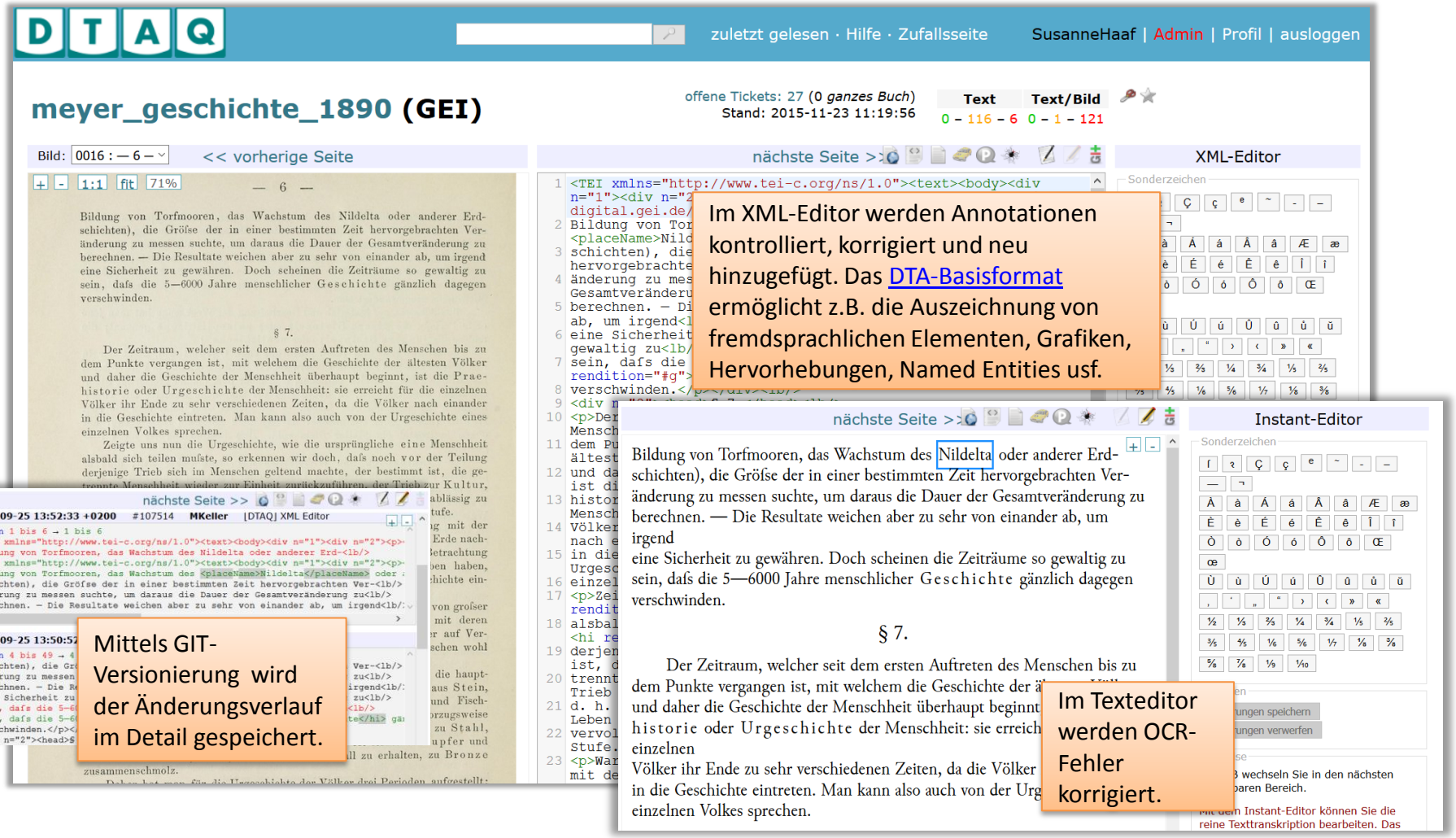
Die Abbildung unten zeigt den mit dem „ZoningTool“ des DTA „gezonten“ Text, wie er sich nach seiner Konvertierung in das TEI-basierte DTA-Basisformat in der Qualitätssicherungsumgebung des Deutschen Textarchivs (DTAQ) darstellt.



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.



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.

VII. Vergabe von PIDs

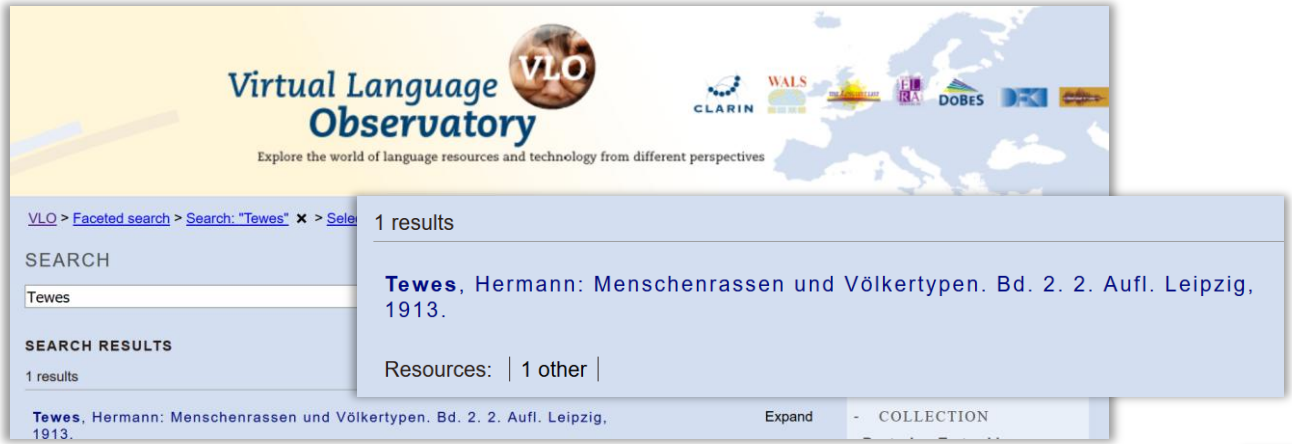
Die Daten werden nun am CLARIN-D Zentrum mit persistenten Identifikatoren (persistant identifiers, PIDs) versehen, die eine Langzeitarchivierung und die dauerhafte Referenzierbarkeit gewährleisten.

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

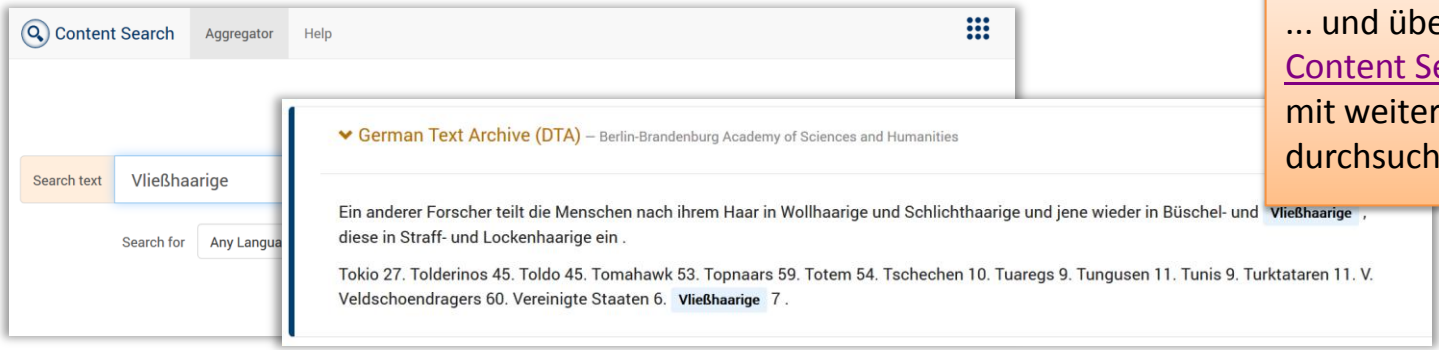
1 Treffer

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.

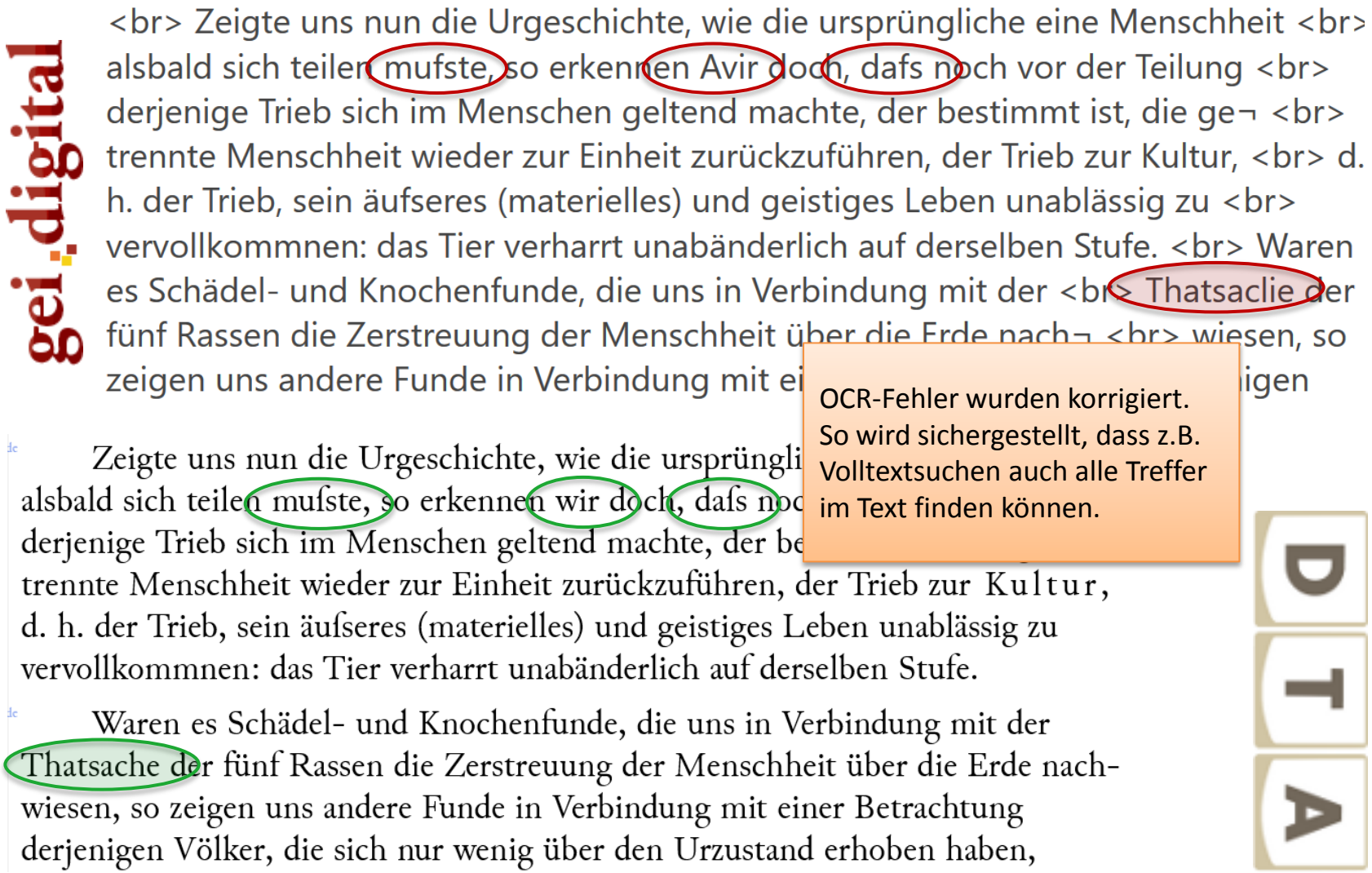


Sie können über CLARINs Virtual Language Observatory (VLO) gefunden …



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

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



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

Bsp. Edmund Meyer, Alte Geschichte, 1890

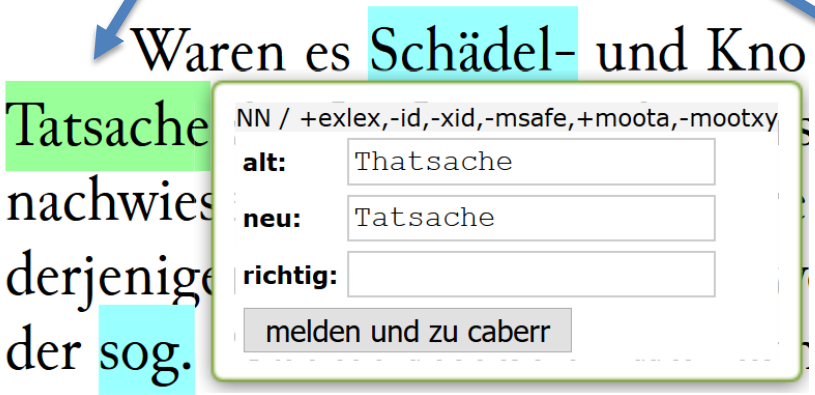
5. Resultate (3): Linguistische Abfragen mit DDC (Suchabfragesprache des DTA/DWDS)

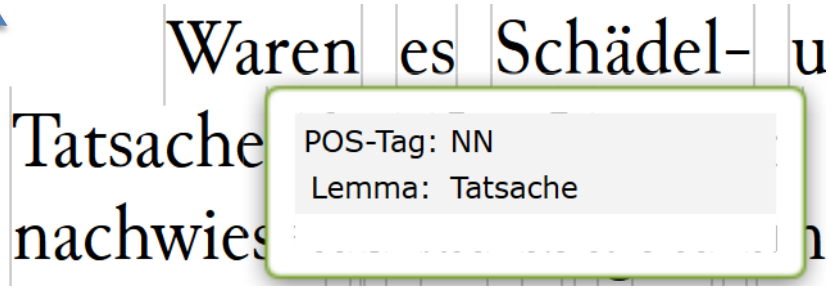
Linguist. Analyse

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.

Zeigte uns nun die Urgeschichte, wie die ursprüngl
alsbald sich teilen mufste, 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 äufseres (materielles) und geistiges
vervollkommnen: das Tier verharrt unabänderlich auf de

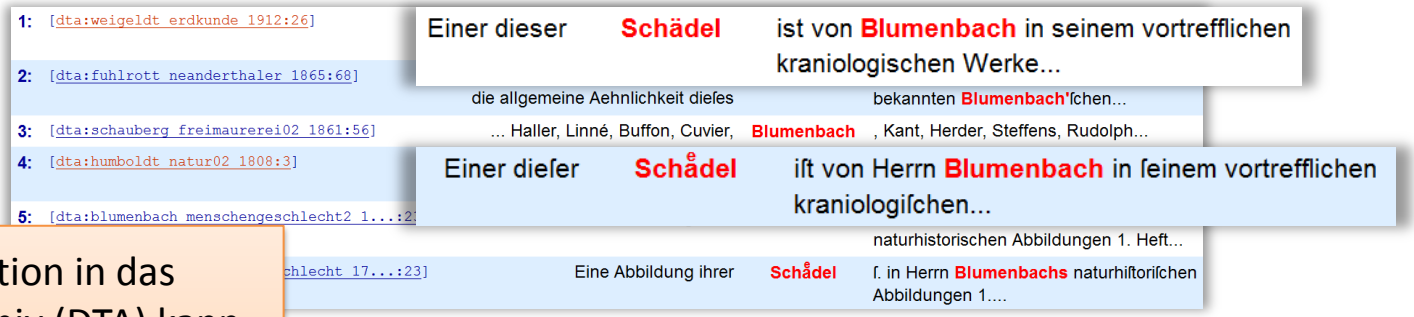
Waren es Schädel- und Knochenfunde, die uns in
Thatsache der fünf Rassen die Zerstreuung der Menschh
wiesen, so zeigen uns andere Funde in Verbindung mit e
derjenigen Völker, die sich nur wenig über den Urzustan
der sog. Naturvölker, wie sich die Völker, ehe sie in die Geschichte ein-
traten, nach bestimmten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



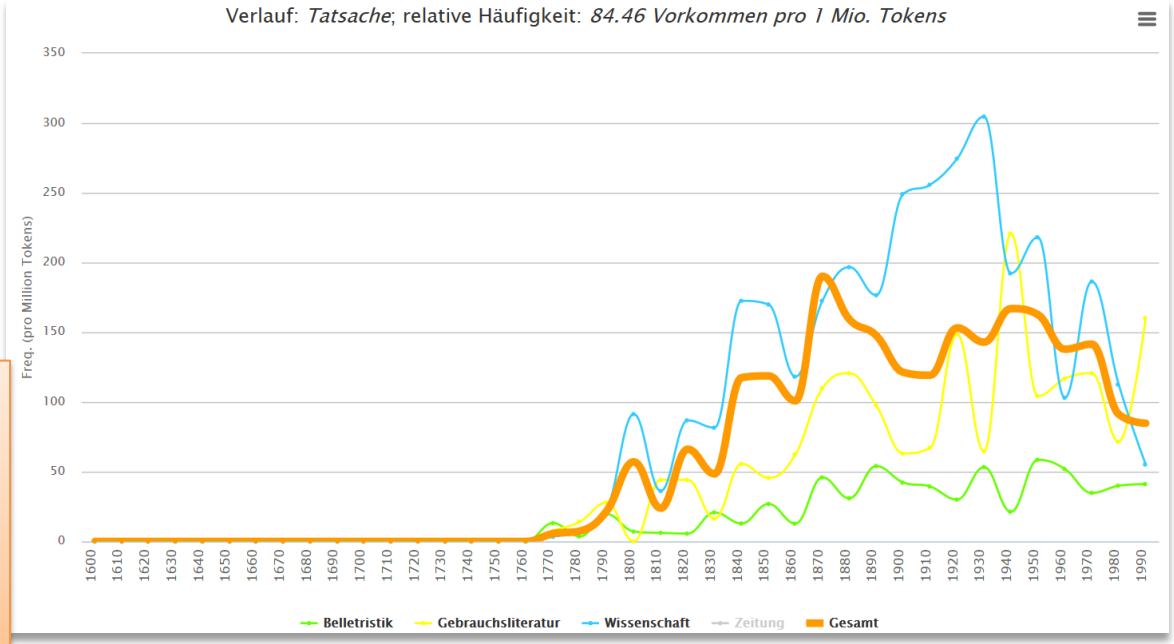


Bsp. Edmund Meyer, Alte Geschichte, 1890

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



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.



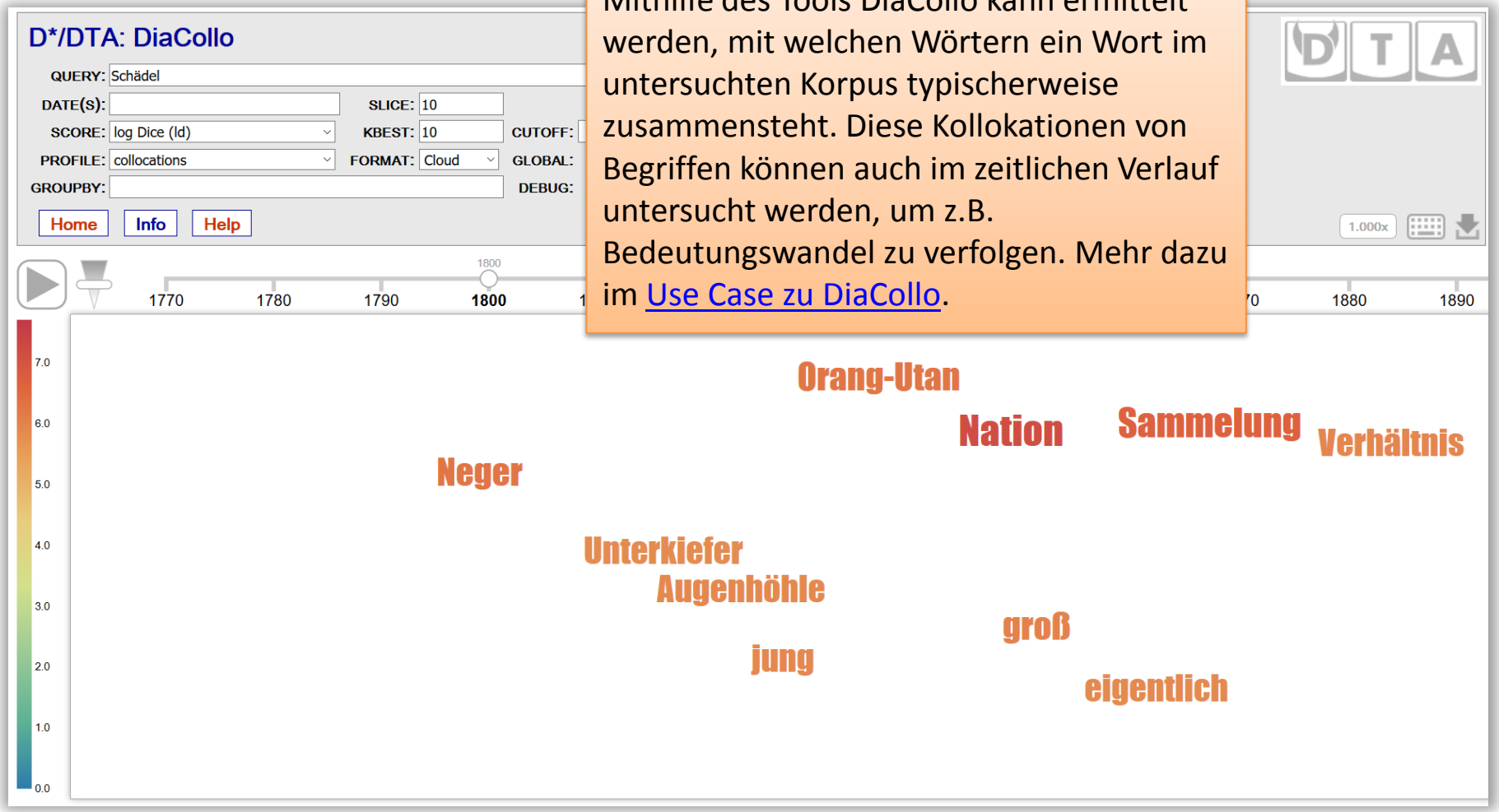
Im DTA lässt sich auch die Konjunktur eines Begriffes über die Zeit und in verschiedenen Subkorpora bestimmen.

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

CLARIN-D