Fünf Semantic Web Paradigmen

SEMINAR PAPER

25. Oktober 2014

Simon Heimler
heimlersimon@gmail.com
Master of Applied Research in Computer Science

Informationssysteme
Prof. Dr. Sabine Müllenbach
University of Applied Sciences Augsburg

Abstract

Üblicherweise wird das Thema Semantic Web anhand des Semantic Web (Technologie) Stacks (Tim Berners-Lee 2009) erklärt. Dieser Artikel will einen alternativen Weg gehen: Nicht die technische Umsetzung steht im Vordergrund, sondern die Konzepte und Paradigmen hinter ihnen.

Dazu wurde eine subjektive Auswahl von fünf Semantic Web Paradigmen getroffen: (1) Mensch-Computer Kooperation, (2) die Mächtigkeit von Links (3) Graphenstruktur, (4) die Trennung von Fakt und Interpretation (5) und die Open World Assumption (Offene-Welt-Annahme).

Keines dieser Paradigmen ist ausschließlich im Semantic Web Kontext zu finden. Im Gegenteil: Alle diese Konzepte haben ihren Ursprung in anderen Disziplinen. Selbst wenn der Leser also nicht die ganze Semantic Web Vision „kauft", können diese Paradigmen dennoch eine wertvolle Bereicherung sein, da sie interessante Lösungen zu aktuellen Problemen aufzeigen.

Wenn man die Paradigmen zusammenführt bilden sie (zusammen mit weiteren) die Grundlagen der Semantic Web Vision.

Keywords: Semantic Web, Linked Data, Mensch-Computer-Kooperation, Open World Assumption, Graphenstruktur

Abkürzungsverzeichnis

SW	Semantic Web
MCK	Mensch-Computer-Kooperation
OWA	Open World Assumption

1 Einleitung

Ein bekanntes Zitat, dass nachträglich diversen Autoren zugeschrieben wird, besagt: "*Wer als Werkzeug nur einen Hammer hat, sieht in jedem Problem einen Nagel.*" Der effizienten Lösung von Problemen tut dies natürlich nicht gut. Wie kann man diesem Problem also entgehen?

Die Lösung liegt nahe: Man beschäftigt sich mit anderen, neuen Werkzeugen. Das Wort Werkzeug ist hier natürlich metaphorisch gesprochen: Oberflächlich gesehen könnte damit das einem zu Verfügung stehende Handlungsrepertoire gemeint sein. Doch hinter diesem stehen unsere (oft festgefahrene) Paradigmen, also Denkweisen und Weltanschauungen (die unser Handeln ursächlich bestimmen.

Diese Argumentation sollte auch darlegen warum sich dieser Artikel mehr auf das „Warum?" konzentriert und weniger auf das „Wie?". Die konkreten technischen Werkzeuge ändern sich. Doch die grundsätzlichen Ideen hinter ihnen sind beständiger und für das Verständnis wichtiger.

Das Ziel dieses Artikels ist es also dem Leser einige Paradigmen aus der Semantic Web Community vorzustellen. Sie zeigen alternative Herangehensweisen zu aktuellen Problemen auf, die bisher relativ unbekannt sind.

2 Vorstellung der Paradigmen

2.1 Mensch-Computer-Kooperation

Es ist heute nicht mehr zu verleugnen, dass es einige Gebiete gibt in denen Computer erheblich besser und effizienter arbeiten als Menschen. Aber man kann auch umgekehrt argumentieren: Trotz allen Fortschritten in der KI Forschung gibt es viele Bereiche in denen Menschen nicht ersetzt werden können – oder sollten.

Das Konzept der Mensch-Computer-Kooperation bietet eine weitere Sichtweise: Menschen haben bestimmte Stärken und Schwächen - und Computer die ihre. Wenn beide Seiten miteinander auf produktive Weise zusammenarbeiten und dies berücksichtigt wird, kann das Ergebnis beide Seiten im Alleingang weit übertreffen. (Sankar)

Das Semantic Web kann als Mensch-Computer-Kooperation Initiative für das Web verstanden werden. Wenigen wird dies bewusst sein, doch das aktuelle Web ist für Menschen optimiert und für Maschinen oft nur sehr schwierig und missverständlich zu interpretieren.

Um eine produktivere Kooperation zu ermöglichen müssen Webseiten also so verfasst und erstellt werden, dass sie sowohl von Menschen als auch Maschinen gut verstanden werden können. Metadaten lösen dieses Problem nicht, da sie nur die Dokumente und Dateien an sich beschreiben, nicht aber dessen Inhalt.

Das Semantic Web will dieses Problem durch semantische Annotation lösen. Semantisch bedeutet in diesem Kontext, dass die tatsächliche *Bedeutung* und der Inhalt auf maschinenlesbare Weise erfasst werden. Dies ist natürlich ein sehr ambitioniertes Vorhaben und es bleibt noch abzuwarten in welchem Umfang und Qualität es sich erfüllen wird.

Warum sollte man sich also diesen Aufwand machen?

2.1.1 Aktueller Stand

Semantische Annotation ist der Teil des Semantic Webs der bis jetzt die beste Annahme gefunden hat. Viele große Webunternehmen haben die letzten Jahre begonnen semantische Annotationen zu fördern und zu verarbeiten. Damit wird die Annotation für Webentwickler interessant bis wichtig, da sie bessere SEO und Integration mit diesen Anbietern verspricht. Die Nutzer profitieren dann wiederrum von den neuen Angeboten und Services die aufgrund dieser Technologie möglich oder besser geworden sind.

Schema.org (Google Inc., Yahoo Inc., Microsoft Corporation and Yandex) ist Projekt von Google, Yahoo, Microsoft, etc., dass ein gemeinsames Vokabular für die Beschreibung von Dingen und Vorgängen im Internet definiert. Facebook hat einen eigenen Standard entwickelt, den OpenGraph (Facebook October 20th, 2014).

Auch Forschung und Industrie haben die Semantic Web Technologien aufgenommen um übergreifende Standards zu schaffen, wie etwa das SKOS Modell (Antoine Isaac, Ed Summers 2009) zur Verwaltung von Wissensbeständen.

2.1.2 Technischer Hintergrund

Aktuell gibt es viele unterschiedliche Semantic Web Datenserialisationsformate, einige von ihnen vom W3C standardisiert. Größere Verbreitung haben RDFa (Manu Sporny 2013) und Microdata (Ian Hickson 2014), die beide auf XML basieren. Steigende Verbreitung hat aktuell JSON-LD (Manu Sporny et al 2014), das auf dem einfachen JSON Datenformat aufbaut. Als textbasiertes Format ist Turtle (Gavin Carothers und Eric Prud'hommeaux 2013) zu erwähnen.

Alle diese Formate teilen sich ein zugrundeliegendes Daten-Konzept: RDF (Eric Miller und Frank Manola 2004). Auf dieses wird im Abschnitt 2.2**Fehler! Verweisquelle konnte nicht gefunden werden.** tiefer eingegangen.

2.2 Hyperlinks

Hyperlinks sind nicht erst mit dem Web erfunden worden. Sie haben eine lange Geschichte die mindestens bis zur Memex zurückreicht und waren wichtiger Bestandteil der in den 60er Jahren entstehenden Hypertext Systeme (Nelson).

Wenn heute von Links gesprochen wird, meinen wir meist nur das Ziel dieser Verknüpfung, die URI (Tim Berners-Lee et al. 2014) (Unique Resource Identifier): Eine eindeutige Adresse die eine Ressource identifiziert. Doch ein Link besteht aus mehr Komponenten: Mindestens eine Quelle und ein Ziel. Meist kommt noch eine Relation, die die Beziehung zwischen beiden definiert, hinzu.

Der Erfinder des WWW, Tim-Berners Lee entschied sich dafür eine vereinfachte Version der Links zu verwenden: Links sind unidirektional, die Quelle ist implizit (die Seite auf der der Link angegeben wird) und die Relation zwischen Quelle und Ziel spielt keine große Rolle.

Dadurch waren Links im Web einfacher als der meisten konkurrierende Hypertext Systeme. Die unidirektionale Natur der Verknüpfungen ermöglichte die dezentrale Struktur des Internets, da keine zentrale Datenbank aller Verknüpfungen nötig ist und Links ohne vorherigen Konsens gesetzt werden können. Rückblickend wird diese Entscheidung mit für den Erfolg des Webs verantwortlich gemacht. (Hendrik Arndt 2006, S. 153–154)

Die Links im WWW sind also eine bewusst vereinfachte Implementierung. Das Semantic Web setzt an dieser Stelle ein und baut das Konzept der Links und URIs weiter aus:

Links können als Tripels verstanden werden: Quelle, Relation und Ziel. Im Semantic Web wird daraus das Konzept von RDF, dass diese in eine minimale grammatische Aussage umformuliert: Subjekt, Prädikat und Objekt. („Franz Meier“ „hat Sohn“ „Rudolf Meier“.) Die Relation bekommt damit eine zentrale Rolle: Erst durch sie bekommt die Aussage auch eine semantische Bedeutung.

URIs werden im Semantic Web allerdings nicht nur verwendet um Adressen zu Webseiten anzugeben. Sie können auch auf abstrakte Dinge oder Beziehungen verweisen. Der Begriff „Resource“ wird hier also ausgeweitet. Dadurch ist es möglich allgemein über Dinge, Konzepte und Beziehungen zu reden und dennoch eine eindeutige Referenz zu haben, die auch über Grenzen, wie verschiedene Webseiten oder Datenbanken, hinweg funktionieren kann.

Somit wird aus einem einfachen, alltäglichen Konzept des aktuellen Webs die Grundlage für eine maschinenlesbare Sprache und damit für die in Abschnitt 2.1 erwähnte semantische Annotation.

2.3 Graphen- und Netzwerkstrukturen

Es gibt eine Vielzahl von Datenstrukturen: Von einfachen Listen über Tabellen und Baumstrukturen hin zu Graphen. Die komplexeren Datenformate sind in der Regel die Obermenge der einfacheren. Graphen stehen in der Hierarchie ganz oben und können alle einfacheren Datenformate abbilden und verlustfrei integrieren.

Ein einfaches Gedankenexperiment zeigt dies: Sobald zwei Baumstrukturen miteinander verbunden werden erhält man entweder zwei getrennte Bäume die keine Gemeinsamkeiten haben oder aber die Baumstruktur wird durch zirkuläre Referenzen aufgelöst und man erhält einen Graphen (Hitzler 2007, S. 43). Umgekehrt kann jedoch jede Baumstruktur in einem Graphen abgebildet werden, da ein Baum nur eine spezielle Form eines Graphen ist.

Man könnte die These aufstellen, dass die meisten Strukturen in der Welt graphenorientiert sind, wie etwa das Gehirn, soziale Gefüge, das Ökosystem. Aus Gründen der Vereinfachung und Abstraktion werden sie auf Bäume und einfachere Strukturen bewusst (oder unbewusst) reduziert.

Doch diese Reduktion kann auch problematische Seiten haben: Es werden Details verloren und die Reduktion auf einen kleineren Nenner zwingt Design und Strukturentscheidungen auf, die meist subjektiv sind. Es müssen oft Einschränkungen in Kauf genommen werden und nachträgliche Änderungen an der Struktur können sehr aufwendig werden.

Deswegen kann es sehr sinnvoll sein die Komplexität einer Graphenstruktur in Kauf zu nehmen. Dies macht vor allem Sinn wenn die Daten ihrer Natur nach graphenorientiert sind.

Das Semantic Web ist graphenorientiert. RDF Triples bestehen aus URIs. Sobald mehrere Aussagen (Triples) sich auf die gleiche URI bezieht, verknüpfen sich diese Aussagen zu einem gerichtetem, benannten Multigraphen (labeled, directed multi-graph) (Wikipedia-Autoren 2013).

Dazu ist es nicht nötig vorher ein Schema, also eine Struktur festzulegen. Die Struktur entsteht durch die Verknüpfungen und den Beziehungen von selbst und kann organisch mit dem Datenbestand mitwachsen.

2.3.1 Technischer Hintergrund

Die Graphenstruktur des Semantic Webs kann über die verschiedenen RDF Serialisierungsformate mit einen regulären Webserver ausgeliefert werden. Es gibt allerdings auch dedizierte Datenbanken die auf der Triplestruktur von RDF aufbauen: Triplestores.

Diese Datenbanken unterstützen üblicherweise die von W3C standardisierte, graphenorientierte Abfragesprache SPARQL (Andy Seaborne und Steven Harris 2013).

Auch unabhängig vom Semantic Web gewinnen Projekte wie Neo4J (Neo Technology Inc 2014) und graphenorientierte Abfragesprachen wie Gremlin (TinkerPop) aktuell stark an Popularität.

2.4 Trennung von Fakt und Interpretation

In Abschnitt 2.3 wurde bereits erwähnt, dass ein Graph kein Schema bzw. Strukturvorgaben benötigt. Dies macht Graphen zu einem relativ neutralen Datenformat, da Fakten nicht vorher transformiert werden müssen um

in die Struktur zu passen. Sie können so abgelegt werden wie sie sind. Im schlechtesten Fall ergeben sich viele isolierte Fakten die nicht miteinander in Zusammenhängen stehen, aber es ist möglich sie so zu erfassen.

Ein Schema kann allerdings Sinn ergeben um Zusammenhänge aufzuzeigen und herzustellen. Im Semantic Web gibt es ein sehr mächtiges Konzept um dies zu erreichen: Ontologien.

Ontologien basieren auf der Beschreibungslogik (Markus Krötzsch, František Simancík, Ian Horrocks). Sie erlauben es auf einer Meta-Ebene die „Struktur der Welt" zu definieren aus denen Ableitungen möglich sind. Dadurch kann aus dem Faktenbestand auch neues, implizites Wissen geschlossen werden.

Ein interessanter Aspekt, der hier im Fokus stehen soll ist folgender: Da die Fakten ohne Schema gespeichert werden können sind sie (relativ) neutral. Für weitere Interpretation kann eine Ontologie als zusätzliche Ebene darüber gespannt werden. Fakten und Interpretation sind also getrennt. Es ist möglich auf derselben Faktendatenbank unterschiedliche Ontologien zu entwickeln die zu unterschiedlichen Schlussfolgerungen kommen, da sie andere Meinungen und Herangehensweisen vertreten.

Das schlussgefolgerte Wissen kann ebenfalls getrennt gespeichert werden. Damit kommen wir zu einem sauber getrennten dreiteiligem Prozess: (1) Faktenwissen wird durch (2) eine Ontologie interpretiert, woraus (3) neue Schlussfolgerungen entstehen.

Diese Trennung hat einige Vorteile: Es könnte sich herausstellen, dass die Interpretation falsch ist. Wenn nun eine Datenstruktur ein Schema (und damit eine Interpretation) voraussetzt bedeutet dies, dass alle Daten eventuell unter falschen Voraussetzungen erfasst worden sind und damit im schlimmsten Fall unbrauchbar sind. Ist beides getrennt, kann man eine neue Interpretation entwickeln und mit den alten Fakten zu anderen Schlüssen kommen.

Es kommt häufig vor, dass ein Modell über die Zeit wächst. Der Datenspeicherung macht dies aufgrund der Graphenstruktur nichts aus und die Ontologie kann dynamisch mit den Daten mitwachsen und eventuell auch entstehende Inkonsistenzen durch neue Regeln ausbügeln

2.4.1 Technischer Hintergrund

Auch hier gibt es mehrere W3C Standards: Das einfachere RDF Schema (RDFS) (Dan Brickley und Ramanathan Guha 2004) hat die Grundlagen gelegt. Für komplexere Ontologien wurde OWL (Markus Krötzsch et al. 2012) entwickelt, dass auch in verschieden Versionen mit unterschiedlicher Aussagestärke und Komplexität existiert.

2.5 Die Open World Assumption

Die OWA, (Offene Welt Annahme) ist ein alternatives Konzept zur CLOSED WORLD ASSUMPTION. Die OWA hat als Grundannahme, dass eine Wissensbasis immer potentiell unvollständig ist (Hitzler 2007, S. 150). Dies hat viele Konsequenzen: In einer geschlossenen Welt ist alles „falsch", was nicht explizit als wahr bekannt und eingetragen ist. In einer offenen Welt gilt dieser Rückschluss nicht: Reicht die Datenlage nicht aus um sicher auf richtig oder falsch zu schließen ist das Ergebnis unbekannt. (Michael K. Bergman 2009). Alles was nicht ausgeschlossen wurde ist auch möglich.

Die offene Welt Logik ist dadurch weniger restriktiv und mehr vermittelnd: Teilen sich etwa zwei Datensätze eine Aussage die einmalig sein muss (wie eine ID), so werden diese Datensätze zusammengeführt, sofern dies nicht vorher ausgeschlossen wurde.

Dies hat natürlich zur Folge, dass die Logik deutlich komplexer werden kann und man mit der Möglichkeit rechnen muss, kein definitives Ergebnis zu bekommen. Dadurch eignet sich die OWA nicht sehr gut für Systeme die bewusst geschlossen sein müssen und harte Validation.

Doch die OWA ist spielt ihre Stärken aus, wenn es um die Aggregation von Informationen aus sehr unterschiedlichen Quellen geht. In der Realität ist Wissen auch fast immer unvollständig. Wiedersprüche entstehen, bzw. müssen aufgelöst werden. Ein System, dass dies von Anfang an mit in Betracht gezogen hat kann hier wesentlich bessere Ergebnisse erzielen.

3 Schlussbetrachtung

Semantic Web Technologien haben also interessante Lösungen für aktuelle Probleme die bisher noch keine weit

- Semantic Web Technologies have interesting solutions for current problems which are mostly unsolved.
- Standards and Implementations will change but the problems behind not.
- Understanding the concepts behind may prove to be useful, no matter if the Semantic Web Vision gets the adoption it needs.

Literaturverzeichnis

Andy Seaborne; Steven Harris (2013): SPARQL 1.1 Query Language. W3C. Online verfügbar unter http://www.w3.org/TR/2013/REC-sparql11-query-20130321/, zuletzt aktualisiert am 2013, zuletzt geprüft am 14.10.2013.

Antoine Isaac, Ed Summers (2009): SKOS Simple Knowledge Organization System Primer. Online verfügbar unter http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/, zuletzt aktualisiert am 18.08.2009, zuletzt geprüft am 31.10.2014.

Dan Brickley; Ramanathan Guha (2004): RDF Vocabulary Description Language 1.0: RDF Schema. W3C. Online verfügbar unter http://www.w3.org/TR/2004/REC-rdf-schema-20040210/, zuletzt aktualisiert am 2004, zuletzt geprüft am 14.10.2013.

Eric Miller; Frank Manola (2004): RDF Primer. W3C. Online verfügbar unter http://www.w3.org/TR/2004/REC-rdf-primer-20040210/, zuletzt aktualisiert am 2004, zuletzt geprüft am 14.10.2013.

Facebook (October 20th, 2014): Open Graph protocol. Online verfügbar unter http://ogp.me/, zuletzt aktualisiert am October 20th, 2014, zuletzt geprüft am 25.10.2014.

Gavin Carothers; Eric Prud'hommeaux (2013): Turtle. W3C. Online verfügbar unter http://www.w3.org/TR/2013/CR-turtle-20130219/, zuletzt aktualisiert am 2013, zuletzt geprüft am 18.10.2013.

Google Inc., Yahoo Inc., Microsoft Corporation and Yandex: schema.org. Online verfügbar unter http://schema.org/, zuletzt geprüft am 24.10.2014.

Hendrik Arndt (2006): Integrierte Informationsarchitektur. Die erfolgreiche Konzeption professioneller Websites: Springer.

Hitzler, Pascal (2007): The semantic web. Grundlagen. Berlin, Heidelberg, New York, NY: Springer (Lecture notes in computer science, Vol. 4825). Online verfügbar unter http://dx.doi.org/10.1007/978-3-540-33994-6.

Ian Hickson (2014): HTML Microdata. Hg. v. W3C. Online verfügbar unter http://www.w3.org/TR/microdata/, zuletzt aktualisiert am 23.06.2014, zuletzt geprüft am 24.10.2014.

Manu Sporny (2013): HTML+RDFa 1.1. W3C. Online verfügbar unter http://www.w3.org/TR/2013/REC-html-rdfa-20130822/, zuletzt aktualisiert am 2013, zuletzt geprüft am 14.10.2013.

Manu Sporny et al (2014): JSON-LD 1.0. Hg. v. W3C. Online verfügbar unter http://www.w3.org/TR/json-ld/, zuletzt aktualisiert am 16.01.2014, zuletzt geprüft am 24.10.2014.

Markus Krötzsch; Pascal Hitzler; Bijan Parsia; Peter Patel-Schneider; Sebastian Rudolph (2012): OWL 2 Web Ontology Language Primer (Second Edition). W3C. Online verfügbar unter http://www.w3.org/TR/2012/REC-owl2-primer-20121211/, zuletzt aktualisiert am 2012, zuletzt geprüft am 14.10.2013.

Markus Krötzsch, František Simancík, Ian Horrocks: Description Logics.

Michael K. Bergman (2009): The Open World Assumption: Elephant in the Room. Online verfügbar unter http://www.mkbergman.com/852/the-open-world-assumption-elephant-in-the-room/, zuletzt aktualisiert am 21.12.2009, zuletzt geprüft am 01.11.2014.

Nelson, T. H.: Complex information processing. In: Lewis Winner (Hg.): the 1965 20th national conference. Cleveland, Ohio, United States, S. 84–100.

Neo Technology Inc (2014): Neo4j Graph Database. Online verfügbar unter http://neo4j.com/, zuletzt aktualisiert am 31.10.2014, zuletzt geprüft am 31.10.2014.

Sankar, Shyam: The rise of human-computer cooperation. Online verfügbar unter http://www.ted.com/talks/shyam_sankar_the_rise_of_human_computer_cooperation?language=en, zuletzt geprüft am 31.10.2014.

Tim Berners-Lee (2009): The next web. Hg. v. L. L.C. TED. Online verfügbar unter http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html, zuletzt aktualisiert am März 2009, zuletzt geprüft am 08.10.2013.

Tim Berners-Lee et al. (2014): Uniform Resource Identifier (URI): Generic Syntax. Online verfügbar unter http://tools.ietf.org/html/rfc3986, zuletzt aktualisiert am 11.10.2014, zuletzt geprüft am 31.10.2014.

TinkerPop: Gremlin. Online verfügbar unter https://github.com/tinkerpop/gremlin/wiki, zuletzt geprüft am 31.10.2014.

Wikipedia-Autoren, siehe Versionsgeschichte (2013): Resource Description Framework - Wikipedia, the free encyclopedia. Hg. v. Wikipedia. Online verfügbar unter http://en.wikipedia.org/w/index.php?oldid=580735405, zuletzt aktualisiert am 08.11.2013, zuletzt geprüft am 12.11.2013.