Herausforderung "Big Data" in der historischen Forschung

Kruse, Sebastian; Schmaltz, Florian; Stiller, Juliane; Wintergrün, Dirk

Max-Planck-Institut für Wissenschaftsgeschichte

Datenintegration und die Verfügbarmachung großer Textkorpora als Quellen für die zeitgeschichtliche Forschung, insbesondere in der Wissenschaftsgeschichte, stellen immer noch eine große Herausforderung dar. In unserem Beitrag stellen wir Methoden und Tools vor, die dieser Herausforderung begegnen und Lösungsansätze aufzeigen sollen. Vorgestellt wird dieses am Beispiel des auf 7 Jahre angelegten Forschungsprogramms zur Geschichte der Max-Planck-Gesellschaft (MPG), das im Juni 2014 begonnen wurde.[1] Ziel des Forschungsprogramms ist die Geschichte der Max-Planck-Gesellschaft von ihrer Gründung im Jahre 1948 bis zum Ende der Präsidentschaft von Hubert Markl 2002 aufzuarbeiten. Ziel ist es hierbei nicht eine additive Geschichte der 80 existierenden und 20 geschlossenen Institute der Max-Planck-Gesellschaft zu schreiben, sondern im Zentrum stehen institutsübergreifende Fragestellungen zu Themenfeldern wie Periodisierungen, Innovationen, Internationalisierung, Forschung und Wirtschaft, Gender und Wissenschaft sowie Konkurrenz und Kooperation.
Ein weiteres Ergebnis des Forschungsprogramms wird es sein, konzeptionelle und epistemologische Perspektiven aufzeigen, wie aus der elektronischen Quellen- und Datenüberlieferung ein digitales Forschungsarchiv generiert werden kann. Damit werden der MPG als Wissenschaftsorganisation neue selbstreflexive Erkenntnismöglichkeiten aus diesen digitalen Wissensspeichern eröffnet.

Für dieses Forschungsprogramm sollen große Aktenbestände und Publikationen, wie Tätigkeitsberichte und Jahrbücher der Max-Planck-Gesellschaft, digital erschlossen werden. Darüber hinaus sollen vorhandene digitale Quellen in einer virtuellen Forschungsumgebung integriert werden und kollaboratives Arbeiten unter den Historikern mit digitalen Tools ermöglicht und unterstützt werden.
Der Umfang der zu digitalisierenden Quellen mit unterschiedlichsten Provenienzen und zu grundliegenden Datenstrukturen erfordert technische Lösungen für Erfassung, Speicherung, Zugriff, Verwaltung und Analyse der Daten.

---

[1] Siehe http://www.mpiwg-berlin.mpg.de/en/research/projects/DEPT1_458_HistMPS

Die erste Herausforderung ist es, eine auf einem flexiblem Datenmodell basierende Infrastruktur zu schaffen, die sich im weiteren Projektverlauf an sich verändernde Anforderungen anpassen lässt und zugleich ermöglichen soll, Datenbestände unterschiedlicher Provenienz zu integrieren. Diese Datenbestände umfassen Normdaten, bibliographische und biographischen Datenbanken, eine Bestandsdatenbank relevanter Archive, sowie als zusätzliche Herausforderung "digital born" Daten verschiedener Disziplinen und Wissenschaftsbereichen der Institute der Max-Planck-Gesellschaft.

Die zweite Herausforderung ist die quantitative Dimension des Forschungsprogramms zur Geschichte der MPG. Es muss eine große Menge zur Zeit noch analog vorliegender, sehr heterogener Quellen digital verfügbar gemacht werden. Der Umfang der Quellen, in der Größenordnung von 4.500 laufenden Regalmetern Akten, macht es unmöglich, diese Quellen im Zeitraum der vorgegebenen Projektdauer mit traditionellen archivfachlichen Methoden zu erschließen. Aus zeit- und arbeitsökonomischen Gründen müssen daher digitale Methoden entwickelt werden, welche die Wissenschaftler bei der Quellenauswahl und -analyse unterstützen.

In unserem Paper werden wir den Ansatz zur Modellierung der Datenstrukturen genauer diskutieren, den Umgang mit den Datenmengen beschreiben und Tools der Digital Humanities vorstellen, die die Auswertung der Datenbestände erleichtern.

Bei der Modellierung der Personendatenbank muss mit widersprüchlichen Informationen und sich verändernden Werten umgegangen werden können. Dazu gehören beispielsweise sich im Zeitverlauf ändernde Namen und Aufenthaltsorte und institutionelle Zugehörigkeiten von Personen. Zugleich müssen strittige oder unsichere Informationen über die in der Datenbasis verwalteten Objekte dokumentiert werden können. Quelle und Urheber der Informationen und Einträge müssen wissenschaftlich nachvollziehbar bleiben. Schließlich sollen unterschiedlichen Versionen eines Eintrages verwaltet werden können.

Zur Umsetzung wurde ein RDF/OWL (Resource Description Framework/Web Ontology Language) Modell für die Daten entwickelt, das angelehnt an CRM (Conceptual Reference Model)[2] auf Grundlage einer ereignisbasierten Ontologie die unterschiedlichen Datenbestände integrieren soll. Hierbei benutzen wir OWL zunächst einmal pragmatisch als optimales Hilfsmittel zur Beschreibung der Daten und Datenstrukturen und sehen es nur als sekundär an, eine formale Konsistenz zu erreichen.

Dabei muss jedoch mit dem Problem umgegangen werden, dass ein Großteil der vorliegenden externen Daten, wie etwa die GND[3] zwar als RDF vorliegen,

---

[2] http://erlangen-crm.org/

[3] http://datahub.io/dataset/dnb-gemeinsame-normdatei

aber nicht kompatibel mit CRM sind. Zur Lösung dieses Problem wird ein hybrider Ansatz analog zu EDM (Europeana Data Model) verfolgt.

Zusätzlich muss der Kontext der Datenerstellung in der Datenbasis nachvollziehbar sein. Zu diesem Zwecke sollen Named Graphs eingesetzt werden, analog zu den Vorschlägen für eine Ontologie und eine API zur Verwaltung von Versionen und Provenienzen, wie sie von Kai Eckert[4] vorgeschlagen wurden. Die konkrete Implementierung erfolgt hierbei in einem Triple Store, das mit einer angepassten API und einem Frontend in Django realisiert wird.

Im zweiten Teil des Vortrages werden wir das Vorgehen zur Digitalisierung der für das Forschungsprogramm notwendigen Quellen näher darstellen. Der Umfang der Quellen, die historisch auszuwerten sind, macht es unmöglich diese in der klassischen Art und Weise durch „Close Reading“ zu sichten. Deshalb müssen Methoden gefunden werden, die mittels computer-gestützter Instrumente es ermöglichen, zu bestimmten historischen Fragestellungen eine Vorauswahl aus dem Gesamtkonvolut der digitalisierten Quellen zu treffen, bzw. Auswertungen mit Hilfe neuer Werkzeuge der Digital Humanities direkt vorzunehmen.

So sind von den einzelnen Akten im Archiv nur die Beschriftung der Aktendeckel bekannt. Mehr als 46.000 Aktenordnerrücken wurden dazu digitalisiert und in einer Datenbank erfasst. Diese Datenbasis dient zur Vorauswahl der jetzt komplett zu digitalisierenden Akten und ermöglicht eine grobe Zuordnung des Inhaltes zu einem Themenfeld.

Dazu sollen Methoden aus der Korpuslinguistik und statistische Modelle herangezogen werden um thematische Zuordnungen in großem Maßstab automatische zu ermöglichen. Um diese Methoden zu testen wurden in einem Vorprojekt mehr als 100.000 Seiten digitalisiert und der Text automatisch mit einer OCR-Software erschlossen. Die Ergebnisse werden in unserem Vortrag diskutiert, um aufzuzeigen, wie sie auf das Gesamtprojekt, das ein Umfang einer quantitativ sehr großen Datenmenge digitalisierter Akten bewältigen muss, angewandt werden können.

Das Vorprojekt hat gezeigt, dass trotz hoher Fehlerraten im jetzigen OCR-Verfahren mittels Named Entity Recognition erhebliche Fortschritte zur Erschließung der Akten und Jahrbücher erzielt werden konnten. Erste Ergebnisse mit Topic Modeling zeigen, dass dieses dazu beitragen kann, die Dokumente näher zu klassifizieren.

Eines der konkreten Ergebnisse der Digitalisierung der Jahrbücher der MPG ist, dass mit dem Aufbau einer möglichst umfassenden Bibliographie der MPG

---

[4] Kai Eckert, „Metadata Provenance in Europeana and the Semantic Web“, Masterarbeit 2012, http://edoc.hu-berlin.de/docviews/abstract.php?id=39630

seit 1949 begonnen werden konnte, die bisher nicht elektronisch vorliegt. Dazu werden die digitalisierten und mit Texterkennung bearbeiteten Jahrbücher mittels eines eigens für das Forschungsprogramm zur Geschichte der MPG entwickelten Annotationswerkzeuges ausgezeichnet. Die aus den erstellten Annotationen resultierenden Referenzen werden in eine Datenbank übertragen. Zurzeit basiert diese Annotationsmethode noch auf einem halb-automatischen Prozess. In unserem Vortrag werden wir über die Ergebnisse der künftig vorgesehenen automatischen Auswertung berichten.

Schließlich kommen unterschiedliche Methoden zur Netzwerk- und Zitationsanalyse in dem Forschungsprogramm zur Anwendung, um die Stellung der MPG innerhalb verschiedener wissenschaftlichen Forschungsgebiete zu analysieren.[5] Die dazu eingesetzten Tools werden wir in unserem Vortrag vorstellen.

Zur Visualisierung von geo-temporalen Abhängigkeiten, wie z.B die Neugründung, Verlagerung und Aufspaltung von Instituten oder die Flüsse von Fördermitteln, wird der Geo-Browser PLATIN[6] eingesetzt, der auf der Grundlage von Stefan Jähnickes GeoTemCo[7] in Zusammenarbeit mit DARIAH und dem Exzellenzcluster TOPOI weiterentwickelt werden konnte.

---

[5] Laubichler MD, Maienschein J, Renn J. 2013. Computational Perspectives in the History of Science. Isis. 104:119-130.

[6] https://github.com/skruse/PLATIN

[7] https://github.com/stjaenicke/GeoTemCo