Vernetzung von Daten im Deutschen Textarchiv

Susanne Haaf, Matthias Boenig, Christian Thomas, Alexander Geyken, Bryan Jurish, Frank Wiegand

Berlin-Brandenburgische Akademie der Wissenschaften/Deutsches Textarchiv

Einleitung

In den vergangenen acht Jahren wurde im Rahmen des Projekts Deutsches Textarchiv[1] ein historisches Korpus für die Entwicklung der neuhochdeutschen Sprache aufgebaut und interessierten Nutzern in hoher Qualität, unter freien Lizenzen und mit vielfältigen Recherche- und Nachnutzungsmöglichkeiten zur Verfügung gestellt. Die DTA-Plattform zieht mittlerweile Nutzer aus ganz unterschiedlichen Disziplinen und mit vielfältigen Fragestellungen an. Sie analysieren die DTA-Korpora in Hinblick auf sprachhistorische, begriffsgeschichtliche, medienhistorische oder wissenschaftsgeschichtliche Phänomene, oder sie nutzen die Plattform, um ihre eigenen Korpora oder Editionen zu publizieren und weiter zu bearbeiten. Für diese Nutzungsszenarien bedarf es sowohl zuverlässiger Forschungsdaten und Analysetools als auch konkreter Richtlinien und Workflows für die Datenaufbereitung und -integration. Wesentliche Faktoren dabei sind zum einen sehr hohe Anforderungen an die Qualität der angebotenen Daten sowie deren Interoperabilität. So wird es möglich, Daten aus ganz verschiedenen Projektkontexten zusammenzuführen und miteinander zu vernetzen, wie im Folgenden gezeigt wird.

Vernetzung von Daten über das DTA

Das DTA besteht aus einem Kernkorpus aus ~1600 Druckpublikationen (17.–19. Jh.) sowie mehreren Erweiterungskorpora (DTAE) auf Grundlage kuratierter Ressourcen aus externen Projekten und dem CLARIN-D-Netzwerk. Die Texterfassung für das Kernkorpus erfolgte mit sehr hoher Zuverlässigkeit im Double-Keying-Verfahren, die aufgenommenen Textquellen in DTAE wurden gemäß den DTA-Richtlinien und dem im DTA entwickelten DTA-Basisformat (DTABf)[2] aufbereitet und in das Korpus integriert. Diese konsequente Vereinheitlichung von Daten unterschiedlicher Herkunft stellt deren Interoperabilität sicher und erlaubt somit deren Vernetzung.

---

[1] Cf. die Projektwebseite www.deutschestextarchiv.de.

[2] Format und Dokumentation unter www.deutschestextarchiv.de/doku/basisformat.

Beispielsweise unternahm das genuin linguistische Projekt „Volltextdigitalisierung der Staats- und Gelehrte[n] Zeitung des Hamburgischen Unpartheyischen Correspondenten und ihrer Vorläufer (1712–1848)“[3] an der Universität Paderborn die Digitalisierung von 204 Zeitungsausgaben entsprechend dem Workflow und den Richtlinien des DTA; auf ähnliche Weise wurde die “Neue Rheinische Zeitung” (301 Ausgaben) durch die Marx-Engels-Gesamtausgabe an der BBAW digitalisiert. Das am IDS Mannheim produzierte “Mannheimer Korpus Historischer Zeitungen”[4] (652 Ausgaben) konnte aus dem TEI-Format I5[5] fast vollautomatisch in das DTABf überführt werden. Die Daten dieser drei zunächst unabhängigen Projekte werden derzeit im DTA zu einem Spezialkorpus historischer Zeitungen zusammengeführt.

Aus dem Bereich der historischen Wissenschaften wurde im Rahmen des Projekts „AEDit Frühe Neuzeit“[6] ein Korpus von 337 Funeralschriften aus der Universitätsbibliothek Breslau entsprechend den DTA-Richtlinien erfasst und in das DTA integriert. Ergänzt wird es durch 110 Funeralgedichte aus der Staatsbibliothek zu Berlin, die dort unter bibliothekarischen Gesichtspunkten im Rahmen einer OCR-Studie[7] digitalisiert und im Anschluss an die Qualitätsrichtlinien des DTA angepasst wurden.

Das DTA wurde auf diese Weise durch zwei Spezialkorpora um neue Textsorten erweitert. Dabei wurden Daten unterschiedlicher Herkunft miteinander und mit der bestehenden DTA-Infrastruktur (Korpora, Analyse-Tools, Recherchemöglichkeiten) verknüpft, was je nach Format, Aufbereitungsstand und Qualität der Quelldaten mit unterschiedlich hohem Aufwand verbunden war. Voraussetzung für diese Aufgabe waren eindeutige und gut dokumentierte Richtlinien für die Texterfassung und -annotation.

Nutzungsmöglichkeiten historischer Daten im DTA

Das DTA bietet vielfältige Möglichkeiten zur Recherche. Hierfür durchlaufen sämtliche Daten eine automatische linguistische Analyse (inkl. Tokenisierung, Lemmatisierung, morphologischer Analyse und orthographischer Normierung).[8] Mithilfe der linguistischen Suchanfragesprache DDC sind sodann komplexe Suchanfragen möglich (Abb. 1).

---

[3] Näheres zum Projekt unter: http://kw1.uni-paderborn.de/institute-einrichtungen/institut-fuer-germanistik-und-vergleichende-literaturwissenschaft/germanistik/personal/schuster/projekte/volltextdigitalisierung-der-staats-und-gelehrten-zeitung-des-hamburgischen-unpartheyischen-correspondenten-und-ihrer-vorlaeufer-1712-1851.

[4] Näheres zum Projekt unter: https://repos.ids-mannheim.de/mkhz-beschreibung.html

[5] Cf. Lüngen/Sperberg-McQueen 2012.

[6] Näheres zum Projekt unter: http://www.hab.de/de/home/wissenschaft/forschungsprofil-und-projekte/aedit-fruehe-neuzeit-archiv--editions--und-distributionsplattform-fuer-werke-der-fruehen-neuzeit.html.

[7] Cf. Federbusch/Polzin 2013.

[8] Zur ling. Analyse im DTA vgl. auch http://www.deutschestextarchiv.de/doku/software.

1: [anonym_relation_1609:15]	... einziehen/ die vornemsten Türcken aber auff die	**Galleren**	**schmieden**/ auch den Lateinischen Bischoff vbel tractiren ...
2: [albertinus_landtstoertzer01_1615:397]	... gehalten/ aber alle jhre beste vnnd gelehrteste	**Bücher**	**schmiden** sie auff jhrer Bibliothec an eysenen Ketten ...
3: [haarer_bawrenkrieg_1625:85]	... helffen/ denselben Jöckeln ließ er an ein	**Ketten**	**schmieden**/ vnd bey einem Fewer lebendig/ ...
4: [kentz_handwerksboden_1629:78]	... sagt obgemelter Autor insonderheit von den Zeug- vnd	**Waffen**	**Schmiden**/ daß selbige vor andern sind genennet ...
5: [gottfried_historia_1631:239]	... noch im Meerhafen gefangen genommen/ vnd in	**Eisen**	**schmiden** lassen/ vnd damit sie desto grössere ...
6: [gottfried_historia_1631:296]	.../ in Eyl auß Gold vnd Silber etliche	**Rüstung**	**schmiden**/ vnd die Statt allenthalben mit starcker ...
7: [wartmann_germania0203_1650:51]	... dem Königreich Neapels nicht vnterhalten/ vnd keine	**Practicken**	**schmiden**.
8: [wartmann_germania04_1652:291]	... vmb die jenigen Geistlichen Gewissen/ die am	**Frieden**	**schmieden** od' hindern solten/ ob sie nicht ...
9: [wartmann_germania05_1653:118]	.../ welche kein andere Arbeit vornehmen/ als	**Schlußreden**	**schmieden**/ vnd spintisieren/ sonderlich aber vber ...
10: [wartmann_germania05_1653:258]	... sondern man müste ohne Odem schöpffen an diesem	**Eysen**	**schmieden**.
11: [greflinger_krieg_1657:82]	... göldnen Frieden/ Ließ aber mitlerzeit viel grimme	**Waffen**	**schmieden**/ und unter Holckens Hand bey zehen ...
12: [glauber_opera01_1658:376]	... so weich machen/ daß mans auff einem	**Amboß**	**schmieden** kan.
13: [buchholtz_herkules01_1659:393]	... aus tausend Markomiren könne man nicht einen einzigen	**Herkules**	**schmieden**/ welches ich zu dem Ende andeute ...

Abb. 1: Abfrage: "$p=NN schmieden with $p=VVINF" ('Normales Nomen, gefolgt von einer Form des Worts "schmieden", dessen Wortart als infinite Verbform bestimmt wurde')

Des Weiteren können die DTA-Korpora im Zusammenhang und Vergleich mit den synchronen Korpora des DWDS[9] ausgewertet werden (Abb. 2).

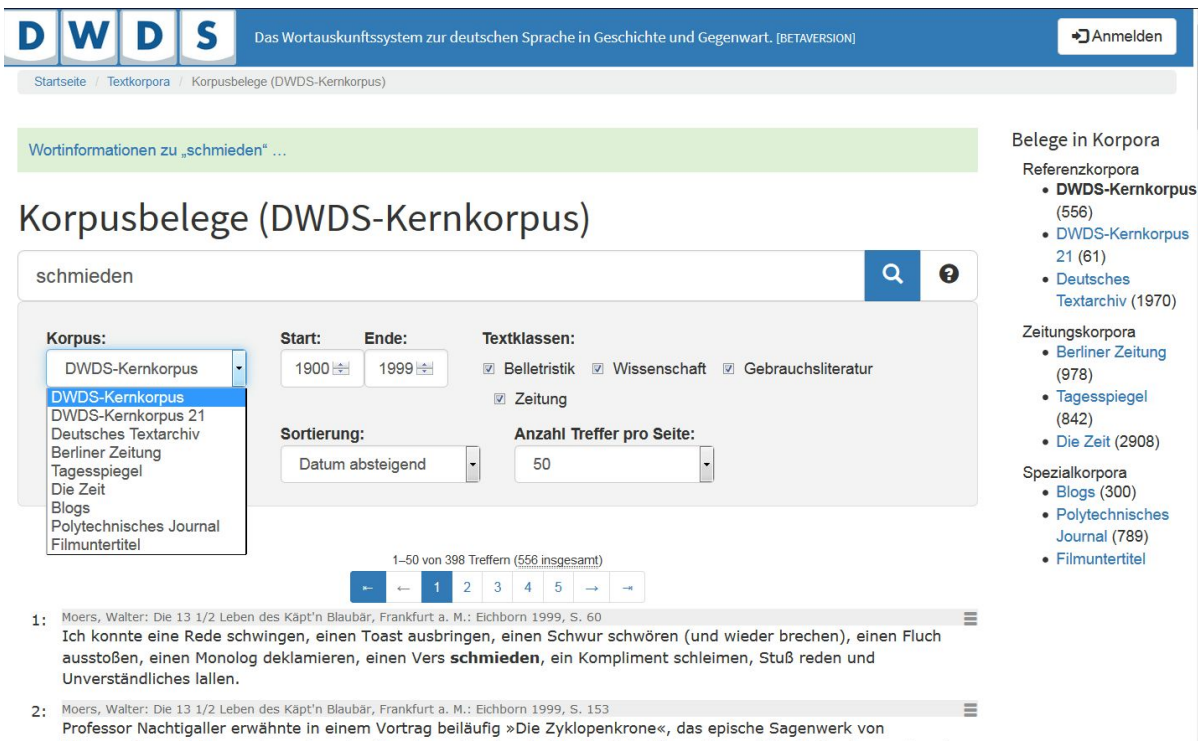


Abb. 2: Suche nach "schmieden" auf http://zwei.dwds.de.

[9] Digitales Wörterbuch der Deutschen Sprache, Website (beta): http://zwei.dwds.de.

Die Analyseergebnisse lassen sich mittels Wortverlaufskurven visualisieren, welche die chronologische Verteilung der Treffer nach Gattungsbereichen darstellen (Abb. 3).

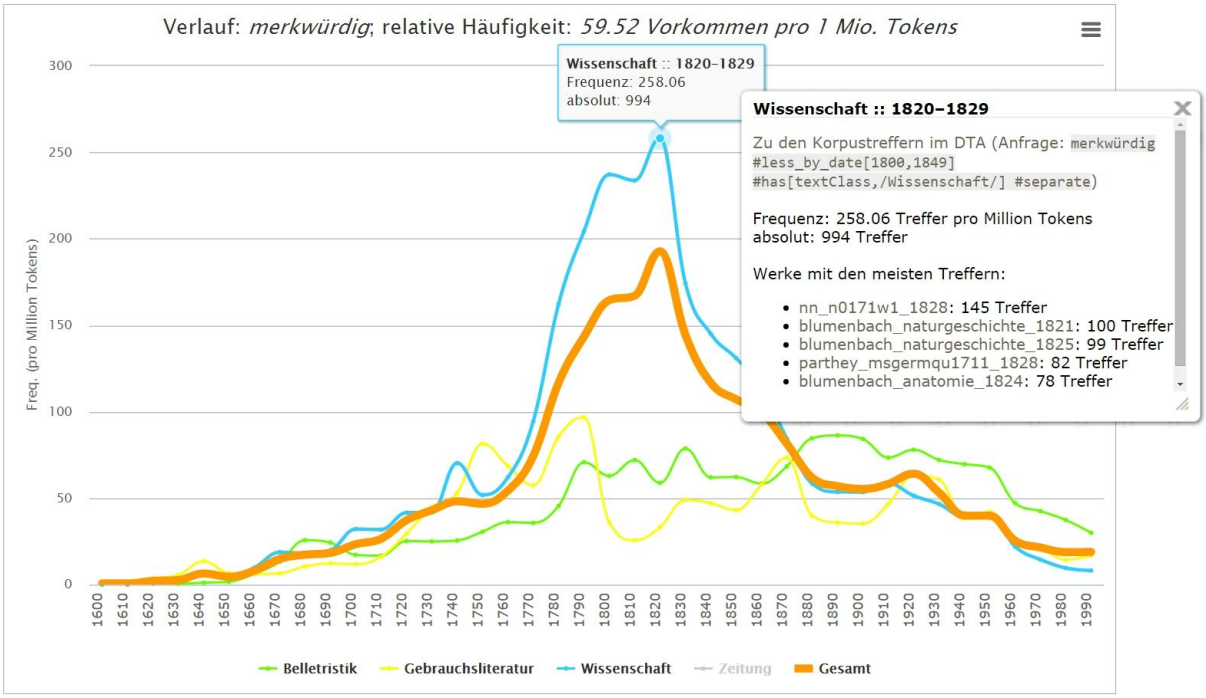
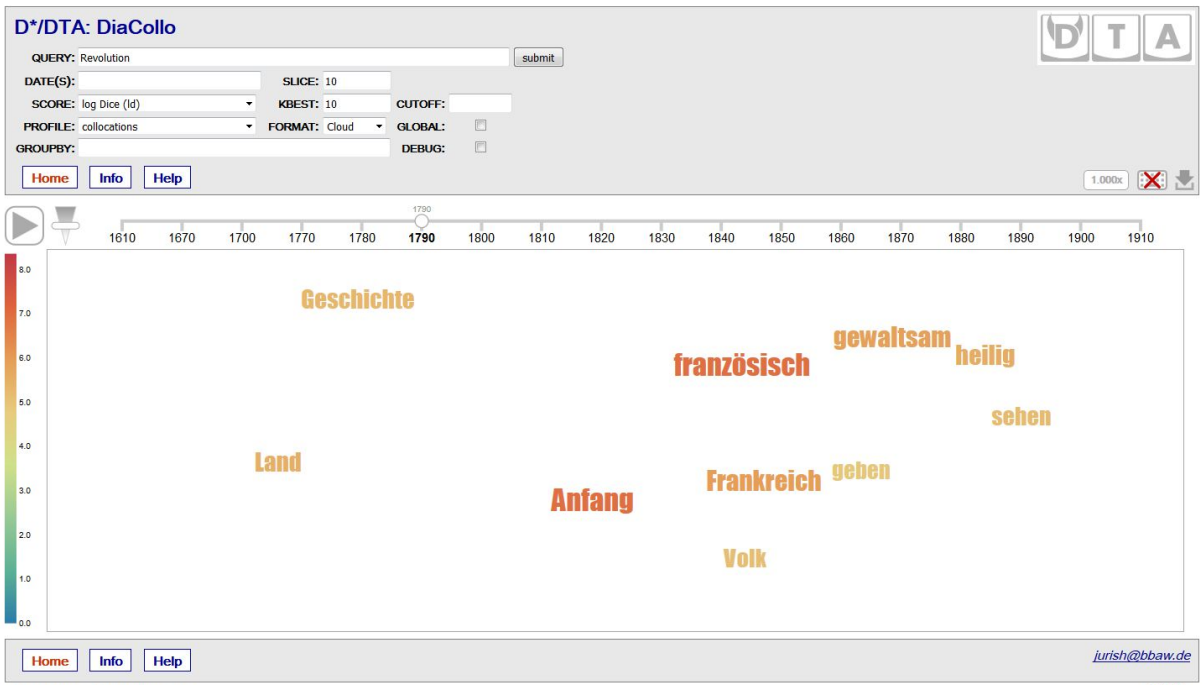


Abb. 3: Wortverlaufskurve DTA+DWDS (1600–2000),
http://www.deutschestextarchiv.de/search/plot/?query=merkw%C3%BCrdig&mode=extended;norm=date%2Bclass&smooth=spline&single=0&grand=1&slice=10&prune=0.01&window=2&wbase=3&logavg=0&logscale=0&xrange=1600%3A2000&totals=0

Das Tool DiaCollo erlaubt schließlich die Betrachtung von Termini im Zusammenhang mit den für sie typischen Kollokationen im zeitlichen Verlauf. Hierbei sind auch vergleichende Analysen von Termini möglich (Abb. 4 & 5).



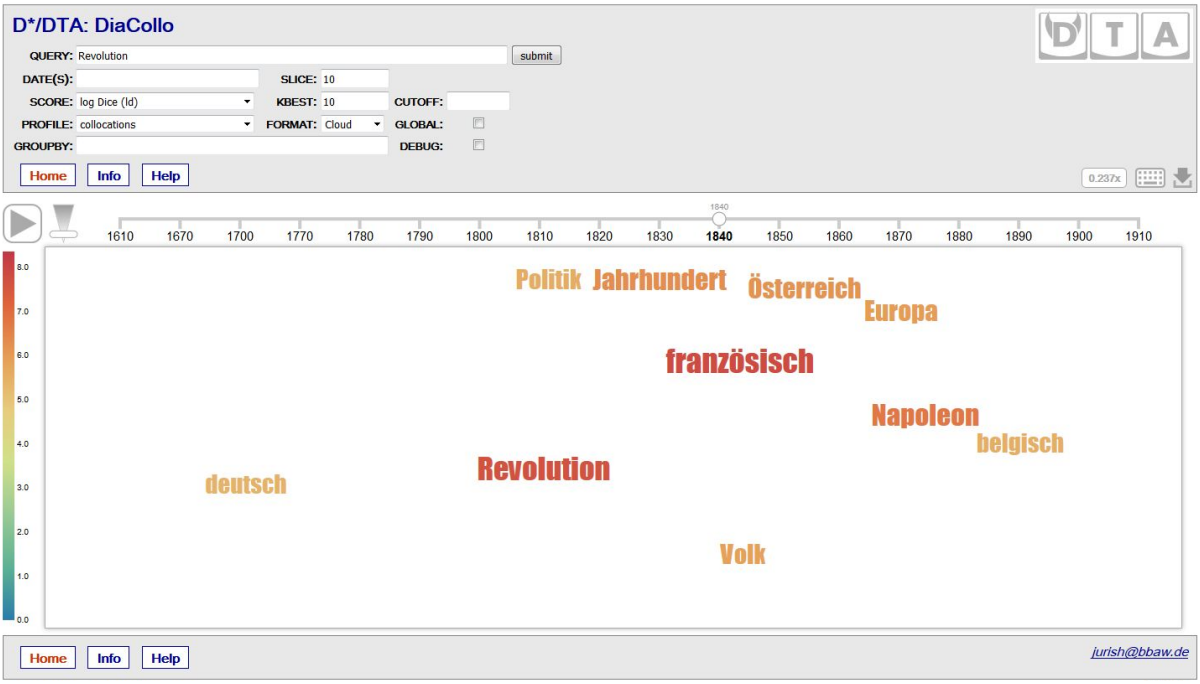


Abb. 4 & 5: DiaCollo-Ansicht "Revolution" mit seinen häufigsten Kollokationen, Zeitschnitte 1790 und 1840

Integration von Daten in die CLARIN-Infrastruktur

Über das Zentrum Sprache der BBAW ist das DTA Partner im Verbundprojekt CLARIN-D.[10] Diese Anbindung an die CLARIN-Infrastruktur eröffnet Nutzern ausgehend vom DTA weitere Recherche- und Analysemöglichkeiten.

Hierfür stellt das DTA sämtliche Texte auch in den CLARIN-kompatiblen Formaten CMDI (Metadaten) und TCF (Textdaten inkl. linguistischer Analyse) bereit. Die Metadaten können beispielsweise durch das CLARIN-übergreifende Virtual Language Observatory[11] geharvestet werden. Die Textdaten sind über CLARINs Federated Content Search[12] zugänglich und somit im Vergleich mit den Korpora anderer CLARIN-Zentren durchsuchbar. Die TCF-Daten können in der CLARIN-Umgebung WebLicht[13] weiter analysiert werden. Ein TEI-to-TCF-Converter ermöglicht die Konvertierung von TEI-Daten in das TCF-Format und vice versa.

[10] Cf. http://clarin.bbaw.de.
[11] Cf. https://vlo.clarin.eu.
[12] Cf. http://weblicht.sfs.uni-tuebingen.de/Aggregator/.
[13] Cf. http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page.

Bibliographie

Federbusch/Polzin 2013: Maria Federbusch und Christian Polzin: Volltext via OCR – Möglichkeiten und Grenzen. Testszenarien zu den Funeralschriften der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz. Mit einem Erfahrungsbericht von Thomas Stäcker aus dem Projekt „Helmstedter Drucke Online“ der Herzog August Bibliothek Wolfenbüttel. Berlin 2013 (=Beiträge aus der Staatsbibliothek zu Berlin – Preußischer Kulturbesitz 43). http://staatsbibliothek-berlin.de/fileadmin/user_upload/zentrale_Seiten/historische_drucke/pdf/SBB_OCR_STUDIE_WEBVERSION_Final.pdf

Lüngen/Sperberg-McQueen 2012: Harald Lüngen und C. M. Sperberg-McQueen: A TEI P5 Document Grammar for the IDS Text Model. In: Journal of the Text Encoding Initiative 3 (2012): TEI and Linguistics. https://jtei.revues.org/508

Für Publikationen zum DTA, zu den Richtlinien und Tools sowie zu angrenzenden Projekten cf. www.deutschestextarchiv.de/doku/publikationen.