Das Deutsche Textarchiv: Vernetzung und Nachnutzung

Veranstalter: Berlin-Brandenburgische Akademie der Wissenschaften Berlin
Datum, Ort: 11.10.2010-12.10.2010, Berlin
Bericht von: Marius Hug, Institut für Kulturwissenschaft, Humboldt-Universität zu Berlin

Der Workshop „Das Deutsche Textarchiv: Vernetzung und Nachnutzung" fand am 11. und 12. Oktober 2010 an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) statt. Ausgehend vom „Deutschen Textarchiv" (DTA; <http://www.deutschestextarchiv.de>) wurden Möglichkeiten des Austauschs von Objekt- und Metadaten sowie Bemühungen um verbindliche Standards vorgestellt. Perspektiven für die Bereitstellung von Schnittstellen zur Anbindung an Repositorien und die Vernetzung mit anderen Projekten wurden diskutiert. Zudem ging es um Fragen der linguistischen Textaufbereitung innerhalb historischer Korpora.

Nach der Begrüßung der circa 60 Teilnehmer/innen durch MANFRED BIERWISCH (Berlin) stellte MATTHIAS SCHULZ (Berlin) die grundsätzliche Zielrichtung des Projekts sowie dessen Arbeitsweise vor. Anschließend beschrieb OLIVER DUNTZE (Berlin) das Auswahlverfahren für die Titelliste des DTA – bisher wurden 650 Titel des Zeitraums 1780 bis 1900 digitalisiert – sowie die Herausforderungen, welche die angestrebte ausgewogene Gewichtung der Titelauswahl „Zwischen Kanon und Nutzeranforderung" bedeutete. Anschließend gab SUSANNE HAAF (Berlin) einen Ausblick auf die für Phase zwei des DTA vorgesehenen, weiteren circa 650 Titel aus dem Zeitraum von 1650 bis 1780. Hinsichtlich der zeitlichen, räumlichen und textsortenspezifischen Gewichtung, weise die nach sprachwissenschaftlich-lexikographischen Gesichtspunkten erstellte Textauswahl die notwendige Diversität auf, um die Grundlage für sprachhistorische, wortschatzbasierte Untersuchungen zu bilden. Auch für Fragestellungen anderer Disziplinen bilde das DTA-Korpus einen guten Ausgangspunkt.

THOMAS GLONING (Gießen) rückte die Perspektive der (Nach-)Nutzung des DTA-Korpus in den Mittelpunkt seiner Überlegungen. Er stellte „Erweiterungs- und Vernetzungsvorschläge zum DTA am Beispiel von Diskursen um 1900" vor. Anhand von Kombinationen des DTA-Korpus mit anderen Korpora, beispielsweise dem des „Digitalen Wörterbuchs der deutschen Sprache" (<http://www.dwds.de>) zeigte Gloning, wie sich mittels vernetzter Ressourcen zeitgenössische Wortfelder zusammenstellen sowie Aspekte von Diskursen und Frames rekonstruieren lassen. Die (unerlässliche) Kombination mehrerer Quellen erlaube es schließlich, die Chronologie der wortgeschichtlichen Entwicklungen zu charakterisieren und zu dokumentieren. Ein einzelnes Korpus könne nicht alle Themenbereiche abdecken, die kulturell und kommunikationsgeschichtlich von Bedeutung (gewesen) seien. Daher sei es wichtig, dass Projekte wie das DTA eine enge Vernetzung mit anderen Vorhaben (zum Beispiel über Kooperationen im inhaltlichen und technischen Bereich, über OAI-Schnittstellen oder parallele Suchroutinen) herstellten.

SEBASTIAN MEYER (Dresden) sprach in seinem Vortrag über die von der DFG im Rahmen ihrer Praxisregeln festgelegten Formatstandards für die Digitalisierung. Meyer stellte den Metadata Encoding & Transmission Standard (METS) und das Metadata Object Description Schema (MODS) vor. Für den Datenaustausch zwischen Institutionen präferiere die DFG das Protocol for Metadata Harversting der Open Archives Initiative (OAI-PMH). Auf Standards für eine textbasierte Digitalisierung – hier wäre v.a. das XML-Format der Text Encoding Initiative (TEI) zu nennen – habe man sich bisher nicht einigen können. Der DFG-Viewer solle als Referenz-Tool für DFG-Digitalisierungsprojekte fungieren. Wie in der Diskussion des Vortrags deutlich wurde, besitzt der DFG-Viewer allerdings einen zu eng begrenzten Funktionsrahmen, um diese Aufgabe zu erfüllen. So fehlten bislang beispielsweise die Möglichkeit zum Speichern von benutzerdefinierten Annotationen oder Bookmarks ebenso wie eine Volltextsuche.

MATHIAS SCHINDLER (Göttingen) stellte im Anschluss mit dem Datenformat „PND-

© H-Net, Clio-online, and the author, all rights reserved.

BEACON“[1] einen Weg zur Verknüpfung von Webseiten, die Personendaten enthalten, vor. Die PND-BEACON-Datei einer Webseite enthält die PND-Nummern, mittels derer Personen innerhalb dieser Seite identifiziert wurden, und einen diesen zugeordneten numerischen Wert ($\geq$1), der die Anzahl der mit der jeweiligen Person verknüpften Einträge auf der Webseite angibt. Das Potential dieser sehr einfachen Struktur bestehe in der Möglichkeit, heterogene Datenbestände zu vernetzen. PND-BEACON-Dateien können sowohl von den Anbietern der jeweiligen Webpräsenz selbst als auch von externen Personen erstellt werden. Schindler wies zudem auf „SeeAlso“[2] hin, ein Verfahren zur kontextbezogenen Einbindung von Links in Webseiten, das sich mit dem BEACON-Format kombinieren ließe.

Im Anschluss stellte BRYAN JURISH (Berlin) einige der Werkzeuge vor, die im DTA zur linguistischen Annotation der XML/TEI-Volltexte dienen. Ein zentrales Anliegen, sowohl zur Optimierung der Suchfunktion innerhalb des DTA als auch für korpuslinguistische Analysen über dessen Textbestand, stelle die fehlerfreie Abbildung verschiedener graphematischer Varianten auf das ihnen gemeinsame Lemma dar. Historische Texte würden durch das Fehlen konsistenter orthographischer Konventionen eine besondere Herausforderung für gängige Sprachverarbeitungstechnologien darstellen – insbesondere, sofern sie auf Bezugnahme zu statischen Lexika basieren. Für das DTA hat Jurish computerlinguistische Methoden kombiniert, um zunächst die Suchanfrage mittels phonologischer Ähnlichkeitsmaße auf eine Menge möglicher Wortformen des gesuchten Lemmas zu expandieren und anschließend eine kontextabhängige Disambiguierung der Treffer zu erreichen.

CHRISTOPH RINGLSTETTER (München) präsentierte Erfahrungen aus seiner Arbeit für das EU-Projekt IMPACT (<http://www.impact-project.eu>), in dem unter anderem Ergebnisse automatisierter Texterkennung per OCR (Optical Character Recognition) mittels linguistischer Hilfsmittel optimiert werden sollen. Ringlstetter sprach sich für den Einsatz von evidenz-basierten Lexika als Ergänzung zu so genannten ‚hypothetischen Lexika‘ aus. Letztere enthielten Heuristiken zur Abbildung historischer Varianten auf heute gültige Wortformen. Am CIS München (<http://www.cis.uni-muenchen.de>) würden auf der Grundlage von umfangreichen Korpusdaten Evidenz-basierte Lexika erarbeitet, auf die die Software während des OCR-Vorgangs zugreifen könne. Das von Ringlstetter vorgestellte Test- bzw. Referenzkorpus enthält Texte von circa 1499 bis 1949 und könnte, auch mit Blick auf den bis 2013 vom DTA zu erwartenden Bestand älterer deutscher Texte, durch eine Kombination mehrerer Korpora entscheidend ergänzt werden.

Den Abschluss des ersten Tages markierte ein Abendvortrag EBERHARD KNOBLOCHs (Berlin), der einige zentrale Werke der Mathematik der Aufklärung im DTA aus wissenschaftshistorischer Sicht vorstellte. Knobloch lobte die Bemühungen des DTA, deutschsprachige Texte, die für die Entwicklung wissenschaftlicher Disziplinen einflussreich wurden, in Erst- bzw. frühen Ausgaben bereitzustellen. Diese seien oft nur noch in wenigen, schwer zugänglichen Exemplaren vorhanden.

ALEXANDER GEYKEN (Berlin) stellte zum Auftakt des zweiten Veranstaltungstages ein noch in Arbeit befindliches Paket von Applikationen vor, mit dem die Nutzer/innen des DTA in Zukunft auch kumulativ bzw. kooperativ mit dem Korpus arbeiten können. Bereits jetzt sei es möglich, Bildausschnitte aus den Digitalisaten (nach Anmeldung mit einem persönlichen Account) in einem eigenen Bereich zu speichern, zu verschlagworten und so in Beziehung zu anderen Fundstellen inner- oder außerhalb des DTA zu setzen. Diese Möglichkeiten sollen in den kommenden Monaten auch für Textpassagen realisiert werden. Literaturmanagementprogramme würden durch das Angebot von COinS (ContextObjects in Spans, <http://www.ocoins.info/>) unterstützt. Mit dieser Methode zur Einbindung von bibliographischen Metadaten in HTML-Seiten können die Metadaten des DTA au-

[1] Vgl. Wikipedia:PND/BEACON, <http://de.wikipedia.org/wiki/Wikipedia:PND/BEACON> (12.01.2011).

[2] Vgl. SeeAlso – Verbund-Wiki GBV, http://www.gbv.de/wikis/cls/SeeAlso, (12.01.2011).

© H-Net, Clio-online, and the author, all rights reserved.

tomatisch von weit verbreiteten Anwendungen wie Citavi oder dem Firefox-Add-On Zotero gespeichert werden. Generell werde für das userseitige (Meta-)Datenmanagement Kompatibilität mit verbreiteten externen Services angestrebt, anstatt einzelne Funktionen bereits gut funktionierender Dienste im DTA nachzubilden.

MICHAEL SOLF und CHRISTIAN THOMAS (beide Berlin) zeigten Strategien zur Ermittlung möglicher Kandidaten zur Erweiterung des DTA-Korpus um einzelne, lexikographisch ergiebige Werke. Allerdings sei mit Blick auf die aufwändige Aufbereitung der Volltexte für das Korpus Vollständigkeit schlichtweg nicht zu erreichen, weshalb sich jede ‚Nachnominierung' umso besser rechtfertigen lassen müsse. Die Ergänzungen müssten mit Blick auf laufende Digitalisierungsvorhaben und Möglichkeiten der Kooperation mit anderen Projekten ausgewählt werden. Neben der Nachnutzung von Bilddigitalisaten aus anderen Kontexten sei man im DTA besonders interessiert am Import von Volltexten. Eine erste Kooperation in diesem Bereich werde das DTA mit dem Editionsprojekt Sandrart.net eingehen.[3] Diese kommentierte Online-Edition biete eine seiten- und zeilengetreue Transkription der ‚Teutschen Akademie' Joachim von Sandrarts in XML/TEI-Lite, das in das XML/TEI-P5-Schema des DTA konvertiert werden solle. Es bleibe zu hoffen, dass die inzwischen weit verbreitete Anwendung der TEI-Richtlinien in Zukunft weitere Kooperationen auf Ebene des Volltextes ermöglichen werde.

Michael Solf beschrieb anschließend Erweiterungsmöglichkeiten des DTA auf der Grundlage des Quellenverzeichnisses des Deutschen Wörterbuchs. Diese etwa 4.000 Titel umfassende Liste von Werken, die für die Neubearbeitung des Deutschen Wörterbuchs von Jacob und Wilhelm Grimm (DWB) mehr oder weniger ausführlich exzerpiert wurden, wurde bereits vor Beginn der Projektlaufzeit des DTA für die Zusammenstellung des Korpus herangezogen. Solf illustrierte am Beispiel der Reiseliteratur, dass ein erneuter, detaillierterer Abgleich der DWB- und der DTA-Titelliste helfen könne, einige ‚blinde Flecke' bei der Fokussierung *beider* Projekte zu identifizieren. Die kultur- und sprachhistorische bzw. lexikographische Relevanz der auf diesem Wege ermittelten als Ergänzungsvorschläge für das DTA sei – zumindest für diejenigen Werke, die sich als besonders ergiebige Belegquellen für das DWB erwiesen hätten – durch ihre Verwendung in diesem bedeutenden Wörterbuch gesichert. Um künftig einen automatischen oder zumindest halbautomatischen Vergleich bibliographischer Verzeichnisse mit den DWB-Quellen zu ermöglichen, sei die Verwendung möglichst offener Datenformate und eindeutiger Referenzen – etwa normierter Ansetzungsformen wie den oben bereits erwähnten PND-Identifiern für Personen – unerlässlich.

CHRISTIAN KASSUNG (Berlin) zeigte mit seinem Vortrag Möglichkeiten der Kooperation zwischen vornehmlich textbasiert arbeitenden Digitalisierungsprojekten auf. Mit seinem exemplarischen Zugriff auf den Bestand des DTA wie auch des digitalisierten Polytechnischen Journals (<http://www.polytechnischesjournal.de>) konnte er vorführen, dass sich durch eine Zusammenarbeit beider Projekte sowohl auf inhaltlicher wie auch technischer Ebene Synergien ergeben würden. Während das Korpus des DTA einen disziplinübergreifenden Kernbestand deutschsprachiger Texte vor 1900 beinhaltet, ist Dingler's Polytechnisches Journal das zentrale deutschsprachige Organ für die technischen Wissenschaften des 19. Jahrhunderts. Bei entsprechend koordinierten Zugriffsmöglichkeiten könne das DTA den kulturellen Entstehungsherd verschiedener technischer, im „Dingler" beschriebener Innovationen liefern. Andererseits seien beide Projekte mit ähnlichen technischen Herausforderungen konfrontiert, denen man zum beiderseitigen Nutzen gemeinsam begegnen könne. Hier seien beispielsweise die Verschlagwortung bestimmter Texteinheiten sowie Algorithmen zur Lemmabildung zu nennen. Schließlich lenkte Kassung die Aufmerksamkeit auf die Bedeutung der usability. So sei eine projektspezifische Präsentationsschicht der entscheidende Garant dafür, mit dem Digitalisat auch wissenschaftlich arbeiten zu können.

THOMAS STÄCKER (Wolfenbüttel) prä-

[3] Vgl. Sandrart.net: <http://www.sandrart.net> (Projektwebsite) und <http://ta.sandrart.net/> (Edition), (12.01.2011).

© H-Net, Clio-online, and the author, all rights reserved.

sentierte aus Sicht der Herzog August Bibliothek Wolfenbüttel (HAB), wie sich durch die massenhafte Digitalisierung von Bibliotheksbeständen sowie allgemein durch den Einsatz digitaler Medien und Informationen bei der Erarbeitung und Veröffentlichung von Forschungsergebnissen auch das Selbstverständnis der „Bibliothek als Partner der Forschung" verändern müsse. Die Langzeitarchivierung von Forschungsergebnissen, aber auch von Primär- und Rohdaten falle in die Zuständigkeit der Bibliotheken. Stäcker stellte das Projekt „Helmstedter Drucke Online"[4] der HAB vor, in dem die Forschungsbibliothek erstmals nicht nur Imagedigitalisate und bibliographische Metadaten, sondern auch elektronische Volltexte einer großen Zahl der in Helmstedt gedruckten Schriften aus dem Bestand der HAB erstellen wird. Etwa 120.000 Seiten würden zunächst per OCR bearbeitet, nachkorrigiert und anschließend als XML-Volltexte angeboten. Die Verzeichnung erfolge im Verbundkatalog sowie im VD16 bzw. VD17.[5] Stäcker zeigte sich optimistisch für den erfolgreichen Einsatz von OCR auch für die älteren Drucke, da in Helmstedt eine überschaubare Zahl verschiedener Frakturtypen genutzt wurde. Um die Nachnutzung und jeweils projektspezifische Anreicherung der von verschiedenen Projekten erarbeiteten Volltexte und Strukturdaten zu erleichtern, schlug Stäcker ein TEI-Basisformat vor, das in groben Zügen bereits dokumentiert sei.[6]

STEFAN GRADMANN (Berlin) skizzierte entlang des Weges „Von Xanadu zu ORE und Linked Open Data" die Herausforderungen netzbasierten Arbeitens, die in Zukunft zu meistern seien. In der heutigen Wirklichkeit des World Wide Web (WWW) gelte es, sehr große und sehr heterogene Datenmengen zueinander in Beziehung zu setzen. Die Herausforderung bestehe darin, den verlinkten Ressourcen Stabilität in sich und in ihren Verweisen auf andere Ressourcen zu verleihen, um deren ‚Zitierfähigkeit' und damit wissenschaftliche Nutzbarkeit sicherzustellen. Ebenso wichtig für die wissenschaftliche Nutzbarkeit online publizierter Daten sei es, nachvollziehbar zu machen, welches Objekt zu einem bestimmten Zeitpunkt mit welchen weiteren Objekten verknüpft war, um dem flüchtigen ‚Netz der Daten' eine ‚Historie'zu geben. Gradmann plädierte für eine über RDF-Triple organisierte Ontologie und für semantisch basierte Heuristiken, wie sie beispielsweise der Europeana (<http://www.europeana.eu>) zugrunde lägen, um Inhalte des WWW semantisch zu erschließen. In dem so beschriebenen Raum vernetzter, frei zugänglicher (Forschungs-)Daten (‚Linked Open Data') habe das DTA unbedingt einen Platz.

Insgesamt wurde bestätigt, dass das ausgewogene Querschnittskorpus des DTA einen interessanten eigenen Untersuchungsraum für sprachgeschichtliche und computerlinguistische Analysen eröffnet. Das DTA offenbart sein Potential jedoch noch eindrucksvoller in der Verbindung mit externen Datenbeständen. Die auf dem Workshop vorgestellten und diskutierten Pläne des DTA-Teams, neue Möglichkeiten der Vernetzung über geeignete Schnittstellen zu schaffen, Nutzer/innen semantische Beziehungen aufzuzeigen und eine intensivere Interaktion mit dem Datenbestand zu erlauben, wurden ausdrücklich begrüßt.

Konferenzübersicht:

Begrüßung
Manfred Bierwisch (Berlin)

Matthias Schulz (Berlin): Das Deutsche Textarchiv – Einblick und Ausblick

Themenblock 1: DTA – Korpus
Oliver Duntze (Berlin): Zwischen Kanon und Nutzeranforderung – Die Textauswahl des Deutschen Textarchivs für den Zeitraum 1780–1900

Susanne Haaf (Berlin):„Von den Natürlichen und übernatürlichen Dingen". Die Texte des 17./18. Jahrhunderts im Deutschen Textarchiv

Thomas Gloning (Gießen): Wortgebrauch, Textcorpora und Lexikologie/Lexikographie.

[4] Helmstedter Drucke Online, http://www.hab.de/forschung/projekte/helmstedterdrucke.htm, (12.01.2011).

[5] VD 16 (bzw. VD 17) – Verzeichnis der im deutschen Sprachbereich erschienenen Drucke des 16. (bzw. 17.) Jahrhunderts, , <http://www.vd16.de> (bzw. <http://www.vd17.de>), (12.01.2011).

[6] Vgl. HAB-WDB Dokumentation, <http://www.hab.de/bibliothek/wdb/doku/index.htm>, (12.01.2011).

© H-Net, Clio-online, and the author, all rights reserved.

Erweiterungs- und Vernetzungsvorschläge zum DTA am Beispiel von Diskursen um 1900

Themenblock 2: Standardisierung
Sebastian Meyer (Dresden): Formatstandards für die Digitalisierung – Wo sind wir und wo wollen wir hin?

Mathias Schindler (Göttingen): Der digitale Fugenkitt. PND BEACON als Minimallösung für die automatische Vernetzung personenbezogener Internetdienste

Themenblock 3: Volltexterschließung
Bryan Jurish (Berlin): More Than Words: Orthographic Standardization in the Deutsches Textarchiv

Christoph Ringlstetter (München): Korpusbasierter Lexikonaufbau zur Erschließung älterer Texte

Abendvortrag
Eberhard Knobloch (Berlin): Wissenschaftsgeschichte und Mathematik der Aufklärung im Deutschen Textarchiv

Themenblock 4: Das DTA als aktives Archiv
Alexander Geyken (Berlin): Kumulatives Arbeiten mit dem Deutschen Textarchiv

Michael Solf (Berlin): Erweiterungsmöglichkeiten des Deutschen Textarchivs auf der Grundlage des Quellenverzeichnisses des Deutschen Wörterbuchs

Christian Thomas (Berlin): Deutsches Wörterbuch von Jacob Grimm und Wilhelm Grimm, Berlin-Brandenburgische Akademie der Wissenschaften

Themenblock 5: Vernetzung und Infrastrukturen
Christian Kassung (Berlin): Was ist ein Cylinder? Die Dampfmaschine im Deutschen Textarchiv und im Polytechnischen Journal

Thomas Stäcker (Wolfenbüttel): Die Bibliothek als Partner der Forschung: Szenarien der Archivierung, Vernetzung, Distribution und Publikation wissenschaftlicher elektronischer Dokumente und Quellen

Themenblock 6: Vernetzung und Infrastrukturen
Stefan Gradmann (Berlin): Von Xanadu zu ORE und Linked Open Data: Granular vernetzte Textobjekte und semantisch basierte Heuristiken

Tagungsbericht *Das Deutsche Textarchiv: Vernetzung und Nachnutzung*. 11.10.2010-12.10.2010, Berlin, in: H-Soz-u-Kult 19.04.2011.

© H-Net, Clio-online, and the author, all rights reserved.