Thomas Gloning, Frederike Neuber, Thomas Stäcker, Stefanie Seim, Christian Thomas

Zwischenbericht
CLARIN-D Kurationsprojekt 1 der F-AG 1

Kurationsprojekt zur Integration und Aufwertung historischer Textressourcen des 15.–19. Jahrhunderts in einer nachhaltigen CLARIN-Infrastruktur

---

Partner im Kurationsprojekt

- Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)
- Universität Gießen
- Herzog August Bibliothek Wolfenbüttel (HAB)
- Institut für Deutsche Sprache (IDS)[1]

Berichtszeitraum

01.09.2012–28.02.2013

Informationen zum Kurationsprojekt auf der CLARIN-D-Webseite

http://www.clarin-d.de/de/fachspezifische-arbeitsgruppen/f-ag-1-deutsche-philologie/kurationsprojekt-1.html

[1] Am IDS ist die Spiegelung der in die Infrastruktur integrierten Textdaten vorgesehen; diese Arbeiten werden im Bericht nicht eigens aufgeführt.

Arbeiten der Gießener Arbeitsgruppe im Berichtszeitraum

Die Gießener Arbeitsgruppe um Prof. Dr. Thomas Gloning setzt sich aus Hannah Glaum, Melanie Henß und Marc Kuse sowie – als Mitglieder der F-AG1 – Jurgita Baranauskaite und Stefanie Seim zusammen. Die Gruppe wird zusätzlich von dem an der JLU Gießen beschäftigten Christoph Wagenseil unterstützt. Im Berichtszeitraum wurden – wie von dem im Projektantrag aufgestellten Arbeitsplan vorgesehen – insbesondere Tätigkeiten zur Umsetzung von Arbeitspaket 1 (*Ressourcen identifizieren und verzeichnen*) und Arbeitspaket 4 (*Konvertierung und Integration textueller Ressourcen in die CLARIN-Infrastruktur*) bzw. dem zugehörigen Teilbereich 4.1 (*Erfassung der Texte in einem standardisierten, skalierbaren und TEI-konformen Basisformat*) durchgeführt.

Um diese Tätigkeiten innerhalb der Arbeitsgruppe zu koordinieren und gleichermaßen die Möglichkeit eines Austausches über die individuellen Fortschritte in der Texterfassung, Korrektur sowie Konvertierung nach XML entsprechend dem Basisformat des Deutschen Textarchivs (DTABf)[2] zu ermöglichen, fanden wöchentliche Arbeitstreffen statt. Im Rahmen dieser regelmäßigen Besprechungen wurden auf der Grundlage einer Ausdifferenzierung der Ressourcenfunde die zu konvertierende Texte jeweils dem für die jeweilige Textsorte zuständigen Bearbeiter zugewiesen, der dann die Transkription des Textes vornahm bzw., wo vorhanden, eine existierende Transkription aus dem Ausgangsformat nach TEI-XML konvertierte und ggf. korrigierte. Dabei konnten folgende projektbezogene Integrationsszenarien thematisiert, evaluiert und realisiert resp. erprobt werden:

(i) Nutzung von Quellen aus älteren Forschungsprojekten,
(ii) Nutzung von Quellen aus laufenden Projekten (Kontakt zur Forschungsgruppe *Nürnberger Texte des 15. Jahrhunderts* um PD Dr. Heike Sahm (Universität Siegen)),
(iii) Nutzung von Quellen aus dem laufenden Promotionsprojekt von Anna Michalak (Texte zur ersten Frauenbewegung),
(iv) Nutzung der Quellen von thematisch orientierten Arbeitsgruppen (ebenfalls Texte zur ersten Frauenbewegung) sowie
(v) Nutzung der Quellen von EinzelforscherInnen.

Wie die tabellarische Übersicht im Anhang zeigt, hat die Gießener Arbeitsgruppe bisher eine Konvertierung von 2340 Seiten Text nach XML vorgenommen.

[2] Zum DTA-Basisformat siehe www.deutschestextarchiv.de/doku/basisformat.

Arbeiten der Wolfenbütteler Arbeitsgruppe im Berichtszeitraum

Im Rahmen des vom Niedersächsischen Ministeriums für Wissenschaft und Kultur geförderten Forschungsprojekt „Obrigkeitskritik und Fürstenberatung: Die Oberhofprediger in Braunschweig-Wolfenbüttel 1568–1714“ zwischen der Herzog August Bibliothek Wolfenbüttel (HAB) und dem Interdisziplinären Institut für Kulturgeschichte der Frühen Neuzeit (IKFN) der Universität Osnabrück wurden in einem umfassenden Digitalisierungsanteil wichtige Schlüsselquellen der wissenschaftlichen Forschung zur Verfügung gestellt. Die Digitalisierung der ausgewählten Drucke des 16. bis 18. Jahrhunderts aus den Beständen der HAB umfasst zum einen Werke Wolfenbütteler Oberhofprediger aber auch zentrale theologische Texte sowie Schriften zur regionalen Landes- und Kirchenverfassung. Insgesamt wurden über 300 frühneuzeitliche Drucke digitalisiert, von denen mehr als zwei Drittel mit einem Umfang von ca. 35.000 Seiten (bzw. Images) nicht nur in Bildform sondern als maschinenlesbare Volltexte in der Wolfenbütteler Digitalen Bibliothek[3] zur Verfügung stehen.

Die Volltexte wurden von der Grepect GmbH per Double-Keying-Verfahren erstellt, dabei wurden elementare Textstrukturen sowie Referenzen zu den Faksimiles des jeweiligen Druckes ausgezeichnet.
Durch semiautomatische Transformationsprozesse wurden die vom Dienstleister gelieferten Texte in das Basisformat der Wolfenbütteler Digitalen Bibliothek entsprechend den Regeln von TEI-P5 konvertiert. Ebenfalls nach den Regeln von TEI-P5 wurden die Metadaten für die Texte und die Präsentation und Steuerung der Texte in den WDB obligatorischen XSLT-Stylesheets und mets.xml-Dateien erstellt. Durch die Speicherung der Volltexte in der XML-Datenbank eXist/Lucene kann eine Volltextsuche angeboten werden. Innerhalb der WDB-Ansicht werden die Texte zusammen mit den dazugehörenden digitalen Faksimiles präsentiert, wobei verschiedene Ansichten möglich sind.
Digitalisate und Volltexte sind im OPAC der HAB und zudem im Forschungsportal des Projektes nachgewiesen.[4]

[3] http://www.hab.de/de/home/bibliothek/digitale-bibliothek-wdb.html.
[4] Siehe http://www.oberhofprediger.de.

Seit dem 2. Januar 2013 arbeitet Marcus Baumgarten als wissenschaftliche Hilfskraft im Kurationsprojekt. Nach einer kurzen Einarbeitungsphase korrigierte er die von der Herzog August Bibliothek Wolfenbüttel für das Projekt in die Korpora des Deutschen Textarchivs (DTA)[5] eingebrachten Volltexte aus dem Oberhofprediger-Projekt der Bibliothek. Unter Berücksichtigung des DTA-eigenen TEI-Schemas wurden offensichtliche Fehler, die in den insgesamt 40 Texten hervorgehoben waren, entsprechend korrigiert. Dabei handelte es sich in der Hauptsache um unleserliche Zeichen und Abbreviationen, die nach Möglichkeit aufgelöst wurden.
Zurzeit arbeitet die Wolfenbütteler Arbeitsgruppe an der Korrektur des ersten „Theatrum Europaeum"-Bandes von Abelinus und der Transformation der jüngeren Übersetzung des Petrus de Crescentiis aus dem LaTeX-Format in das TEI-Format zur Integration in die Wolfenbütteler Digitale Bibliothek.

Arbeiten der Berliner Arbeitsgruppe im Berichtszeitraum

Die Berliner Arbeitsgruppe, bestehend aus dem Koordinator des Kurationsprojekts Christian Thomas und den beiden studentischen Hilfskräften Elena Kirillova und Frederike Neuber, entwickelte eine Matrix zur Bewertung und somit zur kriteriengestützten Auswahl geeigneter Textressourcen (Arbeitspaket 2). Auf Grundlage dieser Bewertungsmatrix wurden im Anschluss die zu integrierenden Texte ausgewählt (Arbeitspaket 3).

Die Berliner Arbeitsgruppe leistete und leistet den anderen Projektpartnern außerdem Hilfestellung bei der Konvertierung von Texten in das Basisformat des Deutschen Textarchivs (DTABf)[6]. Erste Probetranskriptionen der Gießener Gruppe wurden begutachtet und intensiv kommentiert bzw. korrigiert.
Die Berliner Arbeitsgruppe arbeitet in jedem Arbeitsschritt eng mit dem Deutschen Textarchiv (DTA) an der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) zusammen. Die zur Integration ausgewählten Texte wurden in das DTABf konvertiert und als Teilkorpora in das Erweiterungsmodul des Deutschen Textarchivs DTAE[7] aufgenommen (Arbeitspaket 4). Die Texte stehen dadurch zugleich über die webbasierte Plattform zur Qualitätssicherung DTAQ[8] zur Verfügung. Die zu den via DTAE/DTAQ integrierten Texten gehörigen Bilddigitalisate wurden – sofern nötig – in einzelne JPEG- bzw. TIFF-Dateien

[5] Siehe dazu den Abschnitt zur Berliner Arbeitsgruppe
[6] Zum DTA-Basisformat siehe www.deutschestextarchiv.de/doku/basisformat.
[7] Deutsches Textarchiv – DTAE, www.deutschestextarchiv.de/dtae.
[8] Deutsches Textarchiv – DTAQ. www.deutschestextarchiv.de/dtaq..

segmentiert, mit fortlaufenden Dateinamen versehen und auf diese Weise mit den `@facs`-Attributen der TEI-XML-Transkriptionen aliniert. Die Bilddigitalisierung einzelner, ausgewählter Titel, die aus dem Sachmittel-Budget des Kurationsprojekt finanziert wird, wurde von der Berliner Arbeitsgruppe koordiniert. Die Bilddigitalisate werden – ebenso wie sämtliche die in DTAE/DTAQ integrierte XML-Volltexte – an der BBAW vorgehalten.

Eine Übersicht über die von der Berliner Arbeitsgruppe konvertierten Texte findet sich auf http://www.deutschestextarchiv.de/dtae, wo die Teilkorpora „Gutenberg-DE", „gutenberg.org", „Gutzkow Editionsprojekt", „HAB", „Wikisource", „Monumenta Culinaria" und einzelne Texte unter „DTAE" zum Kurationsprojekt zählen.[9]

Insgesamt wurden im Berichtszeitraum auf Berliner Seite 42 247 Textseiten mit 71 672 863 Zeichen nach TEI-XML entsprechend dem DTABf konvertiert und in DTAE/DTAQ integriert (Stand: 11.2.2013). Damit ist zwar die im Projektplan für das Kurationsprojekt angestrebte Zahl von „mindestens 35 000 Druckseiten" für das Kurationsprojekt bereits zu diesem Zeitpunkt überschritten, es ist dabei allerdings zu bedenken, dass „Kuration" im hier gemeinten Sinne noch weitere Arbeit an den Texten beinhaltet, die *nach* der Integration der Texte in DTAQ erfolgt, um die Möglichkeiten, die diese Plattform gerade für die kollaborative Arbeit an Texten bietet, nutzen zu können.
Die große Zahl bisher integrierter Textseiten konnte vor allem dank der ebenfalls im Kurationsprojekt beteiligten HAB Wolfenbüttel erreicht werden, die bisher allein 22 014 Seiten Text aus dem interdisziplinären Forschungsprojekt „Obrigkeitskritik und Fürstenberatung: Die Oberhofprediger des Fürstentums Braunschweig-Wolfenbüttel 1568–1714" zur Verfügung stellen konnte. Metadaten zu den von der HAB Wolfenbüttel übernommenen Texten (siehe oben, Abschnitt Wolfenbütteler Arbeitsgruppe) wurden von der Berliner Arbeitsgruppe als Teile des TEI-Headers übernommen und wo nötig ergänzt. Die Konvertierung in das CLARIN-D-Metadatenformat CMDI erfolgt automatisiert. Die Titelblätter der Bände wurden von den studentischen Hilfskräften der Berliner Arbeitsgruppe mit Hilfe von XML-Tags nachstrukturiert (explizite Kennzeichnung von Autor, Titel, Untertitel, Verlag, Druckort usw.). Die Texte aus dem Oberhofprediger-Projekt werden in der nächsten Zeit noch einige intensive Bearbeitungen erfahren, um den Zielen des Kurationsprojekt bzw. allgemeiner: von CLARIN-D noch besser zu entsprechen. Marcus Baumgarten (HAB) übernahm Verbesserungen am Tagging der Texte, insbesondere die

[9] Zu den Textquellen siehe www.deutschestextarchiv.de/doku/textquellen.

Bearbeitung und Dokumentation unleserlicher Textstellen, die vom Dienstleister im Oberhofprediger-Projekt zunächst nicht transkribiert wurden.
Ein weiterer wichtiger Faktor für die Überschreitung der angestrebten Seitenzahl bereits zu diesem frühen Zeitpunkt der Projektdauer ist die gelungene Kuration von 17 985 Seiten Text aus der freien Quellensammlung Wikisource. Hierbei kam dem Kurationsprojekt vor allem die Expertise der Arbeitsgruppe des DTA, namentlich des im DTA für die Software-Entwicklung und Webapplikation zuständigen Mitarbeiters Frank Wiegand zugute. Unter seiner Anleitung konnte ein Workflow zur Integration von Wikisource-Texten in die DTA-Infrastruktur entwickelt werden, durch den sich der – gleichwohl unumgängliche und noch immer zeitintensive – manuelle Aufwand bei der Konvertierung der sperrigen Wikisyntax nach TEI-XML entsprechend DTABf auf ein Mindestmaß reduzieren ließ.

Ausblick: Für die kommenden Monate zieht die Berliner Arbeitsgruppe mehrere externe Forschungsprojekte als zukünftige Textquellen in Betracht, z.B.:

1. *Kommentierte digitale Gesamtausgabe der Werke und Briefe Karl Gutzkows* (http://projects.exeter.ac.uk/gutzkow/Gutzneu/gesamtausgabe/index.htm)
   Ein erstes Test-Kapitel wurde bereits aus dem seitens der Gutzkow-Edition erstellten HTML in das DTABf konvertiert (www.deutschestextarchiv.de/dtaq/book/show/gutzkow_narren_1832) und dem Editionsprojekt mit detaillierten Kommentaren zur Begutachtung zur Verfügung gestellt. Die Herausgeber, Prof. Dr. Martina Lauster und Prof. Dr. Gert Vonhoff (beide Exeter, GB) sind an einer weiteren Zusammenarbeit interessiert und werden mit dem zuständigen Gremium im April 2013 über eine Kooperation der Gutzkow-Edition mit dem Kurationsprojekt beraten. Im Fall eines positiven Entscheides des Gremiums sind noch viele weitere Bände der Gesamtausgabe[10] zu erwarten. Es sind dies qualitativ hochwertige Transkriptionen nach den Erstausgaben, die (mit vertretbarem manuellem Aufwand) als Subkorpus ‚Gutzkow' integriert werden könnten.
2. *Georg Simmel Online* (http://socio.ch/sim/simmel_pub.htm)
   Mit Prof. Dr. Hans Geser (Universität Zürich, Soziologisches Institut) als Vertreter des Projekts wurde bereits die Übernahme der Texte durch das DTA bzw. das Kurationsprojekt vereinbart, so dass einer zukünftigen Integration von umfangreichen Texten nichts mehr im Wege steht.

[10] Siehe dazu den Überblick der der Digitalen Ausgabe unter http://projects.exeter.ac.uk/gutzkow/Gutzneu/gesamtausgabe/index.htm.

3. Weitere potentielle Quellen
   (geeignete Texte werden anhand der oben erwähnten Bewertungsmatrix ausgewählt)
   - *Ngiyaw-eBooks*, (http://ngiyaw-ebooks.org/)
   - *The Sophie-Project* (http://sophie.byu.edu/)
   - Universität Leipzig, Institut für Allg. Psychologie: Sammlung *Wilhelm Wundt (1832–1920) und die Anfänge der experimentellen Psychologie* (http://www.uni-leipzig.de/~psycho/wundt/opera/)

Außendarstellung des Kurationsprojekt

Neben den oben beschriebenen konkreten Aktivitäten aller Partner zur aktiven Kuration und Integration von Textressourcen ist das Kurationsprojekt seit Projektbeginn im Rahmen von Konferenzen und in Publikationen vorgestellt worden:

- *Full Paper*: Christian Thomas, Frank Wiegand: Making great work even better. Appraisal and Digital Curation of widely dispersed Electronic Textual Resources (c. 15th–19th cent.) in CLARIN-D. Full Paper for the International Conference "Historical Corpora 2012", December 6–9, 2012; Goethe University, Frankfurt, Germany. [urn:nbn:de:kobv:b4-opus-23081 – online-Version 2012-10-31]
- *Newsletter-Beitrag*: Frederike Neuber, Christian Thomas: Vorstellung des Kurationsprojekts 1 der Clarin-D-FAG 1 »Deutsche Philologie«. In: Clarin-D-Newsletter, Nummer 3, 2012, November, S. 11–13. [http://www.clarin-d.de/images/newsletter/CLARIN-D-Newsletter-2012-3.pdf]
- *Vortrag:* Alexander Geyken, Thomas Gloning: A living text archive of 15th–19th c. German. Corpus strategies, technology, organization. International Conference on "Historcal Corpora 2012", 6.–9.12.2012, Johann Wolfgang Goethe-Universität, Frankfurt (Main). (Abstract)
- *Vortrag:* Christian Thomas: Making great work even better: Appraisal and Digital Curation of widely dispersed Electronic Textual Resources (c. 15th–19th cent.) in CLARIN-D. International Conference on "Historcal Corpora 2012", 6.–9.12.2012, Johann Wolfgang Goethe-Universität, Frankfurt (Main)
- *Beiträge zur DTA-/CLARIN-D-Konferenz und -Workshops (18.–19. Februar 2013):* Zum Thema „Historische Textkorpora für die Geistes- und Sozialwissenschaften.

Fragestellungen und Nutzungsperspektiven“ veranstalteten das Deutsches Textarchiv und CLARIN-D eine gemeinsame Konferenz mit begleitenden Workshops. Einer der Workshops schloss an die Erfahrungen des Kurationsprojekts und des DTA-Erweiterungsmoduls DTAE und vermittelte, auf welche Weise existierende oder neue Sprachressourcen mit den Hilfsmitteln des DTA CLARIN-D-konform aufbereitet oder erstellt werden können.

Zusätzlich zu diesem Workshop bot die Berliner Arbeitsgruppe, unterstützt von MitarbeiterInnen aus dem Deutschen Textarchiv bzw. CLARIN-D am 20.2.2013 der Gießener Arbeitsgruppe und Vertretern der Siegener Forschungsgruppe *Nürnberger Texte des 15. Jahrhunderts* ein intensives vierstündiges Tutorial an. Wichtige Hinweise zur DTABf- und CLARIN-D-kompatiblen Transkription und Aufbereitung von Textressourcen sowie die im Vorfeld gesammelten Rückfragen der Partner wurden besprochen. Allgemein wurde vertiefend auf das DTABf eingegangen, insbesondere in die Arbeit mit dem oXygen XML-Editor und dem vom DTA entwickelten speziellen oXygen-Framework eingeführt.