Semantic Web und Linked Open Data in den Geschichts-wissenschaften

Kröger, Bärbel
bkroege@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Störiko, Johanna
johanna.danielzik@stud.uni-goettingen.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Wettlaufer, Jörg
jwettla@gwdg.de
Akademie der Wissenschaften zu Göttingen, Deutschland

Seit der Veröffentlichung des ersten Konzepts eines „Semantic Web“ als Erweiterung des World Wide Web (Berners-Lee und Lassila Hendler 2001) haben sich GeisteswissenschaftlerInnen mit den Möglichkeiten und Grenzen der maschinenlesbaren Modellierung ihrer Daten im Rahmen dieses Entwurfs beschäftigt. Das Datenmodell des Resource Description Framework (RDF) und die Serialisierung in Turtle oder N-Triples hat sich zum Standard in der Modellierung von maschinenlesbaren semantischen Aussagen entwickelt. Obwohl sich eine Reihe von Erwartungen aus der Entstehungszeit des Semantic Web nicht erfüllt haben (umfassende Erweiterung des WWW mit semantischen Daten, Stabilität der Uniform Resource Identifier etc.), bildet das RDF-Datenmodell heute die Grundlage verschiedener Wissensbasen (DBpedia, Wikidata) und weiterer Wissensgraphen (knowledge graphs), die zurzeit in verschiedenen Zusammenhängen entstehen. Aus diesem Grund sind das Semantic Web und die Verlinkung von offen zugänglichen Daten (Linked Open Data) für die digitalen Geisteswissenschaften weiterhin und sogar verstärkt von Interesse (Beretta 2021, Beretta & Alamercery 2020, Hiltmann & Riechert 2020, Meroño-Peñuela 2017, Meroño-Peñuela et al. 2014, Pollin 2017, Wettlaufer 2018, Wettlaufer et al. 2015).

Der ganztägige Workshop bietet eine Einführung in die Thematik „Semantic Web und Linked Open Data“ mit einem Schwerpunkt auf den Geschichtswissenschaften. Das Angebot richtet sich an Teilnehmende ohne Vorkenntnisse im Bereich Semantic Web/Linked Open Data und eignet sich für Forschende aller Fachbereiche. Didaktisch teilt sich der Workshop in vier Teile, wobei die praktische Übung etwa zwei Drittel der Zeit beansprucht.

Zu Beginn des Workshops werden die Grundlagen des Semantic Web, des Resource Description Frameworks sowie damit verbundener www-Standards im Rahmen einer einführenden Darstellung behandelt. Besondere Aufmerksamkeit kommt dabei den Themen Wikidata[1] und SPARQL[2] zu, die in den anschließenden Übungen eine wesentliche Rolle spielen. Folgende Themenblöcke sind für diesen ersten, einführenden Teil vorgesehen:

Die Idee des Semantic Web: kurze Vorstellung der Grundidee, entwickelt aus dem Grundproblem der maschinellen Verarbeitung natürlicher Sprache. Das Resource Description Framework (RDF) als Grundlage für formalisierte Aussagen. Die Bedeutung stabiler URIs für die Idee des Semantic Web. Die Turtle Serialisierung von RDF als Grundlage für die Abfragesprache SPARQL. Namespaces und ihre Bedeutung, auch im Semantic Web. RDF-Schema und Ontologien zur Formulierung komplexerer Aussagen. Linked Open Data, Knowledge Graphs und die LOD Cloud. Wikidata (und DBpedia) als zentrale Knoten der LOD Cloud. Kurzer Exkurs zu alternativen Ansätzen zur Verlinkung von Normdaten: Beacon-Dateien. Übersicht zu Ressourcen im Semantic Web und in der LOD Cloud für die Geschichtswissenschaften.

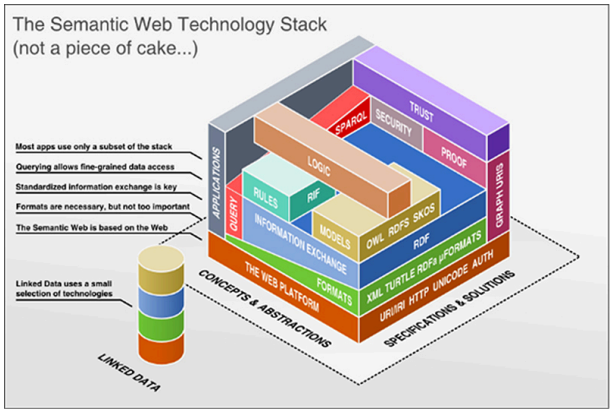

Abbildung 1. The Semantic Web Technology Stack. http://bnode.org/media/2009/07/08/semantic_web_technology_stack.png

Im zweiten Teil des Workshops sollen die Teilnehmenden in die Benutzung der Linked Open Data Plattform Wikidata eingeführt und Grundlagen für die Verwendung der Abfragesprache SPARQL gelegt werden.

Wikidata ist nicht nur der momentan größte frei verfügbare Wissensgraph, sondern bietet Daten unter freien Lizenzen und erlaubt genau wie die Wikipedia die freie Mitarbeit beim Aufbau der Wissensbasis. In der Übung lernen die Teilnehmenden somit eine der relevantesten Datenquellen für das Semantic Web kennen (Jacobsen et al. 2018). Außerdem ermöglicht Wikidata mit dem QueryService[3] einen niederschwelligen Einstieg in SPARQL, für den keine lokalen Installationen oder technischen Kenntnisse notwendig sind. Für die Übungen wird lediglich ein internetfähiges Gerät (vorzugsweise Laptop) sowie ein Wikidata-Account benötigt. Auch die graphische Benutzeroberfläche der Wikidata eignet sich gut für die Vermittlung der theoretischen Konzepte des Semantic Web, ohne dabei Kenntnisse der Informatik voraussetzen zu müssen.

Der erste Übungsblock erläutert zunächst die Datenstrukturen der Wikidata. Anhand eines Beispieleintrags (Item) werden die theoretischen Konzepte aus dem ersten Teil des Workshops in ihrer Anwendung innerhalb der

Wikidata gezeigt. Der Beispieleintrag repräsentiert ein Item, das über Properties mit weiteren Items verbunden werden kann. Durch eine solche Verknüpfung entsteht ein Statement. Dieses kann wiederum durch sogenannte Qualifier näher beschrieben werden. Qualifier sind für die Nutzung der Wikidata in der Geschichtswissenschaft besonders relevant. Sie ermöglichen unter anderem die Modellierung der Herkunft einer Information. Als „Referenz“ können so Internetressourcen verlinkt werden, aus denen die Information entnommen wurde. Qualifier wie „Startzeitpunkt“ und „Endzeitpunkt“ erlauben die Spezifikation eines Zeitraumes, innerhalb dessen eine Information gültig ist. Neben dem Aufbau der Wikidata-Items behandelt dieser Teil des Workshops auch eine Einführung in das Benennungssystem der Wikidata und die verschiedenen RDF-Namespaces, die dort verwendet werden. Schließlich kann ein Item der Wikidata mit mehreren Labels ausgezeichnet werden, die eine Benennung und Beschreibung in unterschiedlichen Sprachen ermöglichen.

Eingeübt wird der Umgang mit Wikidata anschließend anhand von Datensätzen aus dem Forschungsprojekt Germania Sacra[4], welches sich mit der Erforschung kirchlicher Institutionen und Personen des Mittelalters und der Frühen Neuzeit beschäftigt. Für den Workshop werden aufbereitete Forschungsdaten zur Verfügung gestellt, welche die Teilnehmenden selbstständig in die Wikidata einpflegen können. So ergänzen sie bestehende Datensätze zu Bischöfen des Alten Reiches um weitere Informationen wie ihren Begräbnisort. Die händische Eingabe der Daten vertieft das Verständnis für die Datenstrukturen, ist mit größeren Datenmengen aber nicht praktikabel. Als Ausblick auf den Einsatz von Wikidata im Forschungsalltag werden daher mit den Tools „Quickstatements“[5] und „OpenRefine“[6] Möglichkeiten zur seriellen Eingabe von größeren Datenmengen vorgestellt.

Der nächste Block der Übung behandelt die Grundlagen der Abfragesprache SPARQL. Ziel ist es, dass die Teilnehmenden ein Verständnis dafür entwickeln, wie geisteswissenschaftliche Fragestellungen als Abfrage formuliert und mit SPARQL auf Wikidata umgesetzt werden können. Dafür muss zunächst recherchiert werden, wie die in der Fragestellung enthaltenen geisteswissenschaftlichen Konzepte in der Wikidata modelliert sind. Anschließend wird eine passende Abfrage formuliert. Diese folgt mit SELECT, WHERE und gegebenenfalls OPTIONAL immer derselben Grundstruktur, die um weitere, komplexere Befehle ergänzt werden kann (siehe Abbildung 2). Diese Grundbausteine von SPARQL werden zunächst mit einfachen Abfragen wie „Finden Sie alle Datensätze zu Bischöfen mit einer WIAG-Kennung“ geübt. Vertiefend behandelt die Übung dann das Verketten von Abfragemustern und das Abfragen von Labels aus der Wikidata. Diese Grundlagen der Abfragesprache SPARQL werden durch Rückbezüge zum theoretischen Teil des Workshops auch mit den formalen Grundlagen des Semantic Web verknüpft. Während der Übung wechseln sich Demonstrationen neuer Konzepte und die Bearbeitung von aufeinander aufbauenden Übungsaufgaben ab. Während der Übungen sind die Teilnehmenden dazu eingeladen, sich mit anderen auszutauschen, Ergebnisse zu vergleichen und Fragen zu stellen. So erarbeiten sie sich Schritt für Schritt die Abfrage der im ersten Teil der Übung eingepflegten Datensätze.

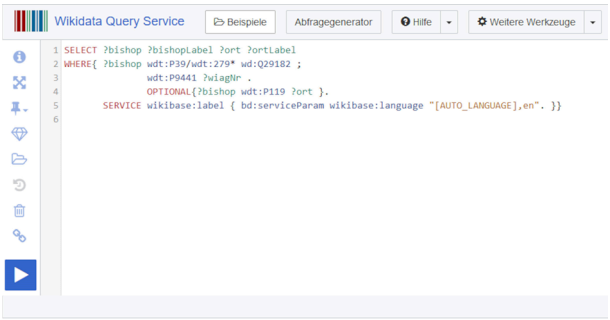

Abbildung 2. SPARQL Abfrage im Wikidata Query-Service. Abrufbar unter https://w.wiki/5FuN

Das technische Framework, in das die Wikidata eingebettet ist, bietet den Nutzenden frei verfügbare Tools, mit denen abgefragte Daten graphisch dargestellt werden können. Dazu zählen ein interaktiver Graph, ein Zeitstrahl, eine Karte, eine Bildergalerie und viele andere Visualisierungsmöglichkeiten. Diese Tools werden zum Abschluss des praktischen Teiles vorgestellt und ausprobiert. Als Ergebnis dieses Hands on Übungsteils entsteht eine Visualisierung der Daten, die die Teilnehmer zu Beginn des Workshops in die Wikidata eingepflegt und mit den erworbenen SPARQL-Kenntnissen abgefragt haben.

Im vierten und letzten Teil des Workshops soll den Teilnehmenden ein Einblick in den Datenbestand der Wikidata zu geschichtlichen Themen und in die plattformspezifische Modellierung dieser Daten vermittelt werden. Daran anknüpfend soll ein Blick auf die Potenziale von wikibase-basierten Wissensgraphen für die Geschichtswissenschaften geworfen werden. Die Vor- und Nachteile einer Datensammlung, die in einem kollaborativen Prozess entsteht, werden diskutiert. Dabei gilt ein kritischer Blick den Fragen der Datenqualität, der Datenmodellierung und der Vollständigkeit der Daten.

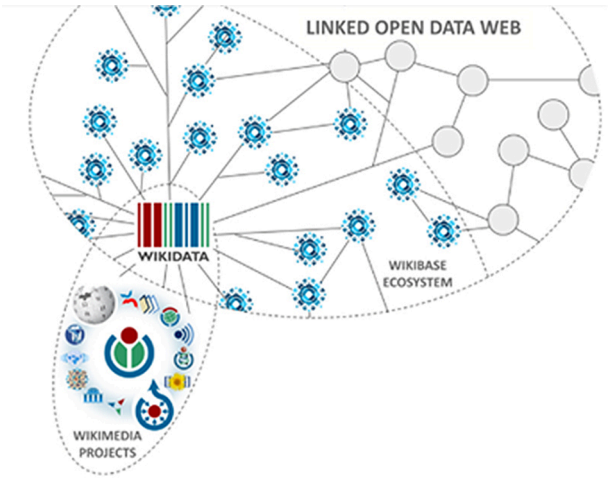

Abbildung 3. Darstellung des Linked Open Data Webs der Wikimedia (Quelle: https://meta.wikimedia.org/wiki/LinkedOpenData/Strategy2021/Joint_Vision)

Auch Alternativen zur Wikidata werden in diesem Teil des Workshops thematisiert. Wikibase[7], das technische Framework, das der Wikidata zugrundeliegt, kann auch

als eigenständige, von Wikidata unabhängige Instanz verwendet werden. Diese Lösung hat das Potenzial, die Vorteile des Systems zur Verlinkung von Daten zu nutzen und gleichzeitig eine unabhängige und durch die Forschenden selbst kuratierte Datensammlung aufzubauen. Hierfür gibt es in den Digital Humanities einige Anwendungsbeispiele, die kurz vorgestellt werden.

Als Grundlage für eine anschließende praxisorientierte Betrachtung eigenständiger Wikibase-Instanzen dienen kleinere Pilotprojekte, die unter anderem im Rahmen von Lehrveranstaltungen an der Universität Göttingen realisiert wurden. Es wurde nicht nur die Wikidata, sondern auch die vom Forschungszentrum Gotha der Universität Erfurt betriebene Wikibase-Instanz FactGrid[8] mit Daten angereichert. Im Fokus standen dabei die Bischöfe des Alten Reiches, die ihre Servitienzahlungen an die päpstliche Kurie mit Hilfe des Florentiner Bankhauses der Familie Medici abwickelten. Mit den im Workshop erworbenen Kenntnissen können die Teilnehmenden diese Daten abfragen und sich einen Einblick in die konkrete Nutzung von Linked Open Data in den Geschichtswissenschaften – und auch für ihre eigene Forschungsfragen – verschaffen.

Zur Nachbereitung des Workshops werden den Teilnehmenden die Übungsaufgaben inklusive einer Musterlösung zur Verfügung gestellt. Sie erhalten außerdem in Form eines „Cheat Sheet“ eine Übersicht über alle in der Übung verwendeten Befehle.

Zielgruppe: HistorikerInnen und GeisteswissenschaftlerInnen ohne Vorkenntnisse in SWT und LOD.

Didaktisches Konzept: Einführende Vermittlung von Grundlagenwissen, Interaktive Übungen, Gruppenarbeit/Hands on Beispiele.

Erwartete Teilnehmerzahl: 5-25

Technische Austattung: Seminarraum mit HD Beamer

Vortragende:
Bärbel Kröger, M.A.
Akademie der Wissenschaften zu Göttingen
Geiststraße 10
37073 Göttingen
bkroege@gwdg.de

Bärbel Kröger arbeitet im Akademievorhaben Germania Sacra und forscht zum Einsatz von Linked Open Data im Bereich der mittelalterlichen Kirchengeschichte. Sie leitet ebenfalls das Linked Data Projekt WIAG (Wissensaggregator Mittelalter und Frühe Neuzeit).

Johanna Störiko, M. Sc.
Georg-August-Universität Göttingen
Institut für Digital Humanities
Nikolausberger Weg 23
37073 Göttingen
johanna.stoeriko@uni-goettingen.de

Johanna Störiko (geb. Danielzik) untersuchte in ihrer Masterarbeit historische Werbeanzeigen in Kulturzeitschriften der Jahrhundertwende mit digitalen Methoden. Sie interessiert sich für den Einsatz von Technologien des Semantic Web und Linked Open Data in den digitalen Geschichtswissenschaften.

Dr. Jörg Wettlaufer
Koordination Digitalisierung und Datenkuration | Digitale Akademie
Akademie der Wissenschaften zu Göttingen
Theaterstraße 7
37073 Göttingen
Germany
jwettla@gwdg.de

Jörg Wettlaufer leitet die Digitale Akademie der Wissenschaften zu Göttingen und forscht zu Themen der Digitalen Geschichtswissenschaft, insbesondere dem Einsatz Semantic Web Technologien in den Digitalen Geisteswissenschaften sowie zur Rechts- und Sozialgeschichte.

Geplanter Ablauf des Workshops:

9:00	Vorstellungsrunde und Einführung in die Veranstaltung (30 min.)
9:30	Teil 1: Einführung Grundlagen des Semantic Web und LOD (60 min.)
10:30	Kaffeepause
11:00	Teil 2: Übung mit Wikidata anhand von Beispielen (90 min.)
12:30	Mittagspause
13:30	Teil 3: Übung SPARQL auf Wikidata (90 min.)
15:00	Pause
15:30	Teil 4: Beispiele für den Einsatz von Wikidata und LOD in den Geschichtswissenschaften (90 min.)
17:00	Ende

Fußnoten

1. https://www.wikidata.org/
2. https://www.w3.org/TR/sparql11-query/
3. https://query.wikidata.org/
4. http://www.germania-sacra.de
5. https://quickstatements.toolforge.org
6. https://openrefine.org/
7. https://www.wikimedia.de/projects/wikibase
8. https://database.factgrid.de

Bibliographie

Beretta, Francesco. 2021. “A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME)”, Semantic Web 12: 2, Special issue on Semantic Web for Cultural Heritage. https://doi.org/10.3233/SW-200416

Beretta, Francesco and Vincent **Alamercery**. 2020. "Du projet symogih.org au consortium Data for History - La modélisation collaborative de l’information au service de la production de données géo-historiques et de l’interoperabilité dans le web sémantique." *Revue ouverte d’ingénierie des systèmes d’information* 1(3):1-15. https://doi.org/10.21494/ISTE.OP.2020.0532

Berners-Lee, Tim, James **Hendler** and Ora **Lassila**. 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American* 284(5), 34-43.

Hiltmann, Torsten and Thomas **Riechert**. 2020. “Digital Heraldry. The State of the Art and New Approaches Based on Semantic Web Technologies.” In *L’édition en ligne de documents d’archives médiévaux*, ed. by Christelle Balouzat-Loubet, Turnhout, 102-125.

Jacobsen, Annika et al. 2018. “Wikidata as an intuitive resource towards semantic data modeling in data FAIRification.” In *Semantic Web Applications and Tools for Health Care and Life Sciences. Proceedings of the 11th International Conference Semantic Web Applications and*

Tools for Life Sciences (SWAT4HCLS 2018). Ed. by Christoper J.O. Baker, CEUR workshop proceedings Vol. 2275. http://ceur-ws.org/Vol-2275/

Meroño-Peñuela, Albert, Ashkan **Ashkpour**, Marieke **van Erp**, Kees **Mandemakers**, Leen **Breure**, Andrea **Scharnhorst**, Stefan **Schlobach** and Frank **van Harmelen**. 2015. "Semantic Technologies for Historical Research: A Survey." *Semantic Web Journal* 6: 539-564.

Meroño-Peñuela, Alberto. 2017. "Digital Humanities on the Semantic Web: Accessing Historical and Musical Linked Data," *Journal of Catalan Intellectual History (JOCIH)* 1(11): 144-149. DOI: 10.1515/jocih-2016-0013

Pollin, Christopher and Georg **Vogeler**. 2017. "Semantically Enriched Historical Data. Drawing on the Example of the Digital Edition of the 'Urfehdebücher der Stadt Basel'", In 2 nd Workshop on Humanities in the Semantic Web (WHiSe), ed. by A. Adamou, E. Daga and L. Isaksen, 27-32.

Wettlaufer, Jörg. 2018. "Der nächste Schritt? Digitale Editionen und Semantic Web." In *Zeitschrift für Digitale Geisteswissenschaften*, Sonderheft "Digitale Metamorphosen", Hg. von Roland S. Kamzelak und Timo Steyer (= Sonderband der Zeitschrift für digitale Geisteswissenschaften, 2). DOI: 10.17175/sb002_007

Wettlaufer, Jörg, Christopher **Johnson**, Martin **Scholz**, Mark **Fichtner**, Sree Ganesh **Thotempudi**. 2015. "Semantic Blumenbach: Exploration of Text-Object Relationships with Semantic Web Technology in the History of Science," Digital Scholarship in the Humanities (DSH), Special Issue 'Digital Humanities 2014'; ed. by Melissa Terras, Claire Clivaz, Deb Verhoeven and Frederic Kaplan, 30 Supplement 1: i187-i198 https://academic.oup.com/dsh/article/30/suppl_1/i187/364720/