Weitere Beiträge

Christian Stein

Linked Open Data – Wie das Web zur Semantik kam

Zusammenfassung: Seit Beginn des Word Wide Web besteht die Idee der universellen Vernetzung von Information in einem einfachen, standardisierten Verfahren. Ein Computer stößt bei der Interpretation von natürlich sprachlicher Bedeutung jedoch schnell an seine Grenzen. Die letzte Generation des Webs, das Semantic Web, schließt diese Lücke und modelliert Semantik explizit und maschinenlesbar – und eröffnet damit neue Möglichkeiten.

Schlüsselwörter: Semantic Web; Linked Open Data; Ontologie; LOD; RDF; Europeana; Semantik

Linked Open Data – How the Web Came to Semantics

Abstract: Since the beginning of the World Wide Web, there is the idea of universal networking of information in a simple, standardized way. However, computers are still very limited in the interpretation of natural language. The last generation of the Web, the Semantic Web, closes this gap and models semantics explicit and machine-readable – and opens up new possibilities.

Keywords: Semantic Web; Linked Open Data; ontology; LOD; RDF; Europeana; semantics

DOI 10.1515/bfp-2014-0055

Inhalt

1 Die Idee des Webs

Als Vannevar Bush im Jahr 1945 die Idee zu seinem System MEMEX präsentierte, konnte er sich noch nicht vorstellen, dass Hypertext-Systeme kein halbes Jahrhundert später die Welt verändern sollten. Er beschreibt es als „a sort of mechanized private file and library. It would use microfilm storage, dry photography, and analog computing to give postwar scholars access to a huge, indexed repository of knowledge – any section of which could be called up with a few keystrokes.“[1] Auch wenn Technologien wie Mikrofilm oder analoges Rechnen heute antiquiert wirken mögen, wurde damit doch eine Grundidee konzipiert, deren erste technische Realisierung sich bereits ein paar Jahre später mit dem ARPANET ergeben sollte. Das „Advanced Research Projects Agency Network“ war ein Kooperationsprojekt zwischen nordamerikanischen Universitäten und dem Militär und gilt als Vorläufer des Internets. Mit einem Schlag war es möglich geworden, Informationen zwischen den Forschungsstätten elektronisch zu durchsuchen, anzufordern und zu übermitteln. Während am Anfang nur ein paar wenige Standorte vernetzt waren, spannte sich 1977 bereits ein respektables Netz zwischen Ost- und Westküste und immer mehr Forscher erkannten den Wert dieser Technik, mit der Information sich spürbar von ihren materiellen Datenträgern gelöst hatte. Dieser Ablösungsprozess bedeutete einen Schub für die kollaborative Forschung, hatte aber zu diesem Zeitpunkt noch einen entscheidenden Nachteil: ARPA- und später Internet wurden von der Kommandozeile aus bedient, man musste wissen, bei welchen Servern man welche Informationen finden konnte und die Informationssuche war dementsprechend immer noch schwierig, zeitaufwändig und teuer. Diese Tatsachen machten das Netz zu einem teuren Forschungsinstrument, das nahezu ausschließlich von Experten genutzt wurde – bis das World Wide Web kam.[2]

Es kam mit Tim Berners-Lee, der als Physiker und Informatiker am CERN in der Schweiz arbeitete. Er kannte das Internet in seiner damaligen Form und war begeistert von den sich bietenden Möglichkeiten. Aber er sah auch, welche Einschränkungen und Inkompatibilitäten es gab und dachte darüber nach, wie sich das Internet zu dem offenen, einfachen, standardisierten und vielfältigen In-

Dr. Christian Stein: christian.stein@hu-berlin.de

1 Wardrip-Fruin, Noah; Montfort, Nick (eds.): The New Media Reader. Cambridge MA, London 2003, S. 35.

2 Marshall T. Poe: A History of Communications. Media and Society from the Evolution of Speech to the Internet. Cambridge 2010.

formationsnetz entwickeln könnte, das wir heute kennen. Er ersann das World Wide Web. Die Kernprinzipien waren einfach, schlank designt, schnell erlernbar und basierten auf dem, was es schon gab. Das erste war das Übertragungsprotokoll HTTP, das alle Anfragen und Antworten zwischen Client und Server regelte und die Informationspakete an die richtige Stelle navigierte. Dieses „Hypertext Transfer Protocol" war in erster Linie dazu gedacht, Hypertext zu versenden. Um solchen Hypertext schreiben zu können, entwickelte Berners-Lee HTML gleich mit. Die Hypertext Markup Language, die heute in der Version 5 gerade ihr letztes großes Update bekommen hat und nach wie vor das Web beherrscht, machte es leicht, Dokumente zu erstellen, zu formatieren und – vor allem – sie über Hyperlinks miteinander zu verbinden. HTML machte dabei keine saubere Trennung zwischen Inhalt und Darstellung, was später bemängelt und korrigiert wurde, was das Erlernen aber ungleich einfacher machte und so wesentlich zur Verbreitung der Sprache beitrug. Um nun tatsächlich Verbindungen zwischen Dokumenten aufbauen zu können, entwickelte Berners-Lee die URL, den Uniform Ressource Locator, der Dokumente (und Teile davon) einheitlich referenzierbar machte.[3]

Damit war alles Notwendige vorhanden, um Informationen einheitlich, technologieübergreifend und einfach auffindbar zu machen, zu präsentieren und zu verbinden. Denn das war die Idee: Kein zentrales Netzwerk in der Hand einer einzelnen Autorität, keine Einstiegshürde durch komplizierten Zugang, keine Abhängigkeit vom Hersteller einer Technologie, keine teuren Lizenzkosten. Vor allem aber: Wissen gibt es nicht isoliert, Informationen stehen immer im Kontext. Diesen Kontext mit zu modellieren, die Verbindungen der Informationen zueinander explizit zu machen, das sollte das Wissensnetzwerk der Zukunft werden, in dem Computer und Menschen enger und besser zusammenarbeiten können sollten.

Schnell folgte die erste Webseite info.cern.ch, die es auch heute noch gibt. Und es folgte die erste Suchmaschine, die serverübergreifend einen einfachen Zugang zu dieser neuen Informationswelt bot. Dieses Web, heute häufig Web 1.0 genannt, wird auch das „Web of documents" genannt – ein Netz aus mit Hyperlinks verknüpften HTML-Dokumenten. Dieses Netz fand Anwender auf der ganzen Welt und bald auch jenseits der Universitäten. Nur so konnte es zu einem globalen Informationsnetz in all seiner Vielfältigkeit werden. Bei aller Einfachheit jedoch hatte es seinen größten Sprung an Nutzerzahlen noch vor sich: 1999 wurde das Web interaktiv und sozial. Informationskonsumenten wurden selber zu Produzenten, faktisch jeder konnte Inhalte beisteuern, verbessern, kommentieren und vernetzen. Mit diesem Web 2.0, wie es dann genannt wurde, explodierte die Zahl der verfügbaren Inhalte genauso wie die Nutzerzahlen. Wieder war das Prinzip das der massiven Vernetzung und eines einfachen, partizipativen Zugangs für alle, der auf offenen, gemeinsamen Standardtechnologien beruhte.

2 Semantic Web

Heute stehen wir an der Schwelle zur nächsten Generation des Webs, dem Web 3.0. Eigentlich haben wir sie schon überschritten. Denn das „Semantic Web" ist bereits überall. Nur ist das noch nicht ganz so bewusst, da viel von dem, was das Semantic Web ausmacht, sich im herkömmlichen Web-Gewand zeigt. Die Versionsnummern des Webs bilden gewissermaßen eine Konzentration auf Aspekte, die bereits in der ersten Idee des Webs vorhanden waren. So wie das Web 2.0 auf die Einfachheit und Heterogenität der Partizipation fokussierte, nimmt das Web 3.0 jetzt die Bedeutung der Information, die Semantik, ins Visier. Denn wenn wir von Informationen reden, meinen wir nicht im eigentlichen Sinne Dokumente, die immer schon Aggregationen von Informationen darstellen, sondern basale Fakten: Die Bevölkerungszahl einer Stadt, den Namen eines Gebäudes, die Lage einer Straße, die Bestandteile eines Stuhls, das Motiv eines Gemäldes, die Gattung eines Tieres, die Richtigkeit einer Aussage. Es sind gewissermaßen die atomaren Bestandteile von komplexen Aussagen, wie sie in einer Diskussion oder einem Dokument gemacht werden: Minimalaussagen, die sich zu komplexen Aussagen zusammenfügen. Mit diesen Minimalaussagen ist es dabei tatsächlich ähnlich wie mit Atomen: Sie tauchen immer wieder in unterschiedlichen Zusammenhängen auf, bleiben in sich aber strukturell gleich. Und was sich wiederholt, kann man wiederverwenden. Ganz im Sinne des ursprünglichen Gedankens des Webs, der vielfältigen Verbindung von Informationen, die einen Bezug aufweisen, ist das Semantic Web die technische Modellierung von Minimalaussagen und deren Bezügen zueinander: nicht mehr primär im Sinne von Dokumenten und Hyperlinks zwischen diesen, sondern im Sinne einer bedeutsamen Beziehung. Während der Hyperlink nur eine Verbindung herstellt, ohne etwas darüber zu sagen, wie diese Verbindung beschaffen ist, stellt die semantische Relation eine bedeutsame Verbindung her. Der Hyperlink moduliert: Diese Zeichenfolge in Dokument A ist verbun-

3 Tim Berners-Lee; Mark Fischetti: Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor. London 1999.

den mit diesem Absatz von Dokument B. Eine Folge semantischer Relationen dagegen kann Folgendes modulieren: Diese Zeichenfolge ist der Name einer Person X. Person X arbeitet an der Humboldt-Universität. Die Humboldt-Universität beinhaltet das Jacob-und-Wilhelm-Grimm-Zentrum. Person X hat ein Büro im Jacob-und-Wilhelm-Grimm-Zentrum. Jeder dieser Sätze stellte dabei eine semantische Relation zwischen zwei Instanzen her. Durch die Wiederholung der Bezeichner dieser Instanzen können so komplexere Informationen abgebildet werden, da sich mehr Informationen zu einem Zusammenhang vernetzen lassen. Und diese Vernetzungen können beliebig groß werden und beliebig viele Details enthalten.

In einem Dokument nun, beispielsweise einem Artikel wie diesem, wird ein kleiner Teil dieser Informationen vom Autor danach ausgewählt, was für die jeweilige Argumentation relevant ist. Sie werden in natürlicher Sprache verknüpft und formen so ihre Aussage. Im Semantic Web dagegen befinden sich keine Argumentationen und keine Vorauswahl, sondern nur die basale Information selbst – und das in rauen Mengen und ungefiltert. Was zunächst wie ein Nachteil aussieht ermöglicht aber Erstaunliches: Da dieses geballte Wissen nicht nur von Menschen gelesen werden kann, sondern noch besser von Maschinen, können diese uns auch besser helfen, die für uns relevanten Fakten herauszufiltern. Man kann sich das wie eine riesige, verteilte Datenbank vorstellen, die nicht für einen einzigen speziellen Zweck gebaut wurde, sondern für beliebige Aussagen aus allen Bereichen menschlichen Wissens. So sagt Tim Berners-Lee darüber: „The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation.“[4]

Das Semantic Web macht also Semantik explizit und für Maschinen verständlich. Was für den Computer vormals nur Zeichenketten waren, wird nun zu berechenbarer Bedeutung. Die Bedeutungen natürlicher Sprache, deren Wörter in unterschiedlichen Kontexten immer Unterschiedlichstes bedeuten können, werden eindeutig und formal definiert. Damit ist eine neue Stufe der Informationsmodellierung erreicht.

3 Linked Data

Unter dem Stichwort *Linked Data* nun verbergen sich im Wesentlichen die Ideen und Techniken des Semantic Web. Der Schwerpunkt liegt dabei zunächst weniger darauf, Daten menschenlesbar darzustellen, als vielmehr darauf, sie maschinenlesbar zu machen. Man muss hier unterscheiden zwischen Linked Data, Open Data und Linked Open Data. Linked Data bezieht sich auf die semantische Verknüpfung von Daten im technischen Sinne. Die Verbindungen zwischen den Daten sind also mit einer explizit definierten Bedeutung versehen. Das beinhaltet nicht zwangsläufig, dass diese Daten auch für jedermann offen zugänglich sind; es können ebenso gut nichtöffentliche Daten, beispielsweise eines Unternehmens, gemeint sein. Open Data dagegen referiert darauf, dass Daten offen für jeden zugänglich und unter einer offenen Lizenz publiziert sind. Gerade öffentliche Institutionen stellen ihre Informationen vermehrt den Bürgern zur Verfügung. Das Recht darauf regelt in den USA beispielsweise der *Freedom of Information Act*. Aber auch in Europa werden inzwischen in großem Umfang Daten für die Bevölkerung geöffnet. Über die Verknüpfung der Daten sagt Open Data noch nichts aus. Es kann sich also um unverknüpfte Informationen handeln, deren Bedeutungserschließung der menschlichen Interpretation bedarf. *LOD* oder *Linked Open Data* vereint beide Aspekte der Verknüpfung und der Offenheit und stellt daher auch den Königsweg für alle dar, die im öffentlichen Interesse Daten anbieten wollen. Anstatt Daten in geschlossenen und proprietären Datenbanken verstauben zu lassen bzw. nur über eigene Interfaces anzubieten, können sie nun von beliebigen Interessenten nachgenutzt und erweitert werden.[5]

Alle Daten dieser Art bilden zusammen die sogenannte *LOD-Cloud*. Sie bildet ein riesiges Konglomerat vernetzter Daten und enthielt in der letzten Berechnung vom September 2011 insgesamt 31 634 213 770 Tripel. Inzwischen dürften es weit mehr sein. Zu den Datenprovidern gehören beispielsweise die DBpedia mit den strukturierten Inhalten der Wikipedia, Project Gutenberg mit Metadaten zu lizenzfreien Texten, LinkedGeoData mit weltumspannenden Ortsinformationen, das World Fact Book mit sehr aktuellen Daten über alle Länder der Welt, Open Library mit umfangreichen bibliographischen Daten und viele mehr. Ein Großteil der Daten stammt von öffentlichen Einrichtungen und Regierungsorganisationen. Einen guten Überblick in einer interaktiven Infografik verschafft die Seite lod-cloud.net, auch wenn sie einen nicht mehr ganz aktuellen Stand wiedergibt.

4 Tim Berners-Lee; James Hendler, Ora Lassila: The Semantic Web. In: Scientific American 284 (5) (2001) S. 34–43.

5 Staab, S.; Scheglmann, S.; Leinberger, M.; Gottron, T.: Programming the Semantic Web. In: The Semantic Web: Trends and Challenges (2014) S. 1–5.

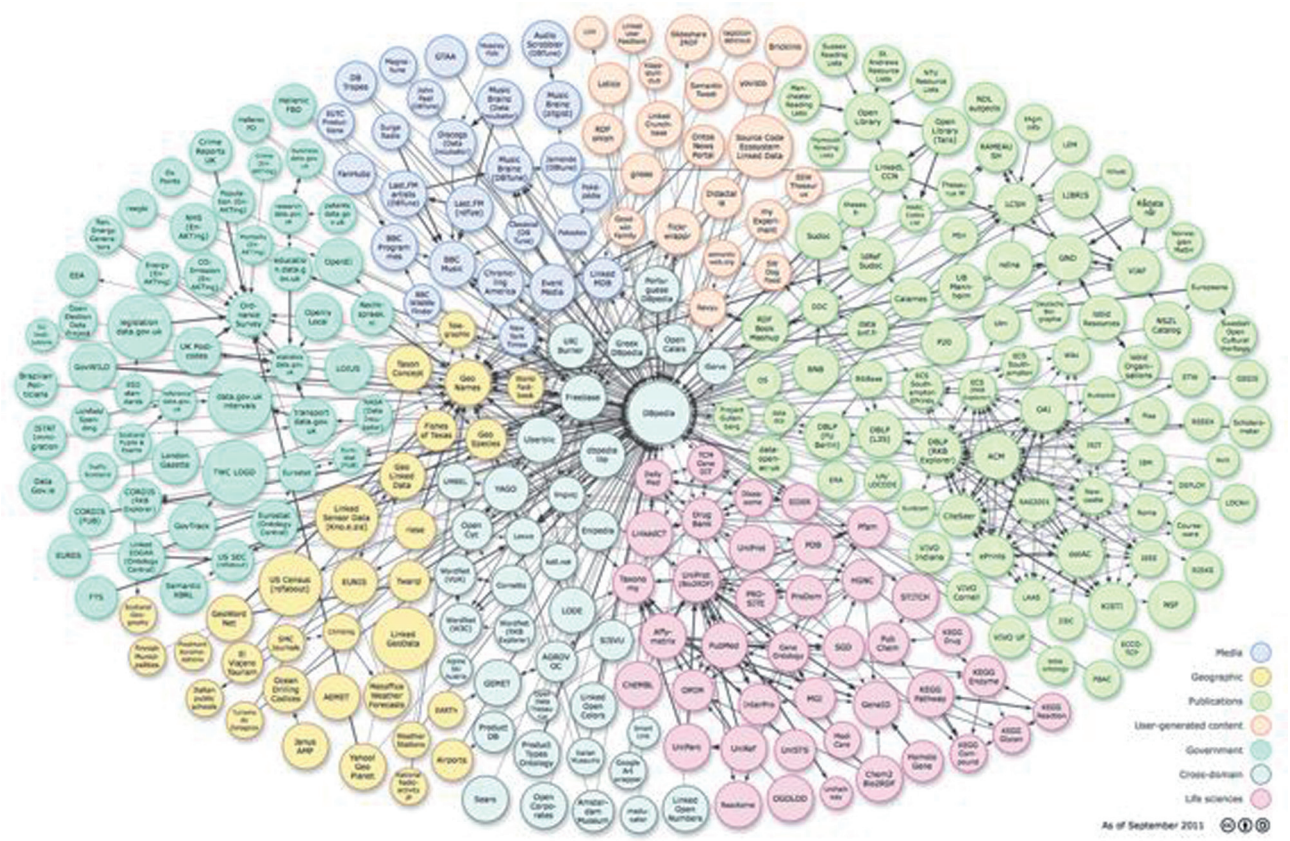

Abb. 1: Linked Open Data Cloud Diagramm[6]

Tim Berners Lee hat für die Distribution von Daten fünf aufeinander aufbauende Kriterien definiert, deren Beachtung gute Datenqualität sicherstellen soll[7]:

Tab. 1: Tim Berners-Lee's 5-Sterne Kriterien für Linked Open Data

★	Available on the web (whatever format) but with an open licence, to be Open Data
★★	Available as machine-readable structured data (e.g. excel instead of image scan of a table)
★★★	As (2) plus non-proprietary format (e.g. CSV instead of excel)
★★★★	All the above plus: use open standards from W3C (RDF and SPARQL) to identify things, so that people can point at your stuff
★★★★★	All the above, plus: link your data to other people's data to provide context

Die dahinterstehende Idee ist die, dass Daten immer wertvoller werden, je stärker sie vernetzt und interpretierbar sind. Um diese Interpretierbarkeit plattformübergreifend zu gewährleisten, sollten die verwendeten Formate ebenso wie die Daten selbst frei nutzbar und offen dokumentiert sein. Dies wird mit der technischen Grundstruktur des Semantic Web gewährleistet, die im folgenden Kapitel basal erklärt wird.[8]

4 Technische Grundstruktur

Die wesentlichen Technologien, die das Semantic Web antreiben, sind in offenen Standards des W3C (World Wide Web Consortium) festgehalten. Der grundlegende Standard ist *RDF*, das Ressource Description Framework, auf dem alle weiteren Standards basieren.[9] Mit RDF lassen sich Aussagen über beliebige Dinge bzw. Ressourcen formen.

6 Cyganiak, Richard; Jentzsch, Anja: Linking Open Data cloud diagram. 2011. URL: http://lod-cloud.net/.
7 Berners-Lee, Tim: Linked data-design issues. 2006. URL: http://www.w3.org/DesignIssues/LinkedData.html.
8 Berners-Lee, Tim: Long live the web. In: Scientific American 303 (6) (2010) S. 80–85.
9 http://www.w3.org/TR/rdf-primer.

Dabei kann es sich sowohl um Dinge der physischen Außenwelt, wie auch digitale, imaginäre oder begriffliche Ressourcen handeln. Die einfachste und wichtigste Form dieser Aussagen konstituiert sich durch sogenannte *Tripel*, die aus Subjekt, Prädikat und Objekt bestehen. Sie funktionieren wie einfache Aussagesätze, z. B.:

:Jacob-und-Wilhelm-Grimm-Zentrum :location :berlin

Dieses Tripel besagt, dass das Jacob-und-Wilhelm-Grimm-Zentrum (Subjekt) in Berlin (Objekt) lokalisiert ist (Prädikat). Subjekt, Prädikat und Objekt werden dabei immer mit einem eindeutigen Bezeichner angegeben, der an natürliche Sprache angelehnt sein kann, aber nicht muss. Wichtig ist nur, dass die Bezeichnung für den Computer eindeutig ist. Genauso gültig wäre also die (gleichbedeutende) Aussage:

:Q1525910 :location :Q64

In diesem Fall sind das Jacob-und-Wilhelm-Grimm-Zentrum (:Q1525910) und Berlin (:Q64) jeweils durch Codes repräsentiert, die sich Menschen nicht mehr unmittelbar erschließen. Die eindeutigen Bezeichner, ob menschenlesbar oder nicht, nennen sich *URI* für Uniform Ressource Identifier. URIs sehen meist aus wie die allseits als Webadresse bekannten URLs. Da diese auf dem Domain Name System basieren, sind sie weltweit eindeutig. Das ist ein großer Vorteil, denn so können URIs aus verschiedensten Quellen in RDF verwendet und kombiniert werden, ohne dass es zu Ambiguitäten kommt. Während wir in den obigen Beispielen eine Kurznotation gewählt haben, bestünden unsere Beispiel-Tripel also eigentlich aus drei kompletten URIs für Subjekt, Prädikat und Objekt:

<http://de.dbpedia.org/resource/Jacob-und-Wilhelm-Grimm-Zentrum>
<http://dbpedia.org/ontology/location>
<http://de.dbpedia.org/resource/Berlin>

Das ist für Menschen zwar etwas langwieriger zu lesen, die komplette Notation der URIs erlaubt uns jedoch zu sehen, wer die entsprechenden Ressourcen ursprünglich definiert hat.[10]

Mit Tripeln wie diesen können in RDF beliebig viele Aussagen getroffen werden, die zusammen einen sogenannten *RDF-Graph* bilden. Die Informationen, die durch solche RDF-Graphen beschrieben werden, können dabei in anderen Graphen wiederverwendet werden. Die Tripelstruktur von RDF und die RDF-Graphen, die dadurch erzeugt werden können, bilden die Basis für alles Weitere. Auf RDF basieren alle Erweiterungen, mit denen Schemata definiert, Abfragen gestellt oder komplexe logische Aussagen getätigt werden können.[11]

5 Erweiterungen

Zur Definition von Schemata dient *RDFS* (RDF Schema), mit der Ober- und Unterklassen sowie die verwendbaren Prädikate und ihre Verwendung definiert werden. Während RDF sich immer auf konkrete Aussagen über konkrete Ressourcen bezieht (z. B. :Berlin, :Jacob-und-Wilhelm-Grimm-Zentrum), wird mit RDFS die generelle Struktur von möglichen Tripeln bestimmt (z. B. :City als abstrakte Klasse für :Berlin). Solche Schemata, auch Vokabularien genannt, existieren bereits zu einer Vielzahl von Anwendungsgebieten und können einfach nachgenutzt werden. Es ist aber auch möglich, eigene Schemata zu definieren.

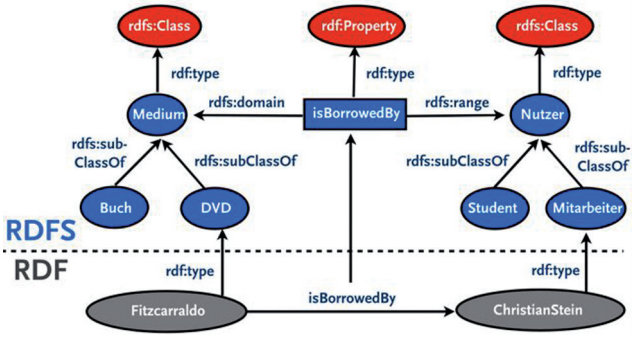

Abb. 2: RDF und RDFS im Beispiel

Im abgebildeten, vereinfachten Beispiel ist im oberen Bereich ein Schema in RDFS modelliert. In ihm sind die Klassen Medium und Nutzer definiert, die jeweils speziellere Unterklassen besitzen (Buch und DVD als Unterklasse für Medium und Student und Mitarbeiter als Unterklasse für Nutzer). Das Prädikat *isBorrowedBy* gibt eine Möglichkeit an, Instanzen beider Klassen zu verknüpfen. Der untere Teil der Abbildung zeigt dann die Anwendung dieses Schemas in RDF, also konkreten Instanzen, bei denen ein

10 In dem obigen Beispiel greifen wir auf die umfangreichen Bestände der DBpedia zu, die die strukturierten Informationen der Wikipedia in RDF bereitstellt: http://wiki.dbpedia.org/.

11 Hitzler, Pascal; Krötzsch, Markus; Rudolph, Sebaastian: Foundations of Semantic Web Technologies. Boca Raton 2011.

konkretes Medium von einem konkreten Nutzer als ausgeliehen definiert ist.

Hat man einen umfangreichen RDF-Graph erstellt, benötigt man einen Weg, spezifische Informationen aus diesem abzufragen. Dazu dient die Abfrage-Sprache *SPARQL* (SPARQL Protocol And RDF Query Language). Sie ähnelt der verbreiteten Datenbankabfragesprache SQL und verwendet ebenfalls Tripel in ihren Abfragen. Mit SPARQL lassen sich beliebig komplexe Abfragen formulieren und entsprechend passgenaue Informationen extrahieren. So könnte man beispielsweise mit SPARQL erfragen, welche Künstler in Berlin zwischen 1900 und 1910 geboren wurden, wann sie gestorben sind und wie sich ihr Name in anderen Sprachen schreibt.[12] SPARQL ist aber nicht nur eine Abfragesprache, sondern auch ein Protokoll zu Kommunikation mit einem Server. Typischerweise verwaltet ein Server für RDF-Graphen diese mithilfe eines *Triplestores*, also einer Datenbank, die für Tripel optimiert ist. Ein Triplestore bietet einen SPARQL-Endpoint, an den SPARQL-Abfragen per HTTP adressiert werden können und der die entsprechenden Ergebnisse maschinenlesbar zurückliefert.

Für die Verbindung zur Dokumentenwelt sorgt der Standard *RDFa* (RDF in Attributes). Mit seiner Hilfe lassen sich RDF-Aussagen in HTML- und XML-Dokumente einbetten. Taucht beispielsweise auf einer HTML-Webseite ein bestimmter Name auf, kann mit RDFa eindeutig auf die entsprechende, tatsächlich gemeinte Ressource verwiesen werden. Findet sich beispielsweise die Zeichenfolge „Paris“ auf einer Webseite, kann RDFa im Hintergrund klarstellen, ob sich diese auf Paris in Frankreich, Texas oder Idaho bezieht oder gar der Stummfilm gemeint ist. Das kann dann auch ein Suchalgorithmus berücksichtigen und bessere Ergebnisse liefern. So kann RDFa dazu beitragen, die Ambiguität natürlicher Sprache zu klären und Eindeutigkeit für Mensch und Maschine gleichermaßen herzustellen.

Mit der Erweiterung *OWL* (Web Ontology Language) schließlich lassen sich komplexe logische Aussagen erstellen, wie sie in der Prädikatenlogik möglich sind. Damit können sogenannte Ontologien aufgebaut werden. So lassen sich beispielsweise implizite Klassen, Beschränkungen, Schnittmengen, inverse Relationen oder explizite Differenzen modellieren. Mit diesen Erweiterungen wird es möglich, Zusammenhänge zu berechnen, die niemals explizit angelegt worden sind, sich jedoch rein logisch ergeben. Eine solche Berechnung nennt sich *Reasoning*. Solche Verfahren werden beispielsweise genutzt, um automatisch Entscheidungen in komplexen Situationen treffen zu können, die Bedeutung von Texten zu analysieren oder versteckte Verbindungen aufzudecken.

6 Schwierigkeiten und Grenzen

Bei allen positiven Aspekten werden in der Forschung aber auch grundsätzliche Probleme der Architektur diskutiert, die zu tieferliegenden Fragen führen. Eine fundamentale Unterscheidung, die für RDF und RDFS konstituierend ist, bezieht sich auf das Verhältnis von Klassen, Instanzen und realweltlichen Objekten. In der Philosophie sind Ontologien definiert als die Ordnung bzw. Einteilung des Seienden. Diesen Anspruch hat die Informatik gewissermaßen übernommen, allerdings in einem simplifizierenden Sinne, der der Komplexität dieses Problems nicht immer gerecht werden kann. Gewissermaßen ist es ein semiotisches Problem, das hier virulent wird, wenn es um eine saubere Unterteilung von Klassen und Instanzen geht. Die Klarheit dieser Unterscheidung stammt aus der objektorientierten Programmierung, einem heute dominierenden Programmierparadigma. In diesem wird davon ausgegangen, dass Klassen gewissermaßen Konstruktionspläne für Instanzen sind, aus denen zur Laufzeit dann beliebig viele Instanzen generiert werden können. Das bedeutet, dass die Klasse immer zuerst da ist und aus ihr dann Instanzen generiert werden. In einer solchen Umgebung ist immer klar, zu welcher Klasse eine Instanz gehört, da sie ja aus dieser generiert wird.

Dieses Prinzip wurde im Semantic Web nun als grundlegendes Paradigma übernommen. Klassen werden hier allerdings definiert als „sets of individuals“[13]. Immer dann, wenn man von etwas physisch Konkretem spricht, beispielsweise dem Eifelturm oder Albert Einstein, verwendet man *Individuals*. Gehören diese zu einer Klasse, spricht man von Instanzen dieser. RDF-Aussagen sind dabei nur zwischen Individuals bzw. Instanzen möglich. Im Vergleich zur Objektorientierten Programmierung ist das Verhältnis von Klasse und Instanz als Beschreibungsmuster für eine bereits existierende Welt und Sprache aber problematisch. In der exemplarischen Wine-Ontologie[14], die immer wieder für auch vom W3C selbst als Beispielontologie herangezogen wird, wird das deutlich. Einzelne Weine, beispielsweise ein Longridge Merlot werden hier als Instanzen definiert, wohingegen der Merlot eine Klasse bil-

12 Entsprechende Informationen stellt beispielsweise die DBpedia für jeden kostenlos und offen zur Verfügung: http://wiki.dbpedia.org/OnlineAccess#h28-6.

13 http://www.w3.org/TR/owl2-syntax/#Classes.

14 http://www.w3.org/TR/owl-guide/.

det. Die Frage die sich hier stellt ist, ob ein Longridge Merlot tatsächlich eine Instanz darstellt oder nicht vielmehr selbst eine Klasse ist, nämlich die aller Jahrgänge und Abfüllungen. Ein anderes Beispiel: Ist der Film Fitzcarraldo eine Instanz oder können wir von Instanzen erst bei einer konkreten DVD oder Filmrolle sprechen?

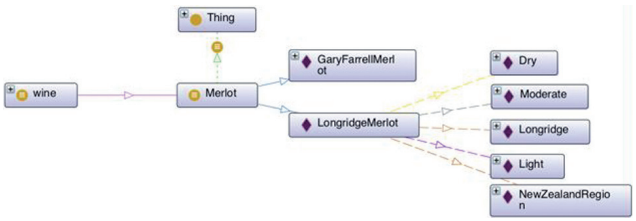

Abb. 3: Merlot als Klasse (gelber Kreis) und Longridge Merlot als Instanz (lila Raute) mit weiteren verknüpften Instanzen in der OntoGraf-Visualisierung des Ontologie-Editors *Protegé*

Noch gravierender wird das Problem mit Begriffen. Ist ein Begriff wie „Tier" eine Instanz, wie es in der weitverbreiteten Ontologie SKOS (Simple Knowledge Organisation System) definiert ist, obwohl der Begriff selbst eine Klasse beschreibt? Was ist mit Begriffen wie „Einhorn", die keine physischen Instanzen aufweisen, aber als Begriff existieren? Sind Begriffe nicht per se Klassen des Systems Sprache, weil sie immer auf unterschiedliche konkrete Sachverhalte anwendbar sind? Wie geht man mit den unterschiedlichen Definitionen eines Begriffs um, die alle in unterschiedlichen Kontexten als korrekt angesehen werden? Die vorgeschlagenen und praktizierten Lösungen dazu bewegen sich im Bereich der Metamodellierung[15] und versuchen, Mischformen zwischen Instanzen und Klassen einzurichten, was ebenfalls problematisch und nicht abschließend geklärt ist. Die Unterscheidung hat für die Modellierung mitunter massive Konsequenzen, denn je nachdem ob etwas eine Klasse oder eine Instanz ist, können unterschiedliche Verknüpfungen angelegt und berechnet werden.

Da Modellierung immer eine Designfrage ist, die unterschiedliche Modellierer unterschiedlich lösen, entstehen verschiedene Varianten semantisch sehr ähnlicher oder äquivalenter Aussagen. Wie gleich aber sind diese wirklich? Immer wieder zu Verwirrungen und Kritik führt beispielsweise das Prädikat owl:sameAs, das auch verwendet wird, um unterschiedliche URIs als inhaltlich identisch auszuweisen. Identität ist aber nur bei physischen Objekten eindeutig zu bestimmen. Werden Begriffe modelliert, muss geprüft werden, ob die jeweils gültigen Definitionen eine identische Bedeutung haben. Diese Entscheidung ist schwierig und subjektiv, führt aber bei einer angelegten Relation dazu, dass alle zukünftigen Berechnungen durch *Reasoner* die Instanzen als identisch behandeln, so dass falsche Schlüsse entstehen.[16] Generell gehen Reasoner davon aus, dass alle verwendeten Prädikate korrekt verwendet wurden. Immer wieder kommt es jedoch zu Fehlern, die sich in sehr großen Graphen kaum noch lokalisieren lassen. Bei allen Vorteilen des Semantic Web wirft das die Frage auf, wie es um die Qualität der Daten bestimmt ist, wie sich diese Qualität messen, einschätzen oder gewährleisten lässt und ob fehlerhafte Daten nicht die Gesamtaussagekraft von berechneten Semantiken in Frage stellen.[17]

Damit verbunden ist die Frage nach den Versionen: Für Dokumente ist es noch relativ leicht, verschiedene Versionen kenntlich zu machen und auch alte Versionen als Referenz aufzubewahren. In vielfach vernetzten, kollaborativ erstellten Graphen jedoch ist jede Ergänzung, Löschung oder Änderung eines Tripels eigentlich eine neue Version des Gesamtgraphen, deren Auswirkungen unter Umständen weitreichend sein können. Wie versioniert man solche Änderungen und wie macht man bestimmte Versionen als Referenz zugänglich? Wie schließlich entscheidet man, ob eine Aussage, die sich auf eine vergangene Version bezieht, auch in einer neuen noch voll gültig ist?

All diese und weitere Fragen beschäftigen die Forschung und fließen in neue Versionen der Standards ein. Auch werden neue Sprachen und Ansätze zur Semantikmodellierung entwickelt, die versuchen, diese Probleme zu lösen. In diesem Sinne trägt die aktive Forschung zum Semantic Web zu einer konstanten Verbesserung der Standards und Prozesse, aber auch zu einem größeren Problembewusstsein bei den Modellierern und Anwendern bei. Auch werden die Grenzen semantischer Modellierung sichtbar: Weder lässt sich alles modellieren, was modelliert werden kann, noch gibt es einen einzigen richtigen Weg der Modellierung. Kontextabhängigkeit und deren Modellierung wird für die Zukunft semantischer Netze genauso wichtig werden wie die Quantifizierung von Datenqualität und die Abschätzung des unbekannten, nichtmodellierten Anteils. Die Verschränkung von Linguistik

15 Glimm, Birte; Rudolph, Sebastian; Völker, Johanna: Integrated Metamodeling and Diagnosis in OWL 2. Technical Report Nr. 3006 Institute AIFB, KIT. http://www.cs.ox.ac.uk/files/3318/TR-GRV-Metamodelling.pdf.

16 Halpin, Harry; Herman, Ivan; Hayes, Patrick J.: When Owl:sameAs Isn't the Same: An Analysis of Identity Links on the Semantic Web. 2010. URL: http://www.w3.org/2009/12/rdf-ws/papers/ws21.

17 Hogan, Aidan; Umbrich, Jürgen; Harth, Andreas; Cyganiak, Richard; Polleres, Axel; Decker, Stefan: An Empirical Survey of Linked Data Conformance. In: Journal of Web Semantics 14 (2012) S. 14–44.

und Informatik zeigt sich hier besonders deutlich, da die Ambiguität natürlicher Sprache eine gewisse Widerständigkeit gegen ihre Formalisierung an den Tag legt, der nicht mit den Mitteln der Informatik allein zu begegnen ist.

7 Anwendung und Anwendungen

Semantic Web Prinzipien und Technologien finden sich heute in einer Vielzahl von Anwendungen. Interessante Projekte sind auch und besonders im Umfeld von Museen, Archiven und Bibliotheken anzutreffen. So sei exemplarisch das British Museum mit seinem *Research Space* genannt.[18] Dabei handelt es sich um eine Webanwendung für kollaborative Forschung zu seinen Beständen, die komplett auf Basis von RDF unter Verwendung des Vokabulars CIDOC CRM realisiert ist[19]. Diese Architektur ermöglicht den Forschern das Auffinden von Zusammenhängen, die sonst verborgen bleiben würden. Ein anderes interessantes Projekt ist das Web-Annotations-Tool *Pundit*[20], das in verschiedenen Forschungsprojekten weiter entwickelt und genutzt wird und in der Lage ist, beliebige Webseiten semantisch zu annotieren. Pundit ermöglicht es, RDF-basierte Tripel direkt im Browser anzufertigen und mit anderen zu teilen, sodass explizite semantische Aussagen über Elemente von Webseiten entstehen. Die BBC verwendet RDF-Architekturen für die Verarbeitung ihrer umfangreichen Daten zu ihren Programmen, Sport, Tierwelt, Musik und Radio. Die deutsche Nationalbibliothek bietet ebenso eine RDF-Schnittstelle wie die UK National Archives. Zahlreiche Bibliotheksverbünde veröffentlichen ihre Daten RDF-basiert und immer mehr Forschungsprojekte setzen Semantic Web Technologien ein und stellen ihre Ergebnisse als Linked Open Data zur Verfügung.

Die Technische Universität Braunschweig entwickelt das ontologiebasierte Terminologiemanagementsystem *iglos* (intelligentes Glossar), das Fachsprache in interdisziplinären Teams visuell modellieren und so Missverständnisse minimieren kann.[21] Die Humboldt-Universität zu Berlin baut in ihrem Exzellenzcluster *Bild Wissen Gestaltung* eine RDF-basierte Ontologie auf, die sehr heterogene wissenschaftliche Arbeitsprozesse modellieren und dokumentieren kann. Ziel ist es, mit Hilfe dieser Ontologie zu einem besseren Verständnis von interdisziplinärer Forschung und ihren Schwierigkeiten zu kommen und die wesentlichen Faktoren für ihr Gelingen zu identifizieren.[22]

Eine Vielzahl von Tools ist in diesem Kontext entstanden und kann zuallermeist kostenlos genutzt werden. Verbreitet sind beispielsweise der Ontologie-Editor *Protegé*[23] der Stanford University oder der Java-basierte Triplestore *Sesame*[24]. Für die Einbindung in alle etablierten Programmiersprachen stehen Frameworks zur Verfügung. Ausgereifte kommerzielle Produkte und professionelle Unterstützung finden sich aber ebenso.

Der Einsatz von Semantic Web Technologie eignet sich besonders für Bibliotheken, Archive und Museen. Hier kann auf viel domänenspezifische Vorarbeiten zurückgegriffen werden. Da für diese Institutionen die Organisation von Information und deren Erschließung für breite Nutzerkreise eine zentrale und komplexe Aufgabe ist, gelangen statische Metadatenschemata an ihre Grenzen.

Die Europeana macht vor, was auf diesem Gebiet möglich ist. Ihre Aufgabe ist es, das vielfältig distribuierte Kulturerbe Europas mit einer zentralen digitalen Schnittstelle zu versehen und zu erschließen. Dazu integriert sie die Metadaten von z. Z. nahezu 150 Providern aus Bibliotheken, Museen, Galerien, Archiven, Projekten und Unternehmen aus ganz Europa.[25] Da diese verschiedenen Datenlieferanten sehr unterschiedliche Inhalte verwalten, liegen auch sehr verschiedene, häufig spezialisierte Datenformate der Ursprungsdaten vor. Die Europeana erstellt dazu ein Mapping, das diese verschiedenen Formate im RDF-basierten Europeana Data Model zusammenführt, ohne die dahinterliegenden Spezialisierungen zu verlieren (EDM steht für Europeana Data Model[26]). Das EDM basiert auf den Vokabularen OAI ORE (Open Archives Initiative Object Reuse & Exchange), Dublin Core, SKOS (Simple Knowledge Organization System) und CIDOC CRM. Darüber hinaus erlaubt das EDM, domänenspezifische Spezialisierungen vorzunehmen, wie beispielsweise im Projekt DM2E (Digital Manuscripts for Europeana), das Spezialisierungen für Manuskripte ergänzt.[27] Alle in EDM vorliegenden oder dahin konvertierten Metadaten werden über die Europeana-Webapplikation aggregiert, wobei das jeweilige CHO (Cultural Heritage Object) im Sinne eines objektorientierten Ansatzes als Basis gewählt wird.

18 http://www.researchspace.org.
19 http://www.cidoc-crm.org.
20 https://thepund.it.
21 http://www.iglos.de.
22 https://www.interdisciplinary-laboratory.hu-berlin.de/de/basisprojekte/architekturen-des-wissens.
23 http://protege.stanford.edu.
24 http://www.openrdf.org.
25 http://europeana.eu/portal/europeana-providers.html.
26 http://pro.europeana.eu/edm-documentation.
27 Dröge, Evelyn; Iwanowa, Julia; Hennicke, Steffen: A Specialisation of the Europeana Data Model for the Representation of Manuscripts: The DM2E model. In: Libraries in the Digital Age (LIDA) 13 (2014).

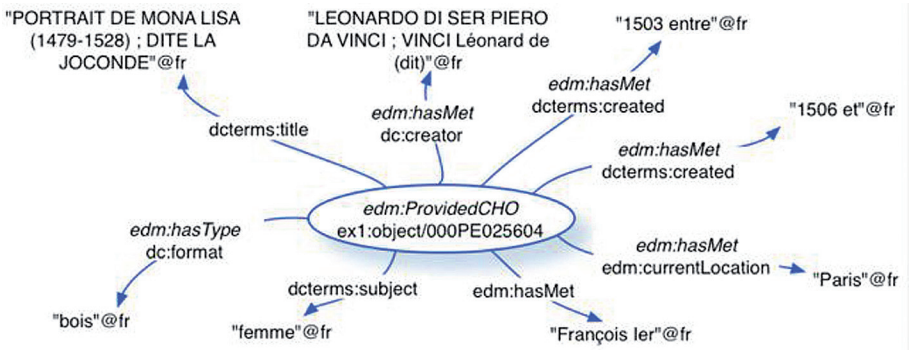

Abbildung 1: Ausschnitt aus einer exemplarischen EDM-Modellierung, die den objektorientierten Ansatz demonstriert.[28]

So wird eine Semantic Web Application zur Verfügung gestellt, die heterogene Datenformate und verteilte Originaldaten semantisch zusammenführt, vernetzt und erschließt.

Gerade das Browsen zwischen vernetzen Objekten und Informationsträgern, das zielgerichtete Filtern, die Nachnutzung bestehender Daten und die Verknüpfung mit externen Daten und anderen Anbietern sind mit semantischen Technologien sehr gut möglich. Auch moderne Suchmaschinen, die sich eher an natürlicher Sprache anlehnen als an reinen Keywords, profitieren stark von semantischen Netzen. Googles Knowledge Graph ist ein Beispiel dafür, auch wenn er intern nicht mit RDF modelliert ist.[29]

Bei allen Vorteilen muss jedoch auch bedacht werden, dass der Umgang mit semantischen Technologien im Detail nicht trivial ist. Gerade wenn es um die Nutzung komplexer Ontologien geht, darf die Einarbeitungszeit nicht unterschätzt werden. Auch gehört der Umgang mit den dazugehörigen Sprachen und Technologien nicht unbedingt zur regulären Ausbildung eines Entwicklers und muss u.U. neu erlernt werden. Die Vielzahl von Semantic-Web-Projekten in unterschiedlichsten Disziplinen und Communities macht einen guten Überblick verhältnismäßig schwierig. Für die Definition einer eigenen Ontologie, die an die spezifischen Bedürfnisse angepasst ist, müssen in der Regel eine Vielzahl existierender Ontologien auf mögliche Nachnutzbarkeit geprüft werden, um möglichst keine Redundanzen zu erzeugen. Das gestaltet sich nicht immer übersichtlich, auch wenn Verzeichnisse wie die Linked Open Vocabularies dabei helfen können.[30]

Wer jedoch die Chance hat, über das Semantic Web Expertise aufzubauen und die eigenen Informationsstrukturen an die nächste Generation des Webs anzupassen, wird daran viel lernen und nachhaltig profitieren. Die Informationsflut – das merken wir alle tagtäglich – steigt ständig an. Nur wer in der Lage ist, die Intelligenz der Maschinen wirksam für Filterung, Vernetzung und Strukturierung einzusetzen, kann seine Wissensschätze für die Zukunft öffnen.

Dr. Christian Stein
Exzellenzcluster Bild Wissen Gestaltung
Unter den Linden 6
D-10099 Berlin
christian.stein@hu-berlin.de

28 EDM Primer, http://pro.europeana.eu/edm-documentation.

29 Eder, Jeffrey Scott: Knowledge graph based search system. In: U. S. Patent Application 13/404,109. URL: http://www.google.com/patents/US20120158633.

30 http://lov.okfn.org/dataset/lov.