LOD entzaubert: was semantische Technologien wirklich können

Matej Ďurčo, Austrian Centre for Digital Humanities

Obwohl der berüchtigte LOD-Cloud[1] unaufhaltsam wächst, und die semantischen Technologien auch schon im kommerziellen Bereich[2] längst angekommen sind, scheint in der (geisteswissenschaftlichen) Forschung der Durchbruch noch auszustehen. Andererseits häufen sich die Anzeichen und die Wende naht. Aber lohnt es sich überhaupt, sich mit diesem Hype zu beschäftigen? Und ist es noch ein Hype?

LOD löst alle unsere Probleme

Wer das Semantic Web und LOD als die (Er)lösung herbeisehnt, wird enttäuscht sein.

LOD ist zwar ein genauso einfacher, wie genialer Ansatz, aber nüchtern betrachtet ist das zugrundeliegende RDF[3] ein Datenaustauschformat und ein radikal einfaches obendrein.

RDF entspricht ungefähr der Feststellung: „Meine Sprache besteht aus Wörter die nach bestimmten Regeln zu Sätzen zusammengesetzt sind“. Wobei diese Regeln bei RDF ausgesprochen einfach sind: Ein Satz besteht aus genau 3 Worten, dem „Triple“: Subjekt Prädikat Objekt. Und der Wortschatz besteht praktisch nur aus frei wählbaren Eigennamen.

Im Grunde ist RDF nur eine Abstraktionsebene unter der Zen-Feststellung „Alles ist eins“ angesiedelt: Es gibt Ressourcen und die haben Eigenschaften oder Beziehungen zu anderen Ressourcen. Eine Ressource kann ein digitales Dokument oder ein Bild sein, aber auch eine Person, eine Stadt, wenn man will auch ein Baum oder eine Körperzelle. Es kann eine identifizierbare eindeutige Entität sein, aber auch eine Klasse von Entitäten, wie z.B. die Klasse aller Dinge *owl:Thing*[4].

Das wohl entscheidende Merkmal der semantischen Technologien ist die globale Referenzierung, wovon auch immer. Dies ist auch die erste Regel des LOD-Paradigmas, vorgeschlagen 2006 von Sir Tim Berners-Lee: “Use URIs as names for things”[5], bzw. noch schärfer formuliert: „If it doesn't use the universal URI set of symbols, we don't call it Semantic Web.”

Was ist dann eigentlich das Beschreibungsvokabular?

Obwohl man also im Prinzip absolute Freiheit bzgl. der „Modellierung“, Strukturierung der eigenen Daten hat, ist es ratsam sich an existierende Vokabularien zu orientieren. Grundsätzlich ist es möglich, beliebige vorhandene Vokabularien miteinander zu kombinieren und auch mit eigenem definierten Vokabular zu ergänzen. Die große Freiheit birgt allerdings die große Gefahr in sich, dass ein unüberschaubarer, inkonsistenter Datenhaufen entsteht.

So wie die Schemas bei XML wurde bei RDF eine Reihe von „Vokabularen“, Schemas, Ontologien für spezifische Bereiche definiert. Einige Verzeichnisse helfen dabei, eine Übersicht in dem Dschungel zu

---

[1] http://lod-cloud.net/
[2] http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html
[3] http://www.w3.org/RDF/
[4] http://www.w3.org/TR/owl-ref/
[5] http://www.w3.org/DesignIssues/LinkedData.html

gewinnen/bewahren, wie zB das Portal LOV - Linked Open Vocabularies[6], das die Vokabularien kategorisiert und auch die Beziehungen untereinander sichtbar macht. Allein das LOV listet 520 solcher Vokabularien auf. Es ist unmöglich, an dieser Stelle näher auf diese Vielfalt einzugehen. Aber zumindest vier seien genannt: *Simple Knowledge Organisation Systems,* oder *SKOS*[7]*,* erlaubt jede Art von Taxonomien und Thesauri abzubilden; *Friend of a Friend, FOAF*[8], ermöglicht Beschreibungen von Personen (Visitenkarten) und ihren Beziehungen untereinander; *dcterms*[9] ist die RDF-Reinkarnation der dublincore terms, des gemeinsamen Nenners vieler Metadaten-Ansätze. *schema.org*[10] ist eine gemeinsame Initiative von kommerziellen Big-Playern, die das Ziel verfolgen durch eine einheitliche semantische Auszeichnung bessere Strukturierung und dadurch Auffindbarkeit von Web-Inhalten zu gewährleisten.

Womit verlinken?

Schlimmer noch als beim Beschreibungsvokabular, herrscht auch bei den existierenden Datensätzen, die als Kandidaten fürs Verlinken dienen könnten, eine unüberschaubare Vielfalt. Die LOD-Cloud repräsentiert zwar nur einen Ausschnitt aus dem riesigen Fundus und die Linien zwischen den Knoten nur die triviale *sameAs* Identitätsbeziehung[11] (die sich bei genauerem Hinsehen als gar nicht so trivial herausstellt[12]). Aber immerhin bietet sie eine anschauliche erste Landkarte des „Web of Data“. Die Hauptstadt ist von Anbeginn dbpedia[13], die (kontinuierlich) in RDF übersetzte Version von Wikipedia. Inzwischen gibt es einige weitere riesige semantische Ressourcen, kompiliert und konsolidiert aus verschiedenen Quellen, wie z.B. YAGO mit 120 Mio. „Fakten“ über mehr als 10 Mio. Entitäten.

Eine besonders active Community im Rahmen der Geisteswissenschaften ist die Linguistik mit der Website Linguistic Linked Open Data (LLOD)[14], die mit eigener LLOD-Cloud und eigenem LingHub[15] eine feiner aufgelöste Sicht auf die Daten in dieser Domäne bietet. Unter diesen spezialisierten Sites sticht das BabelNet[16] hervor, das – auch automatisch kompiliert – eine Reihe von strukturierten Quellen zu einem multilingualen enzyklopädischen und lexikographischen Wörterbuch zusammenführt, das gleichzeitig ein riesiges semantisches Netzwerk darstellt. (14 Mio. synsets, 272 Sprachen)

Besonders relevant fürs Verlinken sind die großen Referenzressourcen. Lange vor dem Semantic Web und sogar lange vor dem Internet und Digitalia überhaupt wurde das Material in Bibliotheken strukturiert, u.a. mit Klassifizierungsschemata, erfasst. Zunehmend werden diese großen traditionellen Referenzressourcen wie die Gemeinsame Normdatei der Deutschen Nationalbibliothek, die Library of Congress Subject Headings (LCSH), die Getty Thesauri, etc. als LOD bereitgestellt.

---

[6] http://lov.okfn.org/dataset/lov/
[7] http://www.w3.org/2009/08/skos-reference/skos.html
[8] http://xmlns.com/foaf/spec/
[9] dublincore.org/2012/06/14/dcterms
[10] http://schema.org/
[11] http://sameas.org/
[12] Halpin, H., & Hayes, P. J. (2010). When owl: sameAs isn't the Same: An Analysis of Identity Links on the Semantic Web. In *LDOW*. Retrieved from http://events.linkeddata.org/ldow2010/papers/ldow2010_paper09.pdf
[13] http://wiki.dbpedia.org/
[14] http://linguistic-lod.org/llod-cloud
[15] http://linghub.lider-project.eu/
[16] http://babelnet.org/about

Eine auch linguistisch sehr relevante Ressource ist EuroVoc[17], der multilinguale Thesaurus der EU.

Das wirklich Beeindruckende ist das „O(open)“ in LOD. All diese riesigen Ressourcen sind frei verfügbar zum Abfragen oder Herunterladen, viele davon auf datahub[18], einem offenen Repositorium für Datensätze aller Art.

Wozu ist es gut?

Das Leitthema der DHd[19] 2016 Tagung in Leipzig[20] bringt es auf den Punkt: “Modellierung – Vernetzung – Visualisierung: Die Digital Humanities als fächerübergreifendes Forschungsparadigma”[21]. Interessanterweise trifft der Dreifachschwerpunkt genauso gut auf semantische Technologien wie auf DH zu.

Ausgehend von den Forschungsobjekten muss also zuerst die Entscheidung getroffen werden, wie diese digital erfasst, modelliert werden. Eine Reihe von (open-source und kommerziellen) Applikationen bietet komfortable Benutzerschnittstellen zum Eingeben von RDF-basierten Daten an: Protégé[22], OntoWiki[23], Topbraid Composer, Semantic Media Wiki, uvm[24]. Es gibt eine Reihe von spezialisierten Tools für die Verwaltung von Thesauri, TemaTres[25], OpenSKOS[26], OpenTheso. Zum Speichern und Abfragen der RDF-Daten werden sogenannte Triple-Stores eingesetzt, auch hier eine breite Auswahl an Lösungen[27] [28]. Es ist aber auch möglich, die Daten in traditionellen Datenbank-Applikationen einzugeben und nur über Schnittstellen eine Export-Möglichkeit für RDF zu schaffen. Wie bereits zuvor erwähnt, ist RDF primär ein Datenaustauschformat.

Der spannendste Teil der Arbeit ist wohl die Annotation, das Anreichern der eigentlichen Forschungsdaten mit Links zu den Referenzressourcen. Dies können Verweise von einem Bild auf Begriffe in einer Taxonomie sein, oder die Auflösung von Bezeichnungen in Texten zu Entitäten. Hierfür steht auch eine Reihe von (tlw. webbasierten) Diensten zur Verfügung: „Reconciliation Services“, wie Babelfy[29], Stanbol[30], DBpedia Spotlight[31], OpenNLP[32]. Obwohl diese Dienste mit riesigen Datensammlungen im Hintergrund und mit beachtlichem Durchsatz arbeiten, ist auch hier Vorsicht geboten, die Trefferquote kann sehr stark variieren. Neben unerkannten Begriffen, ist die Disambiguierung die größte Herausforderung. Idealerweise gehen das automatische und das manuelle Verarbeiten Hand in Hand.

Ein Aspekt, der oft unbeachtet bleibt, ist das „Reasoning“, das logische Schließen, bzw. Inferieren neuer Fakten. Während es traditionell in der Domäne der wissensbasierten Systeme eine zentrale Rolle spielt, begnügt man sich beim Semantic Web oft mit reiner Beschreibung. Ein Grund dafür

---

[17] http://eurovoc.europa.eu/drupal/
[18] http://datahub.io/
[19] http://dig-hum.de/
[20] http://dhd2016.de/
[21] *http://www.dhd2016.de/node/9*
[22] http://protege.stanford.edu/
[23] http://aksw.org/Projects/OntoWiki
[24] http://www.w3.org/2001/sw/wiki/Category:Editor
[25] http://www.vocabularyserver.com/
[26] http://openskos.org/
[27] http://www.w3.org/wiki/LargeTripleStores
[28] http://www.w3.org/2001/sw/wiki/Category:Triple_Store
[29] http://babelfy.org/
[30] http://stanbol.apache.org/
[31] http://spotlight.dbpedia.org/
[32] https://opennlp.apache.org/

könnte sein, dass für diesen Zweck viel strikter definierte Ontologien mit einer Reihe logischer Aussagen über die modellierte Domäne erforderlich ist.

Zu guter Letzt zum wohl interessantesten Aspekt: die Ausgabe, die Darstellung, die Nutzung der aufbereiteten und vernetzten Daten. Jeder Triple-Store ist mit einem SPARQL-Endpoint[33] ausgestattet, der Möglichkeit, die RDF-Daten in der Abfragesprache SPARQL zu erkunden. Wesentlich leichter zu nutzen ist die (auf SPARQL-Endpoints aufsetzende) semantische Suche, die durch die Kombination von Volltext-Suche, facettierter Navigation und weiterführenden Links in den Resultaten die Daten als einen navigierbaren Graphen vor den Benutzern ausbreitet. Dazu kommen zunehmend weitere Visualisierungskomponenten wie Kartensichten, Zeitleisten, dynamische interaktive Netzwerke. Selbstredend passiert das heutzutage alles im Web-Browser, idealerweise offen für die ganze Welt. All das gibt und gab es auch ohne LOD, ergibt aber zusammengenommen einen Werkzeugkasten, der der (geisteswissenschaftlichen) Forschung völlig neue methodische Horizonte eröffnet.

Die semantischen Technologien und insbesondere das LOD-Paradigma liefern also eine konzeptuelle und technische Grundlage für ein globales Wissensnetzwerk. Was uns trotz allem nicht erspart bleibt, ist die interpretative wissenschaftliche Arbeit.

Matej Ďurčo

is a key figure of the Austrian Centre for Digital Humanities who has been contributing substantially to the Austrian DH infrastructure activities and to drafting and building-up of the new institute of the Academy which has been operative as of early 2015. Since then, he has been leading the technical department of the ACDH.

Proceeding from an IT background (Technical University of Vienna), he has been involved in a wide range of humanities computing activities at the Austrian Academy of Sciences and the University of Vienna. He has mainly participated in language oriented projects focusing on the creation, annotation and exploitation of digital corpora. Particularly worth mentioning is his engagement for the European infrastructure consortia CLARIN (Common Language Resources and Technology Infrastructure) and DARIAH (Digital Research Infrastructures for the Arts and Humanities) where he has been involved in a wide range of activities on the Austrian and the European levels.

For his continued commitment to the development of the core components of the CLARIN research infrastructure, he was awarded in October 2014 by the CLARIN-ERIC consortium the Major Achievements Young Scientist Award, an award that was handed out for the first time.

[33] http://www.w3.org/wiki/SparqlEndpoints