# Geisteswissenschaftliche Fachdatenrepositorien im Semantic Web. Modellierung, Vernetzung, Visualisierung.

**Schrade, Torsten**
Torsten.Schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Deutschland

## Implizite und explizite Semantik TEI-basierter Fachdatenrepositorien

Zahlreiche geisteswissenschaftliche Fachdatenrepositorien setzen zur Modellierung ihrer Forschungsdaten auf die Richtlinien der Text Encoding Initiative (TEI) und somit auf XML als primäres Datenformat. XML eignet sich sehr gut zur Lösung editorisch-philologischer Aufgabenstellungen und entspricht den geforderten Kriterien der Interoperabilität und Nachhaltigkeit von Forschungsdaten. Durch die standardkonforme Auszeichnung der Forschungsgegenstände in TEI werden diese formal und inhaltlich erschlossen. TEI-kodierte Daten beinhalten in jeder Hinsicht semantische Bezüge (bspw. Raumbezüge, Personenbezüge, begriffliche und konzeptuelle Bezüge etc.). Aus der Perspektive des *Semantic Web* sind diese Bezüge jedoch zunächst nur implizit und nicht explizit in den Daten vorhanden (Abbildung 1). Im Gegenzug gründen sich *Semantic Web*-Technologien auf das *Resource Description Framework* (RDF) zur Formulierung semantischer Aussagen (*statements*) in Form von Subjekt – Prädikat – Objektbeziehungen (*triples*). Die besondere Stärke von RDF liegt in der automatisiert möglichen Vernetzung (*interlinking*), Zusammenführung (*merging*) und Analyse (*reasoning*) eigentlich separater Datenbestände. RDF ist modellierungstechnisch auf einer höheren Abstraktionsebene anzusiedeln als TEI-kodierte XML-Daten (Abbildung 2; vgl. auch Polleres u. a. 2009).

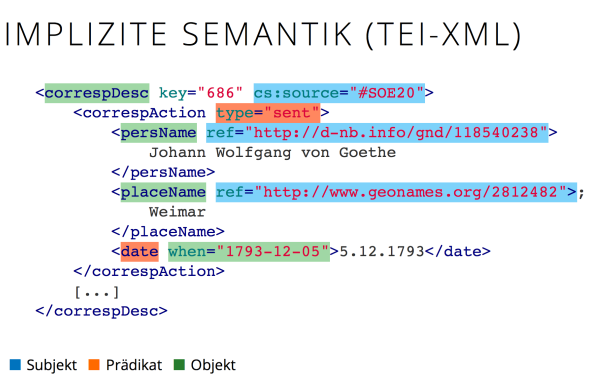


***Abb. 1***

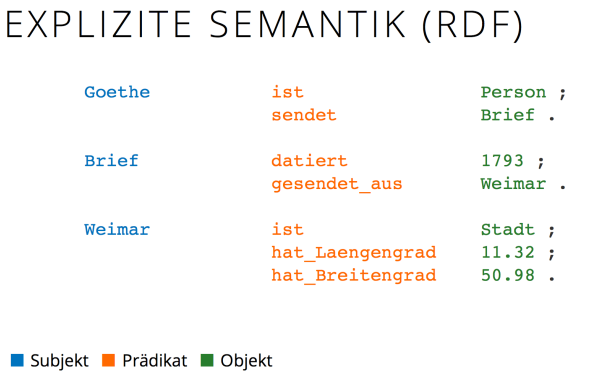


***Abb. 2***

Während die formale Erschließung geisteswissenschaftlicher Forschungsgegenstände mittels XML-basierter Annotationsmethoden mittlerweile als weit fortgeschritten gelten kann, bleibt die semantische Erschließung häufig noch weit zurück. Zwar wird in den Daten oft das Auftreten bestimmter Ortsnamen, Personennamen, Werktitel etc. annotiert. Dennoch gehen diese Annotationen meist nicht darüber hinaus, anzuzeigen, dass eine bestimmte Entität an einer spezifischen Stelle erwähnt ist. Damit bleibt die Semantik der Fachdaten weit hinter den Möglichkeiten zurück, die aktuelle Technologien – insbesondere die des *Semantic Web* und des *Linked Open Data* (LOD) – bieten könnten. Der durch LOD mögliche Zugang auf vernetzte Forschungsdaten eröffnet neuartige Perspektiven der Nutzung bisher isoliert stehender Fachdatenrepositorien. Wesentlich ist dabei, dass LOD und RDF bestehende Standards der *Digital Humanities* (wie bspw. TEI / XML) um die Verwendung gemeinsamer Terminologien und Metadatenschemata erweitern (vgl. Iglesia et al. 2015). Dadurch wird es möglich, auch Bestände verteilter Provenienz und unterschiedlicher Struktur gemeinsam inhaltlich zu beschreiben und zu analysieren.

Insgesamt existiert momentan also noch eine Kluft: Auf der einen Seite die zahlreichen geisteswissenschaftlichen Fachdatenrepositorien mit implizit semantischem

Potential, auf der anderen Seite die Technologien und Datenmodelle des *Semantic Web*, die neue Sichten und Analysemethoden auf die Daten eröffnen könnten. Zwar existieren einige Sprachkonzepte, Methoden und Tools zur Übersetzung zwischen TEI / XML und RDF. Diese sind jedoch ausnahmslos komplex, teilweise technisch veraltet, verfügen nur über prototypische Implementierungen oder sind hochgradig spezialisiert auf einen bestimmten Datenbestand. Während die dem *Semantic Web* zugrunde liegenden Technologien aus informatischer Sicht als erschlossen und anwendbar angesehen werden können (vgl. Lanthaler 2014: 11–35), besteht zum jetzigen Zeitpunkt also ein Bedarf an exemplarischen Bearbeitungen repräsentativer Forschungsdatenbestände aus den Geisteswissenschaften, um die Tragfähigkeit dieser Technologien auch für die geisteswissenschaftliche Forschung zu demonstrieren.

# Semantische Aussagen aus XML mit Hilfe des *XTriples*-Webservices

An dieser Stelle setzt der *XTriples* -Webservice der *Digitalen Akademie* der Mainzer Akademie der Wissenschaften und der Literatur an. Grundgedanke des generischen Dienstes ist das Crawling beliebiger XML-Datenbestände und die anschließende Generierung semantischer Aussagen aus den XML-Daten auf Basis definierter Aussagemuster. Das Prinzip der Explizierung semantischer Aussagen aus XML ist dabei nicht sonderlich komplex: Wird die URI einer XML-Ressource oder eine Dateneinheit in dieser Ressource als das Subjekt einer semantischen Aussage begriffen, können diesem Subjekt über Prädikate aus kontrollierten Vokabularen weitere Werte aus den XML-Daten bzw. URIs zu weiteren Datenressourcen als Objekte zugeordnet werden. Im Übersetzungsvorgang zwischen XML und RDF geht es also vor allem um die Bestimmung semantischer Aussagemuster, die sich gesamthaft auf alle Ressourcen eines XML-Datenbestandes anwenden lassen.

Die Aussagemuster werden in Form einer einfachen, XPATH-basierten Konfiguration an den Dienst übermittelt. Dabei ist es auch möglich, über die Bestände eines spezifischen XML-Repositoriums hinauszugehen und externe Ressourcen oder Dateneinheiten in die Transformation mit einzubeziehen (bspw. aus der GND, der *Dbpedia*, aus *Geonames* u. a *)*. Die technische Realisierung als Webservice hat den Vorteil, dass AnwenderInnen keine weitere Software zur semantischen Übersetzung von Forschungsdaten benötigen. Gleichzeitig kann der Webservice auch als eine Art „externe“ RDF-Schnittstelle (im Sinne eines Proxy) für ein oder mehrere XML-Repositorien eingesetzt werden. Grundvoraussetzung hierfür ist lediglich, dass die jeweiligen Repositorien über HTTP erreichbar sein müssen.

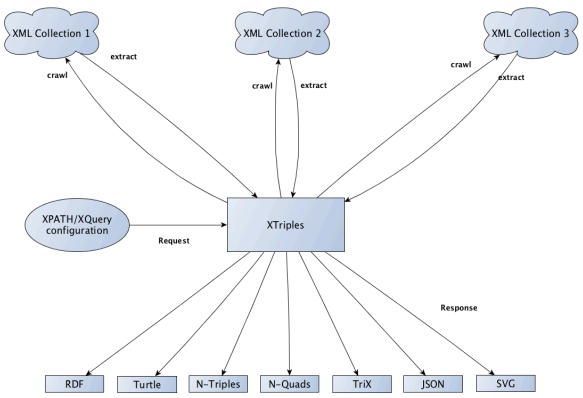


***Abb. 3***

Das Ergebnis einer *XTriples*-Extraktion steht in einer Vielzahl gängiger RDF-Serialisierungen zur Verfügung (Abbildung 3). Neben rein RDF-basierten Formaten ist es auch möglich, die semantischen Bezüge eines Repositoriums mittels SVG darzustellen oder das Extraktionsergebnis zur weiteren Analyse und Visualisierung an *Semantic Web*-Tools weiterzureichen.

# Anwendungsbeispiele

*XTriples* wurde vom Autor im Kontext des Akademievorhabens *Deutsche Inschriften Online* in Verbindung mit dem BMBF-Projekt *Inschriften im Bezugssystem des Raumes* entwickelt und steht der DH-Community in einer stabilen Version unter Open Source Lizenz (MIT) zur Verfügung. Das zugrunde liegende Softwarepaket ist vollständig dokumentiert und auf GitHub veröffentlicht.

Neben den *Deutschen Inschriften* wird *XTriples* aktuell auch in den Akademievorhaben *Regesta Imperii* und *Die Schule von Salamanca* verwendet. Das Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte nutzt *XTriples* für eine CIDOC-CRM basierte, semantischen Modellierung von EpiDoc-Daten im Rahmen des BMBF-Projektes *Relationen im Raum*. Gemeinsam mit der Berlin-Brandenburgischen Akademie der Wissenschaften wird gerade eine Schnittstelle zwischen *XTriples* und *correspSearch* , dem Webservice der BBAW zur dezentralen Aggregation digitaler Briefeditionen, implementiert.

Folgende Beispiele geben einen ersten Überblick über die unterschiedlichen Anwendungsgebiete von *XTriples*:

- Semantische Extraktion und nachfolgende Visualisierung von Familienbeziehungen aus dem *Epidat* Grabstein Corpus für den jüdischen Friedhof in Hamburg-Altona (Ausgangsdaten EpiDoc/TEI): http://xtriples.spatialhumanities.de/examples/dh/epidat/index.html
- Semantische Extraktion und SVG-Visualisierung eines Briefnetzwerks (Teilbestand der Korrespondenz

Goethes aus den in *correspSearch* aggregierten CMI-Daten bei gleichzeitiger *on-the-fly* Einbeziehung der RDF-Schnittstellen von *GND* und *Geonames*): http://bit.ly/1LKK1dv

- Beispielhafte semantische Extraktion und Visualisierung europäischer Kommunikationsnetzwerke auf Basis der *correspSearch* TEI / CMI-Daten: http://metacontext.github.io/presentation-correspsearch-xtriples/viz/map.html

Weitere Beispiele zu den einzelnen Funktionalitäten finden sich auf der *XTriples*-Website unter http://xtriples.spatialhumanities.de/examples.html . Einen schnellen Überblick über die Funktionsweise des Dienstes gibt folgende Präsentation: http://metacontext.github.io/presentation-correspsearch-xtriples .

Ziel des Vortrags ist eine Veranschaulichung der Methoden und Potentiale, die sich aus der semantischen Extraktion, Modellierung, Vernetzung und Visualisierung XML-basierter, geisteswissenschaftlicher Fachdaten ergeben. Neben einer Darstellung der technischen Hintergründe des *XTriples*-Webservices werden auch Fragen der semantischen Modellierung geisteswissenschaftlicher Fachdaten mittels bestimmter Ontologien (bspw. FOAF, CIDOC-CRM u.a.) in den Blick genommen. Weiterhin werden auch beispielhafte Analyse- und Visualisierungsmöglichkeiten für semantisch modellierte geisteswissenschaftliche Fachdaten vorgestellt.

# Fußnoten

1. Projekte wie bspw. SPQR oder das Textual Encoding Framework sind veraltet oder technisch nicht generalisiert. Einen interessanten Ansatz bietet die XSPARQL Language Specification des DERI, die 2009 in Form einer W3C Member Submission niedergelegt wurde. Hier fehlen jedoch praktische Implementierungen. Die Benutzung von RDFa innerhalb von XML-Daten stellt eine weitere Möglichkeit dar, doch verfolgen die wenigsten geisteswissenschaftlichen Fachdatenrepositorien eine so ausgerichtete semantische Markup-Strategie. Auch das bereits 2007 in Form einer W3C Recommendation grundgelegte GRDDL-Framework (Gleaning Resource Descriptions from Dialects of Languages) ist bis heute eine theoretische Spezifikation geblieben. Der *OxGarage* Transformations-Webservice der *Text Encoding Initiative* bietet zwar eine Routine für die Konvertierung von TEI kodierten Daten nach RDF an, legt sich für die Transformation aber auf das CIDOC-CRM als Ontologie fest. Mit *OxGarage* können *out-of-the-box* also keine anderen Ontologien für eine semantische Modellierung benutzt werden. Zudem ist der Webservice nicht darauf ausgelegt, auch weitere, externe Datenrepositorien in eine Transformation mit einzubeziehen oder andere RDF-Serialisierungen jenseits von RDF/XML zurückzugeben.

2. Beispielsweise an den RDF zu SVG Transformations-Webservice oder an die RDF Visualisierungsbibliothek d3sparql .

# Bibliographie

**Akademie der Wissenschaften und der Literatur Mainz** (o. J.a): *Digitale Akademie* http://www.digitale-akademie.de [letzter Zugriff 16. Februar 2016].

**Akademie der Wissenschaften und der Literatur Mainz** (o. J.b): *Regesta Imperii* http://www.regesta-imperii.de/startseite.html [letzter Zugriff 16. Februar 2016].

**BBAW** (o. J.): *Berlin-Brandenburgische Akademie der Wissenschaften* http://www.bbaw.de/ [letzter Zugriff 16. Februar 2016].

**correspSearch** (o. J.): *correspSearch*. Search diverse letter editions http://correspsearch.bbaw.de/index.xql [letzter Zugriff 16. Februar 2016].

**Deutsche Inschriften Online** (o. J.): http://www.inschriften.net [letzter Zugriff 16. Februar 2016].

**Haft, Michael** (2013): "RDF als Verknüpfungsmethode zwischen geisteswissenschaftlichen Forschungsdaten und Geometrien am Beispiel des Projektes 'Inschriften im Bezugssystem des Raumes'",in: *Skriptum* 2,3 http://nbn-resolving.de/urn:nbn:de:0289-2013120622 [letzter Zugriff 14. Oktober 2015].

**i3Mainz / Akademie Mainz** (o. J.): *Inschriften im Bezugssystem des Raumes* http://www.spatialhumanities.de/ibr/startseite.html [letzter Zugriff 16. Februar 2016].

**IBR (Inscriptions in their spatial context) / Academy of Sciences and Literature, Mainz / Institute for Spatial Information and Surveying Technology i3Mainz** (o. J.): *XTriples* http://xtriples.spatialhumanities.de/index.html [letzter Zugriff 16. Februar 2016].

**Iglesia, Martin de la / Moretto, Nicolas / Brodhun, Maximilian** (2015): "Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten." In: Neuroth, Heike / Rapp, Andrea / Söring, Sibylle (eds.): *TextGrid: Von der Community – für die Community*. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften. Göttingen: Universitätsverlag Göttingen 91–102 http://dx.doi.org/10.3249/webdoc-3947 [letzter Zugriff 14. Oktober 2015].

**Lange, Felix, Martin Unold** (2015): Semantisch angereicherte 3D-Messdaten von Kirchenräumen als Quellen für die geschichtswissenschaftliche Forschung, in: *Zeitschrift für digitale Geisteswissenschaften* 1 http://dx.doi.org/10.17175/sb001_015 [letzter Zugriff 14. Oktober 2015].

**Lanthaler, Markus** (2014): *Third Generation Web APIs*. Bridging the Gap between REST and Linked Data. Diss. Institute of Information Systems and Computer Media. Technische Universität Graz http://www.markus-lanthaler.com/research/third-

generation-web-apis-bridging-the-gap-between-rest-and-linked-data.pdf [letzter Zugriff 14. Oktober 2015].

**Polleres, Axel u.a.** (2009): *XSPARQL Language Specification* http://www.w3.org/Submission/xsparql-language-specification letzter Zugriff 14. Oktober 2015].

**Salomon Ludwig Steinheim-Institut für deutsch-jüdische Geschichte** (o. J.): http://www.steinheim-institut.de [letzter Zugriff 16. Februar 2016].

**Schrade, Torsten** (2013): "Datenstrukturierung", in: *Über die Praxis des kulturwissenschaftlichen Arbeitens*. Ein Handwörterbuch. Bielefeld: transcript 91–97.