Graphentechnologien in den Digital Humanities: Methoden und Instrumente zur Modellierung, Transformation, Annotation und Analyse

Jarosch, Julian
julian.jarosch@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Kuczera, Andreas
andreas.kuczera@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Schrade, Torsten
torsten.schrade@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Yousef, Tariq
tariq.yousef@adwmainz.de
Akademie der Wissenschaften und der Literatur | Mainz, Deutschland

Text und Graph

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. Durch die standardkonforme Auszeichnung der Forschungsgegenstände in TEI werden diese formal und inhaltlich erschlossen. TEI-kodierte Daten beinhalten mannigfaltige semantische Bezüge – aus der Perspektive einer Graphmodellierung sind diese Bezüge jedoch zunächst nur implizit und nicht explizit in den Daten vorhanden (Schrade 2013).

Während die formale Erschließung geisteswissenschaftlicher Forschungsgegenstände mittels XML-basierter Annotationsmethoden mittlerweile als weit fortgeschritten gelten kann, kann die semantische Erschließung häufig noch verdichtet werden. 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 Vernetzung der Fachdaten hinter den Möglichkeiten zurück, die Graphentechnologien bieten (Iglesia u.a. 2015; Grüntgens/Schrade 2016).

Graphentechnologien sind hervorragend für die Modellierung, Speicherung und Analyse semantisch vernetzter Daten – auch verschiedener Modalitäten – geeignet. Einerseits sind in Graphen modellierte Daten hinreichend genau und berechenbar, andererseits bietet die Schema- und Hierarchiefreiheit dieser Datenstrukturierung eine ausreichend große Flexibilität zur Erfassung auch komplexer geisteswissenschaftlicher Sachverhalte (Kuczera 2017).

Eine gegenseitig ausschließende oder separate Modellierung von Forschungsgegenständen *entweder* in klassischen Strukturen – linearem Text (etwa in TEI-XML kodiert), hierarchischer Baumstruktur (Ontologien) – *oder* als Graph ist nicht mehr zwangsläufig. Möglich ist auch die Verknüpfung beider Technologien, so dass das geeignetste Datenmodell für jeden Aspekt der Daten zur Anwendung kommt. Dies erlaubt die Synthese und zusammenfassende Analyse verschiedener Daten- und Objekttypen.

Darüber hinaus werden die Grenzen zwischen den Technologien zunehmend überbrückt: Die Extraktion von Graphstrukturen aus (annotierten) Texten ist ebenso möglich wie die Modellierung von annotiertem Text als Graph in Form von *standoff property markup*.

Es breitet sich also ein Spektrum an Möglichkeiten aus: Von der Ableitung von (ephemeren) Graphen aus (führenden) XML-Texten, über die verlustlose Migration von XML zu Graph, bis zu *text-as-a-graph* als führendes Datenformat mit geeigneten Editionsumgebungen.

Die Gangbarkeit jeder dieser Möglichkeiten und ihre Unterstützung durch flexible Werkzeuge soll in diesem Workshop nachvollziehbar gemacht werden, sowie die Grundlagen zur eigenständigen Anwendung gelegt werden.

Werkzeuge

Zur Abdeckung des oben dargestellten Spektrums stellen wir vier Werkzeuge vor, die in der Digitalen Akademie der Akademie der Wissenschaften und der Literatur entwickelt werden. Teilweise eng aufeinander aufbauend, bieten sie einen aktuellen Werkzeugkasten der Graphentechnologien.

XTriples

XTriples (http://xtriples.spatialhumanities.de) ist ein Webservice zur Extraktion von RDF-Statements aus XML-Daten zur Vernetzung von Ressourcen im *semantic web*. Dieses Werkzeug ist insbesondere geeignet zur (einmaligen oder wiederkehrenden) Ableitung von RDF-Graphen aus XML-Daten (*RDF-Lifting*).

Grundfunktion des generischen Dienstes ist das Crawling beliebiger XML-Datenbestände und die anschließende Generierung semantischer Aussagen aus den XML-Daten auf Basis definierter Aussagemuster. Wird eine Dateneinheit in einer 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.

eXGraphs

eXGraphs ist der Ausbau des Grundprinzips von XTriples zur Extraktion von *property graphs*, d.h. Graphstrukturen, die über die Subjekt-Prädikat-Objekt-Tripelstruktur von RDF hinausgehen. Der Dienst ist so weit generalisiert, dass grundsätzlich keine Einschränkungen der Komplexität der extrahierten Graphen bestehen.

eXGraphs basiert auf der Graphdatenbank neo4j, das heißt es importiert entweder die gewonnenen Graphen direkt in eine spezifizierte Datenbank, oder oder gibt sie als Cypher-Abfrage zurück. Das Tool ist somit geeignet, Datenbestände von XML in gerichtete Property-Graphen zu migrieren oder wiederkehrend zur Aktualisierung der Graph-Datenbank aufgerufen zu werden. Die Konfiguration der Extraktion und Transformation wird in einer unkomplizierten XML-Konfiguration spezifiziert, deren hierarchische Struktur direkt mit den notwendigen Extraktionsschritten korrespondiert. Die gesuchten Informationen werden mittels XPath angesteuert.

GRACE

GRACE (***gra**ph **c**ontent **e**ditor*) ist eine Web-App, die das Erstellen und Pflegen von Graphdaten in neo4j-Datenbanken über eine GUI anwenderfreundlich ermöglicht. Unterstützt ist die Suche nach bestehenden Daten, das Verknüpfen von bestehenden Knoten mittels neuer Kanten, das Bearbeiten von Knoten, und die Neuanlage von Knoten. Die Attribute (*properties*) der Knoten werden als Tabelle bzw. bei der Bearbeitung als Formular dargestellt; das Nutzererlebnis ist also durchaus einer klassischen Datenbankeingabe oder Registerpflege vergleichbar.

Gegenüber einer klassischen Sacherschließung innovativ ist, dass Kanten zur Modellierung von Beziehungen, Zusammengehörigkeiten – generell für semantische Relationen verwendet werden können. Die Flexibilität und Aussagekraft ist Querverweisen klar überlegen, da die Kanten grundsätzlich klassifiziert sind und durch Attribute weiter spezifiziert werden können. Die Darstellung und Verwaltung der Kanten ist in die Nutzeroberfläche integriert, so dass die Sacherschließung im Graphen ohne Kenntnisse von Datenbankabfragesprachen aufgebaut werden kann.

SPEEDy

SPEEDy (s*tandoff* p*roperty* edi*tor*, https://github.com/argimenes/standoff-properties-editor) ist ein Editor zur Bearbeitung von *text-as-a-graph* – sowohl zur nativen Erfassung wie auch zur Weiterpflege von Datenbeständen nach Konvertierung.

Bei *standoff properties* werden Text und Annotationen voneinander getrennt gespeichert. Im Unterschied zu Standoff XML sind die in SPEEDy verwendeten *standoff properties* resistent gegen nachträgliche Änderungen, da der Editor die Indizes nach jeder Bearbeitung neu berechnet.

Mit diesem Konzept sind überlappende und auch konkurrierende Annotationshierarchien möglich. Annotationen lassen sich auch in Layern organisieren und in SPEEDy ein- und ausblenden.

Gespeichert werden die Texte im json-Format, wobei in der json-Datei als erstes der reine Text und anschließend die verschiedenen Annotationen abgelegt sind.

Mit dem in SPEEDy realisierten Annotationskonzept mit *standoff properties* werden multiple Annotationshierarchien möglich, die perspektivisch auch den wissenschaftlichen Diskurs abbilden könnten.

Material

Für den Workshop werden Beispieldaten aus den Sozinianischen Briefwechseln[1] herangezogen, die derzeit erfasst und annotiert werden. Charakteristikum dieser Korrespondenzen ist die enge Verzahnung verschiedener Themengebiete – beispielsweise werden astronomische Beobachtungen von Person zu Person entlang akademischer und familiärer Verbindungen weiter berichtet und von Ort zu Ort weitergetragen, um politisch und theologisch interpretiert zu werden … Dieses komplexe Ineinandergreifen der Themenfelder in den Korrespondenzen erfordert eine entsprechend verzahnte Registerstruktur, die die diversen Relationen zwischen Entitäten verschiedener Art adäquat abbilden kann.

Die Brieftexte werden im TEI-Subset DTABf kodiert, das eindeutige Kodierungen und verlässliche Extraktion von Informationen ermöglicht.

Als Gegenstand der Übungen stehen die Korrespondenzmetadaten im correspDesc-Format wie auch das *named entity tagging* und die Sacherschließung im Brieftext zur Verfügung. Ersteres bildet bereits das Netzwerk von Korrespondenten und Orten ab; zweiteres gewährt Einblick in die inhaltlichen und thematischen Verknüpfungen. Die zugehörigen Register, auf die die Annotationen verweisen, werden sowohl im Ausgangs-XML-Format wie auch in Form des Graphregisters Teil der Übungsdaten sein.

Ablauf

Nach einer kurzen einführenden Standortbestimmung der Graphentechnologien in den DH und einer Vorstellung der Beispieldaten werden in den zwei Workshoptagen die vier oben genannten Werkzeuge vorgestellt. Zu jedem Tool zeigen wir Beispielkonfigurationen bzw. -anwendungen, und bieten Übungen an, die praxisnah an die Forschungsziele des datengebenden Projekts angelehnt sind. Darüber hinaus steht es den TeilnehmerInnen frei, auch mit eigenen Daten zu experimentieren.

Am ersten Workshoptag gehen wir auf die Werkzeuge XTriples und SPEEDy ein. Damit zeigen wir Optionen auf, welche ohne eine Migration von XML zum Graph zur Verfügung stehen: die Beibehaltung von XML als führendes Format, oder die native Erfassung in *text-as-a-graph*. In der Übung demonstrieren wir die Erzeugung einer RDF-Datei zur Verknüpfung von Ortserwähnungen mit einer geographischen Normdatenbank.

Am zweiten Workshoptag liegt unser Fokus auf eXGraphs und GRACE. Am Ende dieses Tages sollen die Grundprinzipien des Zusammenspiels von XML und neo4j klar geworden sein, indem Forschungsdaten zu neo4j migriert und aktualisiert werden, und dort in einer für ein breites Nutzerspektrum zugänglichen GUI bearbeitet werden.

Beschlossen werden soll der Workshop neben dem ausleitenden Resümee durch eine Feedback-Runde im Plenum insbesondere zu den neu entwickelten Werkzeugen eXGraphs und GRACE, sowie zur weiteren Entwicklung von SPEEDy.

Lernziele

Ziel des Workshops ist, einen Einblick in die Durchlässigkeit zwischen traditionellen (linearen und hierarchischen) Datenstrukturen und der Modellierung im Graphen zu bieten. Die Verwendung von Transformations-, Migrations- und Bearbeitungswerkzeugen wird praktisch vermittelt und ihre Position im DH-Ökosystem umrissen. Die interaktive Demonstration wird anwendungsnah auf reale Forschungsdaten und Auswertungsziele aufgebaut. Neben der technischen Kompetenz wird das Bewusstsein für implizit vorhandene und semantisch auswertbare Graphstrukturen in bestehenden XML-Daten geschärft.

Die Übungen werden ausgehend von einem Einstiegsniveau konzipiert, mit der Option, zu höheren Komplexitätsstufen weiterzuarbeiten oder die Methoden auf eigene Daten und Fragestellungen zu transferieren.

Zwei der vorgestellten Werkzeuge sind Neuentwicklungen, so dass dies eine der ersten Gelegenheiten zur Schulung in der Anwendung sein wird.

Zahl der TeilnehmerInnen

Maximal 20.

Technische Voraussetzungen

Die Teilnehmenden benötigen nur Laptops. Es muss im Vorfeld keine Software installiert werden.

Beitragende

Julian Jarosch

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

2007–2014 Studium der Allgemeinen Sprachwissenschaft an der Johannes Gutenberg-Universität Mainz. 2011 Auslandssemester an der Bangor University, Wales. Magisterarbeit zu »Typography and Legibility: Do Typeface, Serifs and Justification influence Reading Behaviour?«. Seit 2015 wiederum an der JGU Mainz Promotionsprojekt »Empirical Typography« an der Schnittstelle Sprachwissenschaft–Buchwissenschaft, 2015–2017 als Stipendiat der Stipendienstiftung Rheinland-Pfalz. Seit 2018 wissenschaftlicher Mitarbeiter der Digitalen Akademie im DFG-Projekt »Die sozinianischen Briefwechsel«.

Andreas Kuczera

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

1993-1998 Studium der Physik und Geschichte an der Justus-Liebig-Universität Gießen (Staatsexamen für das Lehramt an Gymnasien), 2001 Promotion »Grangie und Grundherrschaft«. Zur Wirtschaftsverfassung des Klosters Arnsburg als Stipendiat der hessischen Graduiertenförderung. 2001-2006 Mitarbeiter im DFG-Projekt Regesta Imperii Online. Von 2007–2012 leitend in der Projektverwaltung der Akademie Mainz und der Digitalen Akademie tätig. Sachverständiger der IT-

Kommission der Akademie Mainz. Seit 2015 Zuständigkeit im Bereich des Projektes Regesta Imperii.

Torsten Schrade

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

Historiker, Germanist und Anglist, Softwareentwickler und Digitaler Geisteswissenschaftler (seit 2002). Seit 2009 wissenschaftlicher Mitarbeiter der Akademie und Leiter der Digitalen Akademie.

Tariq Yousef

Akademie der Wissenschaften und der Literatur | Mainz
Geschwister-Scholl-Str. 2
55131 Mainz

Bachelor in Computer Science (Softwareentwicklung) an der AlBaath Universität (Syrien), Master in Computer Science an der Universität Leipzig. Forschungsinteressen: NLP, Datenextraktion, Datenaufbereitung, data mining, Visualisierung und Webentwicklung.

Fußnoten

1. http://www.adwmainz.de/projekte/zwischen-theologie-fruehmoderner-naturwissenschaft-und-politischer-korrespondenz-die-sozinianischen-briefwechsel/informationen.html

Bibliographie

Grüntgens, Max / Schrade, Torsten (2016): *Data repositories in the Humanities and the Semantic Web: modelling, linking, visualising*, in: **Adamou, A. / Daga, E. / Isaksen, L. (Hrsg.)** *Hrsg .): Proceedings of the 1st Workshop on Humanities in the Semantic Web (WHiSe), CEUR Workshop Proceedings* . Aachen , S. 53–64.

Iglesia, Martin de la / Moretto, Nicolas / Brodhun, Maximilian (2015): Metadaten, LOD und der Mehrwert standardisierter und vernetzter Daten. In: *Heike Neuroth, Andrea Rapp, Sibylle Söring (Hrsg.): TextGrid: Von der Community – für die Community. Eine Virtuelle Forschungsumgebung für die Geisteswissenschaften* . Göttingen, S. 91–102. DOI: http://dx.doi.org/10.3249/webdoc-3947 [Letzter Zugriff 09 . 01 .201 9]

Kuczera, Andreas (2017): Graphentechnologien in den Digitalen Geisteswissenschaften. abitech 37, S. 179 –196 . https://doi.org/10.1515/abitech-2017-0042 [Letzter Zugriff 09.01.2019]

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