DH's Next Top-Model? Digitale Editionsentwicklung zwischen Best Practice und Innovation am Beispiel des „Corpus Masoreticum“

Liedtke, Clemens
Clemens.Liedtke@hfjs.eu
Hochschule für Jüdische Studien Heidelberg, Deutschland

Einführung: Handschriftenforschung und Digitale Transformation

Wenn sich auch in den Geisteswissenschaften so etwas wie ein „Digital Turn“ oder eine „Digitale Transformation“ (Pousttchi 2017) beobachten lässt, wirft das unweigerlich die Frage auf, inwiefern die Einführung digitaler Prozesse in geisteswissenschaftliche Forschung Konsequenzen für den Aufbau bzw. das Design von konkreten Forschungsprojekten hat.

In folgendem Vortrag soll diese Frage an einem Fallbeispiel aus dem Bereich der Jüdischen Studien entwickelt werden. Hier wurde unlängst von Gerben Zaagsma herauspräpariert, dass durch die nunmehr in großem Umfang verfügbaren digitalen Ressourcen zu jüdischer Geschichte und Kultur sich "die Sicherung, Bereitstellung und Analyse des in alle Welt verstreuten vielsprachigen und mehrschriftlichen Quellenmaterials“ als eine der zukünftigen Schlüsselaufgaben stellt (Zaagsma 2019:3ff.).

Dies lässt sich in besonderem Maße an dem Teilbereich der Handschriftenforschung exemplifizieren: Gerade durch die großen Digitalisierungsinitiativen der letzten Jahre an hebräischen Manuskripten, v.a. National Library of Israel, der Polonsky Foundation in Zusammenarbeit mit der British Library und der Bodleian Library stehen der wissenschaftlichen Community umfangreiche Quellenbestände zur Verfügung, die erst die Grundlage für weitere inhaltliche Tiefenerschließung, Edition und Corpusanalyse bilden.

Betrachtet man weiter das Teilgebiet der wissenschaftlichen Erforschung des hebräischen Bibeltextes, so lässt sich zeigen, dass forschungsgeschichtlich gesehen die Editionspraxis hebräischer Bibelhandschriften bereits teilweise auf digitale Prozesse zurückgreifen kann, etwa in computerlinguistischer Hinsicht durch die langjährigen Projekte des Eep Talstra Centre for Bible and Computer (ETCBC[1]; van Lit 2019) oder durch die elektronische Edition des Westminster Leningrad Codex (WLC[2]). Ob man dies bereits als Symptom eines digitalen Transformationsprozesses betrachten sollte, der jenseits der Einführung einzelner Tools und Verfahren übergreifende Veränderungen in Fragestellung, Methodologie und Methodenkritik sichtbar macht, ist damit noch nicht ausgemacht. Immerhin gilt nach wie vor die Print-Edition der "Biblia Hebraica Stuttgartensia" auf Grundlage der Handschrift Ms. Fircovitch B19a als eine der massgeblichen kritischen Textausgaben für Theologie und hebräische Bibeltextforschung und bildet damit noch den Stand und die Möglichkeiten der analogen Bibeltextkritik ab.

Hinsichtlich der Perspektive von primär digitalen wissenschaftlichen Editionen der Hebräischen Bibel stellt sich infolgedessen die Frage nach der Anwendbarkeit bereits etablierter Verfahren; während sowohl im WLC als auch im "Digital Mishnah Project"[3] XML-Textauszeichnungen zum Einsatz kommen (entweder teilweise oder vollständig entlang der TEI P5 Spezifikationen implementiert), finden XML-basierte "best practices“ im Bereich linksläufiger Schriftsysteme wie dem Hebräischen bislang nur zögerlich Akzeptanz, zumal unter Einsatz von XML-Quelltext-Autorensystemen wie oXygen ganz basale handwerkliche Probleme das Schreiben von rechtsläufigen Tagsets und linksläufigen Schriften zur Herausforderung macht. So konstatiert auch noch das DARIAH Wiki: „Solange dieses Problem nicht grundsätzlich gelöst ist, wird die Akzeptanz von TEI und/oder XML in Hebraistik und Arabistik gering sein.“[4]. Gleichwohl berührt dies eher die Frage, inwiefern sich solche technischen Hürden durch geeignete grafische Benutzerschnittstellen nehmen lassen, die die systemischen Anforderungen bidirektionaler Unicode-Texte von anwendungsseitigen Annotationsebene wegabstrahieren.

Textcodierung: Modelle

Inhaltlich bedeutsamer scheint aber die mittlerweile ausführlich beschriebene Problematik zu sein, dass sich XML als semi-strukturierte, hierarchisch organisierte Markup-Sprache mit seinen strikten Regeln zur Wohlgeformtheit und Validität von Auszeichnungen nur bedingt dazu eignet, Phänomene zu annontieren, die nicht linear/hierarchisch, sondern mit Überlappungen oder sich überschneidenden Sequenzen strukturiert sind. Ebenso zwingt es die Bearbeitenden, die dokumentenzentrierte und die textzentrierte Perspektive einer zu edierenden Quelle durch zwei unterschiedliche Kodierungsstrategien zu lösen (Brüning/Henzel/Pravida 2013; Pierazzo 2017); gleichzeitig belastet die Verarbeitung von internen wie externen Verweisstrukturen im Dokument (Lesartvarianten, Zitate, Querbezüge) die Kodierung

damit, die referentielle Integrität von Links zuverlässig verwalten zu können. Innovative Lösungsansätze werden für dieses Problem unter anderem entlang des Modells von Textvarianten-Graphen beschrieben (Schmidt 2008; Schmidt 2009; Schmidt/Colomb 2008:498) oder unter Verwendung von "Labelled Property Graph"-Systemen wie der Graphendatenbank Neo4J diskutiert (Kuczera 2016a, Kuczera 2016b).

An dieser beispielhaften Gegenüberstellung verschiedener Datenmodellierungsansätze einen Unterschied zwischen Standardverfahren/Best Practice und Innovation auszuloten, der bereits eine implizite Wertung von Innovation als „fortschrittlich" mitmeint, griffe sicherlich zu kurz - gleichwohl spannt sich durch die in der Literatur diskutierten Anwendungsfälle durchaus ein Spannungsfeld auf: Einerseits belastet der Einsatz spezialisierter Datenbankmanagementsysteme[5] die Anforderungen an Offenheit und Langzeitverfügbarkeit von zu speichernden Forschungsdaten; auch die Neumodellierung von zu erhebenden Forschungsdaten schneidet im Sinne der FAIR-Prinzipien (Wilkinson / Dumontier / Aalbersberg, *et al.* 2016) zunächst von Anschlussmöglichkeiten ab, sind doch neuartige Datenmodelle, möglicherweise eben noch nicht „interoperable" und „reusable".

Andererseits bedeutet auch das Anwenden bestehender *best practices* eine interpretative Einschränkung: Mit der Umsetzung standardisierter Auszeichnungsschemata, sei es TEI-XML, eine bestimmte RDF-Ontologie oder Datenbankstruktur lässt sich am Quellenmaterial nur beobachten, was sich innerhalb der Unterscheidungsmöglichkeiten des jeweiligen Schemas bezeichnen lässt. Die Praxis der XML-basierten Quellenannotation zeigt hier, dass gerade bei steigenden Komplexitätsgraden am Material sich der Focus stärker in Richtung auf Einhaltung der Schema-Compliance und weg von der Beschreibung neuer Merkmalskategorien bewegt. Gute Indikatoren für dieses Phänomen sind beispielsweise vermehrter Einsatz von Standoff-Markup, individuelle, d.h. projektbezogene Schema-Erweiterungen, aber auch steigende Mehrdeutigkeiten im Markup bestimmter Phänomene wie etwa Marginalien in Handschriften (Estill 2016).

Dieses hier am Beispiel zweier Modellierungsstrategien angedeutete Spannungsverhältnis zwischen Standardisierung und Innovation lässt sich gerade im Rahmen des Projektdesigns produktiv nutzen, zwingt doch zum einen das Einführen digitaler Methoden in die Quellenerschliessung zu einer strengen Formalisierung des eigenen Forschungsprozesses, zum anderen gewinnt die Perspektive der Datenmodellierung (Owens 2011) eine größere Bedeutung. Beides hat nicht zuletzt auch entscheidenden Einfluss auf die Auswahl der zum Einsatz kommenden Technologie-Stacks.

Fallbeispiel: Corpus Masoreticum

Am folgenden Fallbeispiel soll entwickelt werden, wie die skizzierten Überlegungen in einem konkreten Projekt umgesetzt werden können: Das von der Deutschen Forschungsgemeinschaft geförderte Langzeit-Editionsvorhaben „Corpus Masoreticum", das an der Hochschule für Jüdische Studien Heidelberg angesiedelt ist, befasst sich mit dem sogenannten masoretischen Text in mittelalterlichen Bibelcodices. In der heutigen Forschung meint der Begriff der Masora alle meta-textuellen Elemente zum Konsonantentext der Hebräischen Bibel. Dazu gehören Grapheme, grammatische, syntaktische und statistische Notizen, Referenzen und Verweise. Ab dem 12. Jh. entstehen im Kulturraum Aschkenas (Nord-Frankreich und Deutschland) hebräische Bibel-Kodizes, in denen die Masora mit mikrographischer Schrift in ornamentalen Formen auf der Seite platziert wurde und als Fabelwesen, vor allem aber als zoomorphe Gestalten (Hunde, Pferde, Hasen, Gazellen, Vögel, Fische) und sogar als anthropomorphe Darstellungen gestaltet werden - hierfür wurde der Begriff der Masora figurata zur Unterscheidung von linearer Masora magna geprägt. Sie kann darüber hinaus Zitate aus Kommentarliteratur enthalten, die weit über die üblichen quantitativen und referentiellen Annotationen zum hebräischen Konsonantentext hinausgingen (vgl. Ms Vat. ebr. 14). Als Beispiel für in diesen masoretischen Metatexten häufig enthaltenes Listenmaterial lassen sich die sogenannten „Okhla-Listen" herausgreifen, in denen als bewahrenswert gedachte Textphänomene und Schreibungen in unterschiedlichen Strukturen und Layouts dem Bibeltext mitgegeben werden und ihrerseits auf verschiedene extern überlieferte Rezensionen dieser Listen referieren (als Überblick: Liss/Petzold 2016).

Bereits oberflächliche Untersuchungen an diesem sehr speziellen Quellenmaterial zeigen, dass hier besonders komplexe Anforderungen an das zu definierende Editionsdatenmodell gestellt werden: Zu dokumentieren ist nicht nur linearer Text, sondern hochgradig vernetzte interne und externe Verweisstrukturen nicht nur mit Bezug auf Lesartvarianten, sondern auch auf Kommentarliteratur und Listenmaterial mit spezifischen Listenmustern, die auf extern tradiertes Listenwissen verweisen. Darüber hinaus bedarf die doppelte Lesbarkeit von Masora figurata als Text und Bild gleichermaßen in ihrem Bezug zum Bibeltext eines besonderen Dokumentationsverfahrens.

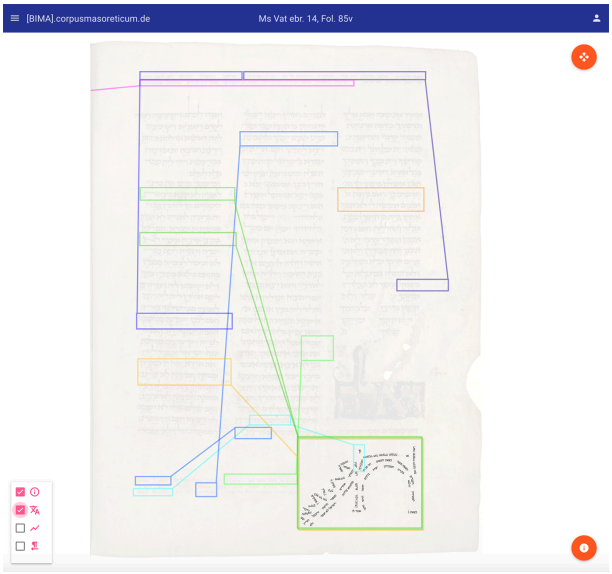

Abbildung 1: Überblick über das beispielhafte mise-en-page von Bibeltext, Masora parva, Masora Magna und Masora figurata in Ms Vat ebr. 14, Fol. 85v. Prototyp einer Visualisierungssoftware für digitale Erschliessung hebräischer Bibelcodices. Quelle: http://bima.corpusmasoreticum.de/figurata/tor

Implementierung von Modellen und Workflows

Durch die netzwerkartige Struktur der in die untersuchten Handschriften eingebetteten Metatexte lag es nahe, die Beschreibung von Text als Daten von vornherein als Graph zu modellieren; der „labelled-property“-Ansatz von Graphdatenbankensystemen wie Neo4J macht es durch seine sogenannte „whiteboard-friendliness“[6] möglich, einfache Modellskizzen rasch in lauffähige digitale Speichermodelle zu implementieren. Im Editionsworkflow wird zunächst der Import von Handschriftendigitalisaten samt Metadaten über IIIF-kompatible Archive realisiert und die Handschriftendaten im Importprozess in Graphendaten als Knoten und Kanten umgewandelt. Ab hier werden über eine grafische Benutzeroberfläche erzeugte Texttranskriptionen kontextbezogen als Datenknoten verlinkt, wobei der Bezug zum Digitalisat über die Kodierung von Text als SVG-Textpfaden erhalten bleibt. Transkriptionen werden bei Bedarf weiter tokenisiert, um weitere Metadaten oder Kontextrelationen in den Graphen integrieren zu können. Aus dem so generierten Text- bzw. Knowledge Graph lassen sich Subsets (Datenaggregate) generieren, die im späteren Prozess sowohl als TEI-XML, RDF-Graph oder auch als angereicherte IIIF-Manifeste (Text, Übersetzung, Kommentar) ausgeliefert werden können.

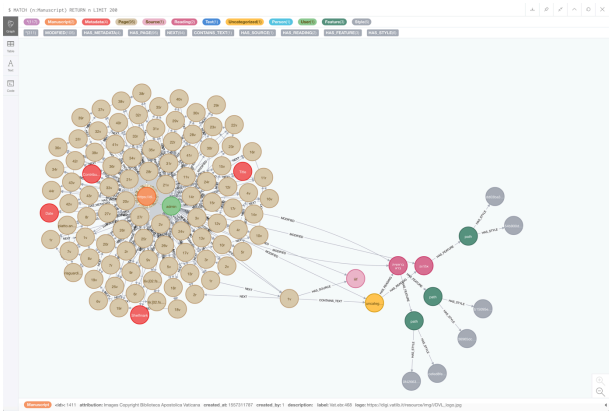

Abbildung 2: Beispielgraph anhand Cod. Vat. Ebr. 468, folio 1v, Darstellung in Neo4J. Quelle: Liedtke 2019 (in Vorbereitung)

Die technischen Komponenten sind stark modularisiert und, wo möglich, als Microservices implementiert, so dass die einzelnen Ressourcen der Edition mit REST-APIs ausgeliefert und abgefragt werden können. Die Softwarearchitektur ist als Docker-Container-Umgebung in einer skalierbaren Cloud-Computing Landschaft realisiert, die vom heiCloud-Service des Rechenzentrums der Universität Heidelberg bereitgestellt wird, wobei die Rahmenbedingungen von Langzeitarchivierung und Nachnutzbarkeit in Zusammenarbeit mit Fachdiensten der Universitätsbibliothek Heidelberg[7] gewährleistet werden.

Ausblick

Bezieht man das Spannungsverhältnis von Standardisierung und Innovation digitaler Prozesse in den architektonischen Aufbau eines geisteswissenschaftlichen Forschungsprojektes ein, lässt sich diese Dynamik produktiv für die Entwicklung eigener Workflows nutzen und öffnet Spielräume für das Modellieren der eigenen Forschungsdaten. Die formale Beschreibung eines digitalen Datenmodells kann dann methodenkritisch dazu verwendet werden, die im Prozess anstehenden Ergebnisse wieder an die Ausgangsfragestellung rückzubinden und Modelle auf ihre Plausibilität zu prüfen. Die Umsetzung in Technologie-Stacks oder digitale Frameworks führt im gezeigten Fallbeispiel zu der Konstruktion einer quasi „hybriden“ Editionsumgebung und beschränkt den digitalen Anteil eines Projektes nicht nur auf das Ausliefern von „Tools“, sondern betrachtet den Aspekt der digitalen Transformation als integralen Bestandteil des gesamten Forschungsprozesses.

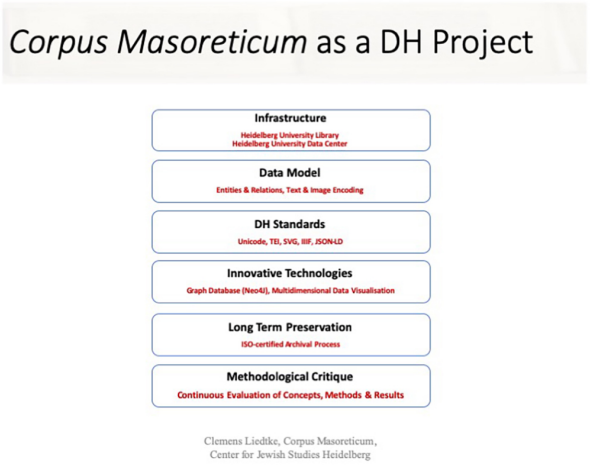

Abbildung 3: Corpus Masoreticum als DH-Projekt (Schema). Quelle: Liedtke 2019 (in Vorbereitung)

Fußnoten

1. http://etcbc.nl
2. http://tanach.us/
3. http://dev.digitalmishnah.umd.edu/
4. https://wiki.de.dariah.eu/display/publicde/3.5+Judaistik+und+Hebraistik
5. Die auch im nachfolgenden Fallbeispiel verwendete Community-Version von Neo4J ist unter einer GPLv.3 Lizenz als Open Source verfügbar, ist also im strengen Sinne keine proprietäre Software – die Entwicklung wird aber massgeblich von Neo4J Inc. als kommerziellem Unternehmen vorangetrieben.
6. https://neo4j.com/developer/guide-data-modeling/#whiteboard-friendly
7. https://heidata.uni-heidelberg.de/

Bibliographie

Estill, Laura (2016): "Encoding the Edge: Manuscript Marginalia and the TEI." *Digital Literary Studies* 1, no. 1. https://journals.psu.edu/dls/article/view/59715/59912.

Kuczera, Andreas (2016a): "Graphbasierte Digitale Editionen." Blog. *Mittelalter*. Interdisziplinäre Forschung Und Rezeptionsgeschichte (blog), April 19, 2016. https://mittelalter.hypotheses.org/7994.

Kuczera, Andreas (2016b): "Digital Editions beyond XML – Graph-Based Digital Editions.", in: *Proceedings of the 3rd HistoInformatics Workshop on Computational History* (HistoInformatics 2016), edited by Marten Düring, Adam Jatowt, Johannes Preiser-Kappeller, and Antal van Den Bosch. Krakow, 2016. http://ceur-ws.org/Vol-1632/paper_5.pdf.

Liedtke, Clemens (2019): 'How am I supposed to read this?' Challenges and Opportunities of Medieval Western Masorah as a Digital Scholarly Edition", in: J. Leipziger/H. Liss/K. J. Petzold (eds.), *Philology and Aesthetics:* Figurative Masorah in Western European Manuscripts (Judentum und Umwelt), Frankfurt am Main et al.: Peter Lang (in Vorbereitung)

Liss, Hanna / Petzold, Kay Joe (2016): "Die Erforschung der westeuropäischen Bibeltexttradition als Aufgabe der Jüdischen Studien. Ein halbes Jahrhundert Forschung und Lehre über das Judentum in Deutschland.", in: *Orchidee oder Mimose.* Versuch einer Standortbestimmung der Jüdischen Studien, edited by Andreas Lehnardt and Guiseppe Veltri. Berlin et al, 2016.

Owens, Trevor (2011): "Defining Data for Humanists: Text, Artifact, Information or Evidence?" *Journal of Digital Humanities* 1, no. 1. http://journalofdigitalhumanities.org/1-1/defining-data-forhumanists-by-trevor-owens/.

Pierazzo, Elena (2017): "Facsimile and Document-Centric Editing.", in: *Creating a Digital Scholarly Edition with the Text Encoding Initiative*, edited by Marjorie Burghart. https://www.digitalmanuscripts.eu/wp-content/uploads/sites/6/2017/09/05-Digital-Facsimiles-EP.pdf.

Pousttchi, Key (2017): "Digitale Transformation.", in: *Enzyklopädie der Wirtschaftsinformatik.* http://www.enzyklopaedie-der-wirtschaftsinformatik.de/lexikon/technologien-methoden/Informatik--Grundlagen/digitalisierung/digitale-transformation/digitale-transformation/?searchterm=digitale%20transformation.

Schmidt, Desmond (2008): "What's a Multi-Version Document?" *Multi-Version Documents* (blog), May 3. http://multiversiondocs.blogspot.com/2008/03/whats-multi-version-document.html.

Schmidt, Desmond / Colomb, Robert (2009): "A Data Structure for Representing Multi-Version Texts Online." *International Journal of Human-Computer Studies* 67, no. 6 (June 2009): 497–514. https://doi.org/10.1016/j.ijhcs.2009.02.001.

Schmidt, Desmond (2009): "'Merging Multi-Version Texts: A Generic Solution to the Overlap Problem.' Presented at Balisage: The Markup Conference 2009, Montréal, Canada, August 11 - 14, 2009," Proceedings of Balisage: The Markup Conference 2009, 2 (2009). https://doi.org/10.4242/BalisageVol3.Schmidt01 .

Wilkinson, M. / Dumontier, M. / Aalbersberg, I. et al. (2016): "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data* 3 **,** 160018. doi:10.1038/sdata.2016.18

Zaagsma, Gerben (2018): "#DHJewish – Jewish Studies in the Digital Age." *Medaon.* Magazin für Jüdisches Leben in Forschung und Bildung, no. 12: 1–11.

Zundert, Joris J. van / Andrews, Tara L. (2016): "Apparatus vs. Graph: New Models and Interfaces for Text.", in: *Interface Critique*, edited by Florian Hadler and Joachim Haupt, 139:183–206. Berlin: Kulturverlag Kadmos.