

MEDEA: Datenkonsistenz mittels Ontologie

Pollin, Christopher

christopher.pollin@uni-graz.at
Universität Graz, Österreich

Vogeler, Georg

georg.vogeler@uni-graz.at
Universität Graz, Österreich

Daten werden laut der Vision der 'High Level Expert Group on Scientific Data' in der Zukunft einen Grad an Ausdrucksstärke und Formen der Selbstbeschreibung erhalten, dass sie in die Lage versetzt werden, ihre eigene Infrastruktur zu stellen (Neuroth, Heike / et al. 2012). Auch die Idee des Semantic Web verspricht eine Zukunft in der Maschinen selbständig mit Daten agieren können (Berners-Lee, Tim 2000). Die praktische Realität - gerade für die digitalen Geisteswissenschaften - ist noch eine Andere. Dennoch steckt in den Methoden des Semantic Webs ein Potenzial, das es mit vernünftiger Kritik zu nutzen gilt. Das Projekt MEDEA (Modeling semantically Enriched Digital Edition of Accounts) versucht dies zu verwirklichen, indem an einem kollektiven Standard zu semantischen Anreicherung digitaler Editionen von historischen Rechnungsbüchern gearbeitet wird. Es wird der Frage nachgegangen, inwieweit Methoden des Semantic Webs bei der Erschließung, Analyse und Darstellung historischer Rechnungsbücher helfen können (MEDEA, 2017).

Ein zentrales Anliegen jedes wissenschaftlichen Projektes ist es, über qualitative und konsistente Daten zu verfügen. Dateninkonsistenz und mangelnde Datenqualität bergen die Gefahr falscher wissenschaftlicher Interpretationen und kritischer Fehlerquellen für die technische Verarbeitung der Daten. Da unterschiedliche TEI-Kodierungen von unikal Quellen zusammengeführt werden, ist es eine besondere Herausforderung im MEDEA Projekt, Workflows zu etablieren, die dieser Herausforderung begegnen. Die Entwicklung einer domänenspezifischen Ontologie zur Formalisierung von historischen Prozessen des Rechnungswesens kann als eine solche potenzielle Lösung betrachtet werden. Die (Bookkeeping Ontologie 2017) formalisiert in ihrem jetzigen Zeitpunkt eine grundlegende Wissensstruktur, um Einträge in Rechnungsbüchern, ihre Transaktionen von Gütern, Dienstleistungen oder Geldbeträgen von einem Akteur oder Konto zu einem anderen, standardisiert beschreiben zu können.

Aus jeder Transaktion, die mittels des Attributes @ana in einem TEI kodierten Text annotiert wurde, wird ein XML/RDF Datensatz erzeugt, der auf Konzepte der in OWL serialisierten Bookkeeping-Ontologie referenziert (Vogeler 2016). Der Ontologie Editor Protégé erlaubt

es, eine Ontologie und die darin enthaltenen Daten (Individuals) einem Reasoning - dem Abarbeiten aller Vorhanden Regeln in einer Ontologie auf Basis der Description Logic - zu unterziehen (Musen 2015). Das Reasoning gilt als ein essentieller Baustein im Design, der Entwicklung, der Wartung und in der praktischen Anwendung einer Ontologie. Das Ergebnis davon sind Inferenzen. Inferenzen sind neu hergeleitete Schlussfolgerungen auf Basis des Reasoning Prozesses (Dentler / et al. 2011). Die Überprüfung strukturierter Daten mittels logischen Schlussfolgerungen kann dazu dienen, größere Datenmengen auf ihre Konsistenz und somit auch auf ihre Qualität hin zu prüfen, da logische Inkonsistenzen als Fehlermeldung angezeigt werden. Die Überprüfung und Zusammenführung der TEI-Kodierungen wird im MEDEA Projekt auf Basis dieser Ontologie durchgeführt. Use Cases für die Bookkeeping-Ontologie im Projekt umfassen:

- Formalisierung und Systematisierung von Rechnungsbüchern in einer maschinenverständlichen Wissensbasis
- Überprüfung der Datenkonsistenz und inhaltliche Zusammenführung der Daten
- Schaffung eines Definitionskonsenses
- Grundlage für semantische Retrieval und Discovery Strategien
- Wiederverwendbarkeit und Erweiterbarkeit
- Interoperables und offenes, sowie transparentes Verteilen von Daten

Dieser Zugang kann mit der Arbeit von (Steffen, Henniche / et al. 2015) verglichen werden, in der die Anwendung von Semantic Web Methoden, im Speziellen des Reasoning, auf geisteswissenschaftliche Daten angewandt wird.

So verlockend die Möglichkeiten einer Ontologie sein können, so kritisch sind diese auch zu betrachten. Ein grundlegendes Problem ist bereits durch den Widerspruch der hermeneutischen Arbeit der Historikerin und der Entscheidbarkeit von OWL gegeben. Sind Ontologien in OWL ausdrucksstark genug, um geisteswissenschaftliche Daten so beschreiben zu können, dass ein logisches Schlussfolgern Ergebnisse erzielt, das die Konsistenz und die Qualität der Daten abbildet?

Bibliographie

Berners-Lee, Tim. (2000): Weaving the Web: The Past, Present and Future of the World Wide Web by Its Inventor. London.

Bookkeeping Ontologie, <http://glossa.uni-graz.at/o:medea.1951/ONTOLOGY> [letzter Zugriff 21.09.2017].

Dentler, Kathrin / et al. (2011): Comparison of reasoners for large ontologies in the OWL 2 EL profile. Semantic Web 2.2, 71-87.

MEDEA, <https://medea.hypotheses.org> [letzter Zugriff 21.09.2017].

Musen, M.A. (2015): The Protégé project: A look back and a look forward. AI Matters. Association of Computing Machinery Specific Interest Group in Artificial Intelligence, DOI: 10.1145/2557001.25757003.

Neuroth, Heike / et al. (2012): Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme. Hülsbusch.

Steffen, Henniche / et al. (2015): Reasoning with Reasoning. Using Faceted Browsers to Find Meaning in Linked Data. Berlin, 1-61, <https://lirias.kuleuven.be/handle/123456789/485851>

Vogeler, Georg (2016): The Content of Accounts and Registers in their Digital Edition. XML/TEI, Spreadsheets, and Semantic Web Technologies, in: SARNOWSKY, Jürgen (Hg.): Konzeptionelle Überlegungen zur Edition von Rechnungen und Amtsbüchern des späten Mittelalters. Göttingen, 13-41.