Corporate Semantic Web

Wie semantische Anwendungen in Unternehmen Nutzen stiften

Bernhard Humm, Börteçin Ege, Anatol Reibold

Hochschule Darmstadt, Technische Universität Wien, Cogia GmbH

1. Semantische Anwendungen haben Business Potential und stiften bereits heute in Unternehmen Nutzen, z.B. in den Branchen Telekommunikation, Logistik, verarbeitende Industrie, Energie, Medizin, Tourismus, Bibliotheks- und Verlagswesen und Kultur.
2. Zahlreiche Linked Data Sets stehen zur Verfügung, aber die fachliche Passung zum Anwendungsproblem muss sorgfältig geprüft werden. Manchmal ist die anwendungsspezifische Modellierung empfehlenswert.
3. Semantic Web Technologien sind reif für den Unternehmenseinsatz, aber je nach Anforderungen können auch andere Technologien für die Entwicklung semantischer Anwendungen empfohlen werden.
4. Semantische Anwendungen zu entwickeln erfordert zusätzlich zum klassischen Software Engineering besondere Fähigkeiten, z.B. die eines Knowledge Engineers.

Das Semantic Web

Semantic Web ist seit der Prägung des Begriffs 2001 durch Tim Berners Lee ein viel benutzter Begriff. Die Vision ist, Daten über Anwendungs-, Firmen- und Ländergrenzen hinweg auszutauschen und wiederzuverwenden. Das World Wide Web Consortium (W3C) hat dazu in den letzten Jahren als Rahmenwerk für das Semantic Web umfangreiche Standards, insbesondere RDF, RDFS, OWL und SPARQL[1], etabliert. Darüber hinaus wurden professionelle Werkzeuge für die Ontologie-Entwicklung sowie für die Speicherung und für den Zugriff auf semantische Daten entwickelt und zur Reife gebracht.

Organisationen weltweit haben zahlreiche so genannte „Linked Open Data Sets" erstellt und veröffentlicht. Die Grundlagen von Semantic Web sind also

[1] Für eine praxisnahe Einführung in Semantic Web Technologien siehe z.B. [1]

Aus Börteçin Ege, Bernhard Humm, Anatol Reibold (Hrsg.): "Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften". Springer-Verlag, 2015.

etabliert und fundiert beschrieben. Auch mangelt es nicht an visionären Artikeln. Überraschenderweise gibt es jedoch wenig Literatur über semantische Anwendungen, die in Unternehmen bereits eingesetzt werden und Nutzen stiften.

Woran liegt das? Gibt es noch gar keine solchen Anwendungen?

Semantische Anwendungen im Unternehmenseinsatz

Tatsächlich gibt bereits zahlreiche semantische Anwendungen, die in Unternehmen bzw. Organisationen verschiedenster Branchen im Einsatz sind, z.B. Telekommunikation, Logistik, verarbeitende Industrie, Energie, Medizin, Tourismus, Bibliotheks- und Verlagswesen und Kultur. Von solchen Anwendungen handelt dieses Buch. Abbildung 1 gibt eine Übersicht über die Branchen und Themenbereiche der in diesem Buch beschriebenen semantischen Anwendungen.

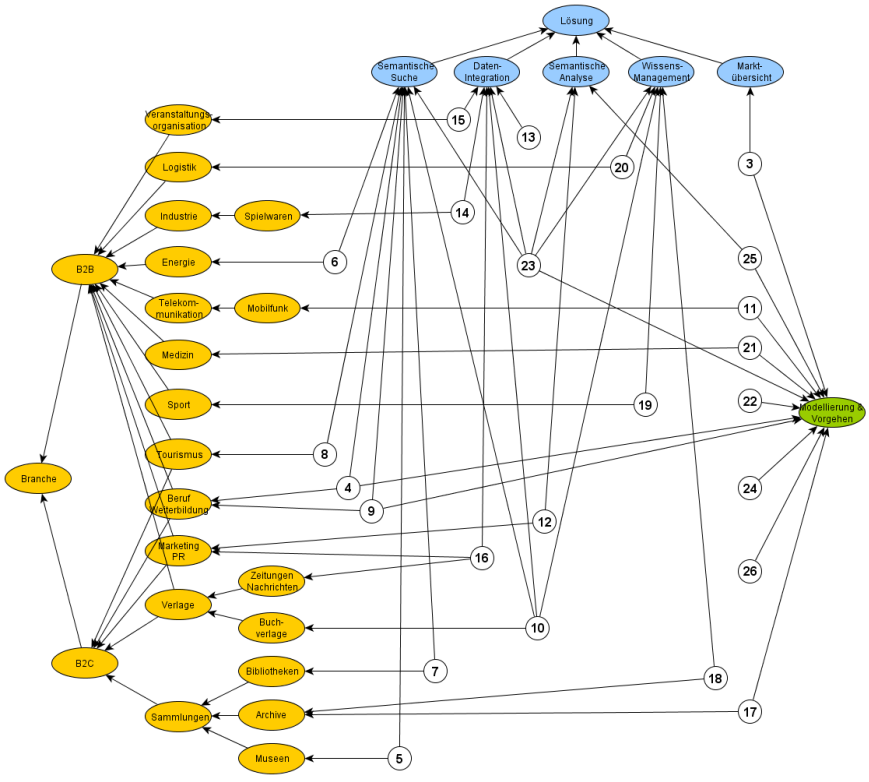

Abbildung 1: Branchen und Themenbereiche der im Buch beschriebenen semantischen Anwendungen. Die Zahlen sind Kapitelnummern (Graphik von Thomas Hoppe).

Die Autoren der einzelnen Kapitel haben umfangreiche Erfahrungen in der Entwicklung von semantischen Anwendungen gemacht – positive wie negative. Sie berichten über Software-Architektur, Methodik, Technologieauswahl, Linked Open Data Sets, Lizenzfragen etc. Manche dieser Erfahrungen und die daraus resultierenden Empfehlungen mögen überraschen – umso wichtiger, dass sie von Praktikern für Praktiker aufgeschrieben sind. Nachfolgend fassen wir wesentliche Erkenntnisse zusammen.

Bereitstellen von Linked Data reicht nicht

Eine der Kernideen des Semantic Web und von Linked Data ist es, das Rad nicht stets neu zu erfinden, sondern Data Sets zwischen Communities auszutauschen und wiederzuverwenden. Damit das immer besser gelingt, stellt Tim Berners Lee Prinzipien für die Veröffentlichung von Linked Data vor [2]. Diese enthalten die Empfehlungen, HTTP-URIs zur Bezeichnung von Objekten zu verwenden, Standards wie RDF und SPARQL zu verwenden sowie zusätzlich zu Objekt-Informationen auch Links auf andere Objekte anzugeben.

Die Festlegung auf Standards ist ein wichtiges Fundament und Design-Prinzipien sind hilfreich, aber reicht das? Wer semantische Anwendungen entwickelt und dabei auf existierende, veröffentlichte Data Sets zugreift, ist mit weiter gehenden Fragestellungen konfrontiert – selbst wenn die o.g. Prinzipien eingehalten wurden.

Da ist zunächst die Frage der *Datenqualität*. Diese kann niemals absolut beantwortet werden, sondern ist immer in Relation zu einem konkreten Nutzungskontext zu sehen. Ein Beispiel: Wie ist die Datenqualität von DBpedia in Bezug auf Künstler und ihre Zuordnung zu kunstgeschichtlichen Epochen? Beispielsweise liefert die Anfrage nach Renaissance-Künstlern nur 19 Resultate[2]. Michelangelo ist vorhanden, Raffael sucht man jedoch vergebens. Statt dessen findet man auch die Resource dbpedia:Leonardo_da_Vinci's_personal_life. Es scheint also, dass derzeit die Datenqualität von DBpedia für semantische Anwendungen im Bereich der Kunstgeschichte nicht ausreichend ist.

Die zweite Frage bezieht sich auf den *Modellierungskontext*. Bowker and Star schreiben ganz richtig in [3], dass Klassifikationen, die in einem Kontext wunderbar natürlich und homogen erscheinen, in einem anderen Kontext gezwungen und uneinheitlich wirken. Es gibt halt nicht „die richtige“ Modellierung eines Sachgebiets – es hängt vom Nutzungskontext ab.

---

[2] SPARQL Query `select ?p where {?p rdf:type yago:RenaissanceArtists}` ausgeführt auf dem SPARQL Endpoint http://dbpedia.org/sparql am 15.7.2014.

Bei der Integration mehrerer Data Sets stellt sich die Frage der *Interoperabilität*. Dies ist eine Frage der der Modellierung, die sich auch stellt, wenn die oben genannten Prinzipien strikt eingehalten wurden. Betrachten wir beispielsweise zwei Thesauri, beide in RDFS modelliert, aber von unterschiedlichen Teams nach unterschiedlichen Modellierungsvorschriften. Im einen Thesaurus sind Begriffe als Instanzen modelliert und die Beziehung zwischen Ober- und Unterbegriffen mittels SKOS [4] (skos:broader bzw. skos:narrower); im anderen Thesaurus sind Begriffe als Klassen modelliert und dieselbe semantische Beziehung über eine andere Property, nämlich rdfs:subclassOf. Werden beide Thesauri integriert, so sind einfache SPARQL-Abfragen über den integrierten Thesaurus nicht mehr möglich. Weitaus schwerwiegender als diese technischen Interoperabilitätsprobleme sind jedoch die fachlichen: Fragen der Redundanz, Konsistenz, Koheränz, des Mappings und der unterschiedlichen Vollständigkeitsgrade der integrierten Thesauri.

Eine global vernetzte Wissensbasis – Fiktion oder Realität?

Zu vielen semantischen Data Sets werden SPARQL Endpoints zur Abfrage bereit gestellt. Beispiele sind DBpedia, DBLP und Gene Ontology Database. Außerdem erlaubt der SPARQL-Standard den Zugriff auf verschiedene Endpoints und die Integration der Ergebnisse. Die Idee ist daher naheliegend, in semantischen Anwendungen auf die lokale Datenhaltung zu verzichten und statt dessen die Daten über verteilte SPARQL Queries zu integrieren – ganz im Sinne einer serviceorientierten Architektur. Das spart lokalen Speicherplatz, erübrigt die Installation eines RDF Triple Stores und garantiert einen stets aktuellen Datenbestand.

In semantischen Anwendungen, die im Praxiseinsatz sind, wird diese Form der Integration jedoch selten gewählt. In diesem Buch werden 18 semantische Anwendungen vorgestellt, die in Unternehmen eingesetzt werden. Lediglich zwei davon greifen über SPARQL Endpoints auf verteilte Data Sets zu.

Wie kommt das? Zum Einen erschweren die oben erwähnten Problemfelder Datenqualität, Modellierungskontext und Interoperabilität eine einfache Datenintegration über SPARQL-Endpoints. Dazu kommen Fragen der Performanz und Verfügbarkeit. Hier gilt, dass die Kette nur so stark wie ihr schwächstes Glied ist. Ist nur einer von vielen Endpoints nicht verfügbar, so steht u.U. die ganze Anwendung. Und Performanz aktueller SPARQL-Endpoints ist meist nicht ausreichend für Online-Anwendungen, in denen Endnutzer-Antwortzeiten von höchstens einer Sekunde gefordert sind.

Aus Börteçin Ege, Bernhard Humm, Anatol Reibold (Hrsg.): "Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften". Springer-Verlag, 2015.

Semantik = RDF?

Eine wesentliche Voraussetzung für den Erfolg der Idee des Semantic Web ist die konsequente Standardisierung von Sprachen wie RDF, RDFS, OWL und SPARQL, die wesentlich vom W3C vorangetrieben wurde. Sie erlaubt erst die verteilte, unabhängige Entwicklung von integrierbaren Data Sets im Rahmen eines Community Prozesses. Sie erlaubt auch erst die Entwicklung von Werkzeugen, z.B. Ontologie-Editoren und RDF Triple Stores, deren Daten austauschbar sind.

Viele semantische Anwendungen verwenden direkt diese Standards und Werkzeuge. Von den in diesem Buch vorgestellten 18 Anwendungen verwenden 8 RDF Triple Stores. Das bedeutet aber auch im Umkehrschluss, dass in etwa der Hälfte der Anwendungen bewusst auf andere Technologien gesetzt wird. Beispiele für die Nutzung anderer Technologien finden sich bei der semantischen Suche für Bibliotheken und Museen, Analyse von Marktdaten und einem Semantic Social Network für Sport-Trainer.

Sind das damit keine semantischen Anwendungen? Wird eine Anwendung semantisch, wenn ein Semantic Web Standard des W3C verwendet wird, und verliert sie diese Eigenschaft, wenn sie dies nicht der Fall ist? Aus unserer Sicht gilt dies keinesfalls. Eine Anwendung kann als semantisch bezeichnet werden, wenn die *Bedeutung* von Inhalten eine wesentliche Rolle spielt. Das betrifft den Anwender und ist unabhängig von der eingesetzten Technologie.

Warum entscheiden sich Architekten solcher semantischer Anwendungen bewusst, bei der Entwicklung auf andere Technologien zu setzen? Dafür gibt es mehrere Gründe. Zum einen können Performanz-Gründe eine Rolle spielen, wenn es sich um sehr große Datenmengen und gleichzeitig hohen Anforderungen an die Antwortzeiten handelt. Weiterhin kann die Funktionalität von RDF Triple Stores für den Anwendungsfall nicht ausreichend sein. Beispiel sind die Möglichkeiten für Volltextsuche, Teilwortsuche, phonetische Suche oder Toleranz gegenüber Tippfehlern. Die RDF-Technologie kann auch das Know-How des Entwicklerteams übersteigen und der Know-How-Aufbau kann als zu kostspielig eingeschätzt werden. In diesem Fall können einfachere Technologien eingesetzt werden, wenn sie dem Anwendungsfall angemessen sind. Sich gegen Semantic Web Technologien zu entscheiden kann aber auch Nachteile haben, z.B. ein erschwerter Austausch von Daten oder die Bindung an eine proprietäre Technologie.

Richtig vorgehen

Wie in jedem Entwicklungsprojekt steht am Anfang der Entwicklung einer semantischen Anwendung die Kundenerwartung. Wer sind die Anwender? Sind sie unternehmensintern oder Endkunden? Was sind die Anwendungsfälle? Und welche

Fragen soll die Anwendung beantworten können. Es empfiehlt sich, solche Kompetenzfragen konkret zu sammeln und aufzuschreiben, z.B. „Wie verteilen sich die Fördergelder für Solarkraftwerke auf die Bundesländer?“.

Gerade für semantische Anwendungen gilt, dass sich zukünftige Anwender vorab gar nicht genau vorstellen können, wie die Anwendung funktionieren soll. Daher ist meist ein agiles Vorgehen empfehlenswert: klein anfangen, rasch Prototypen entwickeln, intensiv die Anwender einbeziehen und dann inkrementell ausbauen.

In Ergänzung zu den Rollen in klassischen Software-Entwicklungsprojekten wird die Rolle des Knowledge Engineers wichtig. Er ist verantwortlich für die Akquisition und Formalisierung des Wissens in Zusammenarbeit mit den Fachexperten. Meist geschieht dies in Form von Interviews. Dabei empfiehlt es sich, die Fachexperten in Form von „narrativen Stories“ frei über ihre Domäne berichten zu lassen und diese Berichte zu dokumentieren. Die Arbeit am Modellierungswerkzeug übernimmt dann meist der Knowledge Engineer selbst. Die Erfahrung zeigt, dass Fachexperten selten mit Modellierungswerkzeugen umgehen können. Häufig verstehen sie nicht die Notation, Konzepte und Vorgehensweisen und zeigen auch keine Bereitschaft, sich dieses Wissen anzueignen. Falls Fachexperten doch Werkzeuge benutzen sollen, dann sollten diese Werkzeuge in der Bedienung so einfach wie möglich sein, z.B. ein Spreadsheet, welches in der täglichen Arbeit ohnehin vertraut ist.

Wichtig ist auch die Qualitätssicherung durch Zurückspielen des entstehenden Modells an die Fachexperten. Wie kann aber Qualitätssicherung des Modells funktionieren, wenn die Fachexperten die Modellnotation nicht verstehen? Hier haben sich Visualisierungswerkzeuge bewährt, wie z.B. hyperbolische Bäume.

Modellieren ist einfach (?!)

Make or buy? Diese Frage stellt sich bei der Entwicklung semantischer Anwendungen nicht nur für technische Produkte und Werkzeuge, sondern auch für Data Sets. Linked Open Data Sets beinhalten hundertausende von Begriffen, in jahre- oder jahrzentelanger Arbeit von Experten formalisiert und qualitätsgesichert. Diesen Schatz gilt es zu heben.

Aber dennoch zeigt es sich in der Praxis immer wieder, dass keines der öffentlichen Data Sets genau auf das aktuelle Problem passt. Und dann muss man sorgfältig die Alternativen abwägen, (a) mit den Schwächen zu leben, (b) verschiedene öffentliche Data Sets zu kombinieren, (c) diese anzureichern – evt. im Rahmen eines Community Prozesses – oder (d) doch ein anwendungsspezifisches Data Set neu zu entwickeln. In der jetzigen relativ frühen Phase der Technologie hat sich auch bewährt, mit dem Hersteller eines Data Set in Kontakt zu treten, um evtl. Anpassungen direkt zu besprechen.

Aus Börteçin Ege, Bernhard Humm, Anatol Reibold (Hrsg.): "Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften". Springer-Verlag, 2015.

Die Entwicklung eines Data Sets fordert das Know-How des Knowledge Engineers. Wichtig ist, dass die Modellierung nicht im luftleeren Raum, sondern stets im Kontext einer konkreten Fragestellung erfolgt. Beispielsweise muss eine Ontologie für die semantische Suche von Hotels Konzepte wie Ausstattungsmerkmale, Sehenswürdigkeiten etc. umfassen. Sie braucht aber keine Aussagen zu beinhalten wie „Im Hotel arbeiten Menschen", „Menschen (Homo Sapiens) gehören zur Ordnung der Primaten, zur Klasse der Säugetiere, etc., haben ein Herz, eine Lunge, eine Leber etc. etc." Diese Aussagen – obwohl wahr – in die Ontologie aufzunehmen, wäre nicht nur unnötig, sondern im Sinne der Anwendung sogar falsch. Sie nutzen nicht dem Anwendungszweck und verursachen bei der Entwicklung und Pflege nur unnötige Kosten.

Wichtig ist also nicht nur, zu wissen, was noch fehlt, sondern auch zu wissen, wann man aufhören kann („Mut zur Lücke"). Werden solche Hinweise beachtet, dann muss Modellieren gar nicht so aufwändig sein, wie gemeinhin vermutet. Mit einem eingespielten Team ist die Modellierung von 4000 Begriffen in ca. 20 Bearbeitertagen realistisch.

Juristische Fragen

Sollen Linked Data Sets in eine semantische Anwendung eingebunden werden, dann stellt sich die Frage, ob die entsprechenden Lizenzen dies auch erlauben. Vernetzte Datensätze und Werke, die auf diesen Datensätzen aufbauen, sind nach dem Immaterialgüterrecht geschützt. Die Datensätze, aus denen sich Linked Data speist, bedienen sich entsprechend ihrer Offenlegungspolitik einer Vielzahl an Lizenzmodellen mit teils sehr unterschiedlichen Terminologien, Geltungs- und Gültigkeitsbereichen. Bei der Wiederverwendung „offener" Daten ist auf die Kompatibilität der Lizenzbedingungen der Datensätze zu achten. Lizenzkonflikte können die Wiederverwendung einschränken und zu juristischen Folgeproblemen führen.

Semantische Anwendungen stiften Nutzen in Unternehmen – nachweislich!

Unternehmen werden nach betriebswirtschaftlichen Gesichtspunkten geführt. Nur was sich lohnt, wird auch gemacht. Semantische Technologien sind noch relativ neu und Know-How-Aufbau kostet Zeit und Geld. Lohnt sich diese Investition? Steht den Kosten ein ausreichender Nutzen semantischer Anwendungen gegenüber? Kann dieser Nutzen gemessen und nachgewiesen werden?

Aus Börteçin Ege, Bernhard Humm, Anatol Reibold (Hrsg.): "Corporate Semantic Web: Wie semantische Anwendungen in Unternehmen Nutzen stiften". Springer-Verlag, 2015.

Die Praxis zeigt, dass eine seriöse Quantifizierung des Nutzens semantischer Anwendungen – hier: semantische Suche – gar nicht so einfach ist. Klassische Maße wie Precision und Recall aus dem Information Retrieval, die auf einem Goldstandard basieren, sind nicht geeignet. Semantische Suche liefert einerseits weniger, dafür aber genauere Treffer; andererseits aber auch zusätzliche Treffer. Die Bewertung einer semantischen Suche muss diesen vermeintlichen Widerspruch auflösen. Daher wird ein neues Maß für die quantitative Bewertung der Effizienzsteigerung bei der Suche empfohlen. Empirische Vergleiche zeigen, dass durch semantische Suche Effizienzsteigerungen seitens des Benutzers in der Größenordnung von 10-15% möglich sind.

Fazit

Semantische Anwendungen stiften schon heute in Unternehmen Nutzen. Die zugrunde liegenden Technologien haben die dafür notwendige Reife erreicht. Auch sind in den letzten Jahren umfangreiche Linked Data Sets entstanden. Die Entwicklung semantischer Anwendungen erfordert aber Know-How in Ergänzung zum klassischen Software Engineering Know-How.

Literatur

[1] Dean Allemang, James Hendler: "Semantic Web for the Working Ontologist - Effective Modeling in RDFS and OWL" 2nd Edition. Morgan Kaufmann Publishers Inc. San Francisco, CA, USA 2011, 978-0-12-385965-5

[2] Tim Berners-Lee: Linked Data. http://www.w3.org/DesignIssues/LinkedData.html . Abgerufen 15.7.2014

[3] Bowker, G. C. and S. L. Star. Sorting Things Out: Classification and its consequences. Cambridge, MIT Press1999

[4] SKOS (Simple Knowledge Organization System): www.w3.org/2004/02/skos/ . Abgerufen 15.7.2014