25. Oktober 2014

SEMINARARBEIT

Fünf

Semantic Web

Paradigmen

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 März 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 und (5) 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

SWT Semantic Web-Technologien

OWA Open World Assumption

CWA Closed World Assumption

SEO Suchmaschinenoptimierung

# Einleitung

Ein bekanntes Zitat, dass nachträglich diversen Autoren zugeschrieben wird, besagt: “*Wer als Werkzeug nur einen Hammer hat, sieht in jedem Problem einen Nagel*.“

Das ist offensichtlich eine Aufforderung sich mit anderen, neuen Werkzeugen zu beschäftigen. Das Wort „Werkzeug“ ist hier natürlich metaphorisch gemeint: Oberflächlich verstanden, könnte 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.

Die konkreten technischen Werkzeuge ändern sich. Doch die grundsätzlichen Ideen hinter ihnen sind beständiger und für das Verständnis wichtiger. Deshalb wird sich dieser Artikel mehr auf das „Warum?“ konzentrieren und weniger auf das „Wie?“.

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

# Vorstellung der Paradigmen

## 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 es kann auch umgekehrt argumentiert werden: 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 die Leistung beider Seiten im Alleingang weit übertreffen. (Sankar 2012)

Das Semantic Web kann als Initiative für eine bessere Mensch-Computer-Kooperation 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 allein lösen dieses Problem nicht, da sie nur die Dokumente und Dateien an sich beschreiben, nicht aber deren 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. „Annotation“ deutet an, dass der bereits bestehende Inhalt durch maschinenlesbare Informationen „kommentiert“ wird.

Dies ermöglicht es Computern die Inhalte eindeutig zu verstehen und nicht nur durch KI Technologien zu „erraten“, was oft zu Fehlern führt und aufwendig ist. Dadurch können Suchmaschinen relevantere und intelligentere Ergebnisse bringen. Wissen kann aus verschiedenen Quellen neu kombiniert und evaluiert werden. Es könnte auch eine neue Generation von Webdiensten ermöglichen, die noch kooperativer zusammenarbeiten als es bisher möglich ist.

Dies ist natürlich ein sehr ambitioniertes Vorhaben und es bleibt noch abzuwarten in welchem Umfang und Qualität es sich erfüllen wird.

### Aktueller Stand

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

Schema.org (Google Inc. et al.) ist Projekt von Google, Yahoo, Microsoft, etc., das ein gemeinsames Vokabular für die Beschreibung von Dingen und Vorgängen im Internet definiert. Facebook hat einen eigenen Standard entwickelt, den OpenGraph (Facebook 06.11.2014).

Auch Forschung und Industrie haben die Semantic Web Technologien aufgenommen um übergreifende Standards zu schaffen, wie etwa das SKOS Modell (Antoine Isaac, Ed Summers 18.08.2009) zur Verwaltung von Wissensbeständen.

### 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 23.06.2014), die beide auf XML basieren. Steigende Verbreitung hat aktuell JSON-LD (Manu Sporny et al 16 January 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 tiefer eingegangen.

## Hyperlinks

Hyperlinks sind nicht erst mit dem Web erfunden worden. Sie haben eine lange Geschichte, die mindestens bis zur Memex zurückreicht (O'Regan 2012, S. 102) und waren wichtiger Bestandteil der in den 60er Jahren entstehenden Hypertext Systeme (Nelson 1965).

Wenn heute von Links gesprochen wird, ist meist nur das Ziel der Verknüpfung, die URI (Unique Resource Identifier) gemeint. URIs identifizieren eine Ressource durch eine eindeutige Adresse. Doch ein Link besteht aus mehr Komponenten: Eine Quelle und mindestens ein Ziel. Falls mehrere Ziele möglich sind, spricht man davon, dass ein Link eine beliebige Arität haben kann (Frank G. Halasz 1990, S. 9). Meist kommt eine Relation, die die Beziehung zwischen Quelle und Ziel 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), nur ein Ziel wird unterstützt und die Relation zwischen Quelle und Ziel spielt keine große Rolle.

Dadurch sind Links im Web einfacher, aber auch nicht so flexibel wie einige konkurrierende Hypertext Systeme. Die unidirektionale Natur der Verknüpfungen ermöglicht die dezentrale Struktur des Internets, da keine zentral verwaltete Datenbank aller Verknüpfungen nötig ist. Links können also ohne vorherigen Konsens gesetzt werden. Rückblickend wird diese Entscheidung mit für den Erfolg des Webs verantwortlich gemacht. (Hendrik Arndt 2006, S. 153f.)

Weblinks 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, das 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.

A Uniform Resource Identifier (URI) is a compact sequence of characters that identifies an abstract or physical resource. (Tim Berners-Lee et al. 2005)

Der Begriff „Resource“ wird hier also gedehnt. Dadurch ist es möglich allgemein über Dinge, Konzepte und Beziehungen Aussagen zu treffen 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 zuvor in Abschnitt 2.1 erwähnte semantische Annotation.

### Technischer Hintergrund

URIs (Tim Berners-Lee et al. 2005) wurden vom der IETF standardisiert. Eine interessante Zusammenfassung (und Symbiose) verschiedener älterer Hypertext Systeme ist das Dexter Modell (Frank G. Halasz 1990).

## Graphenstruktur

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 kommt zu Informationsverlust und die Reduktion auf einen kleineren Nenner zwingt Design und Strukturentscheidungen auf, die meist subjektiv sind. Dabei 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 beziehen, verknüpfen sich diese Aussagen zu einem gerichtetem, benannten Multigraphen (labeled, directed multi-graph) (Wikipedia-Autoren 08.11.2013).

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.

### Technischer Hintergrund

Die Graphenstruktur des Semantic Webs kann über die verschiedenen RDF Serialisierungsformate mit einem 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 31.10.2014) und graphenorientierte Abfragesprachen wie Gremlin (TinkerPop 06.11.2014) aktuell stark an Popularität.

## 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 et al. 2014). Sie erlauben es auf einer Meta-Ebene die „Struktur der Welt“ zu definieren, anhand der neue Ableitungen möglich werden. 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

### 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.

## Die Open World Assumption

Die OWA, (Offene Welt Annahme) ist ein alternatives Konzept zur Closed World Assumption (CWA). Die OWA hat als Grundannahme, dass eine Wissensbasis immer potentiell unvollständig ist (Hitzler 2007, S. 150).

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 21.12.2009). Alles was nicht ausgeschlossen wurde ist auch möglich.

Ein Beispiel: Folgender Fakt ist gespeichert: „Peter“ „hat ein Kind“ „Harry“. „Harry“ „hat das Geschlecht“ „männlich“. In einer geschlossenen Welt wird daraus gefolgert, dass er genau ein Kind hat und alle Kinder männlich sind. In einer offenen Welt kann über beides keine Aussage gemacht werden, da Peter noch mehr Kinder haben könnte, von denen wir nichts wissen. (Baader 2003, S. 72f.)

Die offene Welt Logik ist 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 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 benötigen.

Doch die OWA spielt ihre Stärken aus, wenn es um die Aggregation von Informationen aus sehr unterschiedlichen Quellen geht. In der Realität ist Wissen fast immer unvollständig. Die Rückmeldung, dass ein Sachverhalt unbekannt ist, kann eine wichtige Information sein. Daten können bereits Wiedersprüche enthalten und durch die Zusammenführung verschiedener Quellen dazu noch neue entstehen.

Ein System, dass Unvollständigkeit und den Umgang mit Widersprüchlichkeiten von Anfang an mit in Betracht zieht hat kann hier bessere, bzw. angemessenere Ergebnisse erzielen.

# Schlussbetrachtung

Die einzelnen Paradigmen konnten im Umfang dieser Arbeit natürlich nur oberflächlich angeschnitten werden, denn jedes ist ein Forschungsbereich für sich. Der Autor hofft, dass der Leser einen ersten Überblick und damit einige Gedankenanstöße mitbekommen hat und dadurch sein Repertoire erweitern konnte.

Ob sich das Semantic Web durchsetzt und breitere Adoption findet, steht zum Zeitpunkt dieser Arbeit noch offen. Es ist davon auszugehen, dass die technischen Standards und deren Implementationen sich weiter ändern werden. Einige werden vielleicht durch bessere oder einfachere Ansätze ersetzt.

Vielleicht werden einige der hier vorgestellten Konzepte auch in anderen Kontexten Verbreitung finden. Neue Ideen entstehen bekanntlich oft durch die Kombination von bereits vorhandenen Konzepten, nicht selten aus verschiedenen Disziplinen

Literatur

Andy Seaborne, Steven Harris (2013) SPARQL 1.1 Query Language. W3C. http://www.w3.org/TR/2013/REC-sparql11-query-20130321/. Zugegriffen: 14. Oktober 2013.

Antoine Isaac, Ed Summers (18.08.2009) SKOS Simple Knowledge Organization System Primer. http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/. Zugegriffen: 31. Oktober 2014.

Baader F (2003) The description logic handbook; Theory, implementation, and applications. Cambridge University Press, Cambridge, UK, New York.

Dan Brickley, Ramanathan Guha (2004) RDF Vocabulary Description Language 1.0: RDF Schema. W3C. http://www.w3.org/TR/2004/REC-rdf-schema-20040210/. Zugegriffen: 14. Oktober 2013.

Eric Miller, Frank Manola (2004) RDF Primer. W3C. http://www.w3.org/TR/2004/REC-rdf-primer-20040210/. Zugegriffen: 14. Oktober 2013.

Facebook (06.11.2014) Open Graph protocol. http://ogp.me/. Zugegriffen: 25. Oktober 2014.

Frank G. Halasz (1990) The Dexter Hypertext Reference Model.

Gavin Carothers, Eric Prud’hommeaux (2013) Turtle. W3C. http://www.w3.org/TR/2013/CR-turtle-20130219/. Zugegriffen: 18. Oktober 2013.

Google Inc., Yahoo Inc., Microsoft Corporation, Yandex schema.org. http://schema.org/. Zugegriffen: 24. Oktober 2014.

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

Hitzler P (2007) The semantic web; Grundlagen. Springer, Berlin, Heidelberg, New York, NY.

Ian Hickson (23.06.2014) HTML Microdata. http://www.w3.org/TR/microdata/. Zugegriffen: 24. Oktober 2014.

Manu Sporny (2013) HTML+RDFa 1.1. W3C. http://www.w3.org/TR/2013/REC-html-rdfa-20130822/. Zugegriffen: 14. Oktober 2013.

Manu Sporny et al (16 January 2014) JSON-LD 1.0. http://www.w3.org/TR/json-ld/. Zugegriffen: 24. Oktober 2014.

Markus Krötzsch, František Simancík, Ian Horrocks (2014) Description Logics.

Markus Krötzsch, Pascal Hitzler, Bijan Parsia, Peter Patel-Schneider, Sebastian Rudolph (2012) OWL 2 Web Ontology Language Primer (Second Edition). W3C. http://www.w3.org/TR/2012/REC-owl2-primer-20121211/. Zugegriffen: 14. Oktober 2013.

Michael K. Bergman (21.12.2009) The Open World Assumption: Elephant in the Room. http://www.mkbergman.com/852/the-open-world-assumption-elephant-in-the-room/. Zugegriffen: 01. November 2014.

Nelson TH (1965) Complex information processing:84–100. doi:10.1145/800197.806036.

Neo Technology Inc (31.10.2014) Neo4j Graph Database. http://neo4j.com/. Zugegriffen: 31. Oktober 2014.

O'Regan G (2012) A brief history of computing. Springer, London, New York.

Sankar S (2012) The rise of human-computer cooperation. http://www.ted.com/talks/shyam\_sankar\_the\_rise\_of\_human\_computer\_cooperation?language=en. Zugegriffen: 31. Oktober 2014.

Tim Berners-Lee (März 2009) The next web. http://www.ted.com/talks/tim\_berners\_lee\_on\_the\_next\_web.html. Zugegriffen: 08. Oktober 2013.

Tim Berners-Lee et al. (2005) Uniform Resource Identifier (URI): Generic Syntax. http://tools.ietf.org/html/rfc3986. Zugegriffen: 31. Oktober 2014.

TinkerPop (06.11.2014) Gremlin. https://github.com/tinkerpop/gremlin/wiki. Zugegriffen: 31. Oktober 2014.

Wikipedia-Autoren sV (08.11.2013) Resource Description Framework - Wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?oldid=580735405. Zugegriffen: 12. November 2013.