

1 Die Idee des Semantic Web

In diesem Kapitel werden die grundlegenden Ideen dessen, was das Semantic Web ausmacht, skizziert und die dafür zentralen Begriffe erläutert.

1.1 Das Web

Der Übergang von der Industrie- zur Informationsgesellschaft findet seine wohl markanteste Ausprägung in der rasanten Entwicklung des *World Wide Web* (auch kurz das Web genannt). Es ist aus dem heutigen Alltagsleben kaum mehr wegzudenken und beinhaltet eine das menschliche Vorstellungsvermögen längst übersteigende und ständig anwachsende Menge von Informationen. Dank seiner Organisation und der zugrundeliegenden Infrastruktur bietet das Web gegenüber herkömmlichen Methoden des Informationsaustausches eine Reihe von erheblichen Vorteilen.

Die Chance, die präsenten Informationen jederzeit auf den neuesten Stand zu bringen und die nahezu vernachlässigbare Übertragungszeit zum Empfänger bietet die Möglichkeit einer ständigen *Aktualität* der bereitgestellten Daten. Darüber hinaus besteht eine nahezu universelle *Verfügbarkeit* von Information. Prinzipiell ist es möglich, von jedem Ort aus auf die an einem beliebigen anderen Ort bereitgestellte Information zuzugreifen. Zudem wird durch diesen weniger aufwändigen und kostengünstigeren Weg der Informationsbeschaffung auch breiteren Bevölkerungsschichten mehr Information zugänglich.¹ Bereits heute ist es in manchen entlegenen Andenstädten einfacher, Mails zu lesen und die neuesten Nachrichten abzurufen, als ein Telefongespräch zu führen.

Darüber hinaus hat das Web zu einer Liberalisierung der *Bereitstellung* von Information geführt. War bisher aufgrund des damit verbundenen enormen Aufwandes und der erforderlichen Kosten eine Versorgung eines breiten Publikums mit Informationen vergleichsweise wenigen „Informationsoligopolisten“ (wie beispielsweise Verlagshäusern, Fernsehsendern, großen Unternehmen) vorbehalten, bietet das Web nunmehr auch kleineren Interessengruppen bis hin zu Einzelpersonen die Möglichkeit, zu verschwindenden Kosten Informationen einer großen Masse von Menschen zugänglich zu machen.² Eine weitere essentielle Dimension der Entwicklung des Web stellt dessen fortschreitende Nutzung für kommerzielle Zwecke im weitesten Sinne dar. Wer hat

¹Eine Beherrschung grundlegender Kulturtechniken, wie z.B. Lesen, sei vorausgesetzt.

²Dabei sind sich die Autoren durchaus der potentiellen Probleme bewusst, die diese Entwicklung (beispielsweise hinsichtlich Qualitätssicherung und Zuverlässigkeit der verbreiteten Informationen) mit sich bringt.

etwa noch kein Buch oder keine CD über einen Internetanbieter wie Amazon bestellt? Und auch hier hat die grundsätzlich pluralistische Organisation des Web dazu geführt, die Schranken für den Marktzugang auch für „kleinere“ Akteure drastisch zu senken. Virtuelle Marktplätze oder Auktionsplattformen wie Ebay legen davon ein eindrucksvolles Zeugnis ab.

1.2 1.2 Probleme des Web

Alles in allem zeichnet sich das Bild eines Web mit einer unüberschaubaren Menge präsenter Informationen, deren Repräsentation – und hierin besteht das Problem – auf den Menschen als Endnutzer ausgerichtet ist. So kann ein menschlicher Nutzer die Bedeutung einer Information auf einer Webseite problemlos erfassen, in andere Darstellungsformen transformieren und zu anderen Informationen in Beziehung setzen, während eine Maschine dies in aller Regel nicht zu leisten im Stande ist.

Ein sich daraus ergebendes Problem besteht darin, dass eine bestimmte Information, die zwar prinzipiell verfügbar ist, in der Fülle der vorhandenen Informationen nur schwer gefunden werden kann. Suchmaschinen wie Google, Yahoo etc. leisten hier durch den Einsatz ausgefeilter statistischer Methoden bereits Erstaunliches. Gleichwohl fußen auch diese Techniken letztlich auf der Lokalisierung von Zeichenketten in Text. Eine Recherche nach dem Begriff „Kohl“ wird demnach gleichermaßen Fundstellen liefern, die sich mit Gemüse oder dem deutschen Altbundeskanzler befassen. Wünschenswert wäre hier natürlich eine weniger stichwortbasierte, mehr inhaltliche, eben eine *semantische Suche*.

Weiterhin bringt die dezentrale Struktur und Organisation des Web zwangsläufig eine *Heterogenität* der vorhandenen Informationen auf ganz verschiedenen Ebenen mit sich: angefangen bei unterschiedlichen Dateiformaten bzw. Kodierungstechniken (ASCII vs. Unicode...), über die verschiedenen verwendeten natürlichen Sprachen bis hin zum ganz unterschiedlichen Aufbau privater Homepages. Dies macht es schwierig bis unmöglich, über das Web verteilte Informationen zu einem Gebiet zu sammeln und einheitlich aufbereitet darzustellen bzw. weiterzuverwenden. So wäre beispielsweise eine automatische Extraktion und thematische Gegenüberstellung der Wahlprogramme von den Internetpräsenzen der Parteien (sinnvollerweise ergänzt um Informationen über die tatsächlich erfolgte Umsetzung früherer Wahlversprechen) eine willkommene Entscheidungsgrundlage für den unentschlossenen verantwortungsbewussten Wähler. Hier haben wir es ganz offensichtlich mit dem Problem der *Informationsintegration* zu tun.

Schließlich ist es denkbar, dass Information, nach der ein Nutzer sucht, nicht explizit im Web zu finden ist, sondern lediglich aus einer Reihe von (mögli-

cherweise sogar über das Web verteilten) gegebenen Fakten folgt. So ergibt beispielsweise die Information, dass Samaipata in Bolivien und Dresden in Deutschland liegt, gemeinsam mit der Information, dass die Differenz der lokalen Zeiten von Bolivien und Deutschland fünf Stunden beträgt, dass man aus Dresden möglichst nicht vor dem Mittagessen in Samaipata anrufen sollte. Hier handelt es sich um die Problematik des *impliziten Wissens*.

1.3 Das Semantic Web

Grundsätzlich kann man zwei komplementäre Herangehensweisen zur Lösung dieser Probleme unterscheiden.

Zum einen wäre es denkbar, Methoden der künstlichen Intelligenz anzuwenden, welche die kognitiven Aufgaben, die ein Mensch bei der Verarbeitung von Informationen aus dem Web erfüllen würde, bis zu einem gewissen Grade übernehmen und die primär für Menschen repräsentierten Informationen in formalere, maschinenverarbeitbare Daten verwandeln. Zwar stellt die künstliche Intelligenz nach wie vor ein spannendes Forschungsfeld dar, die bisher erreichten Resultate lassen jedoch eine webweite zuverlässige Anwendung als sehr fragwürdig erscheinen – zumindest in näherer Zukunft.

Eine alternative Herangehensweise besteht im Ansatz des *Semantic Web*. Das Semantic Web steht für die Idee, die Information von vornherein in einer Art und Weise zur Verfügung zu stellen, die deren Verarbeitung durch Maschinen ermöglicht. Was sind nun die Grundvoraussetzungen für ein solches Unterfangen?

Zum ersten ist es nötig, einheitliche, offene *Standards* für die Beschreibung von Informationen zu vereinbaren, die es letztlich ermöglichen sollen, Informationen zwischen verschiedenen Anwendungen und Plattformen auszutauschen und zueinander in Beziehung zu setzen, eine Fähigkeit, die *Interoperabilität* genannt wird. Eine grundlegende Anforderung an diese Standards – neben deren formal klarer Definition – besteht in der Flexibilität und Erweiterbarkeit: für eine auch zukünftige Etablierung muss ein Standard auch unvorhergesehene Anwendungsfälle möglichst gut behandeln, ohne revidiert werden zu müssen.

Genau diesem Ziel der Schaffung solcher Standards für das Semantic Web hat sich das sogenannte *World Wide Web Consortium*³ (W3C) verschrieben und bereits grundlegende Standards für XML, RDF(S) und OWL – Informations-Spezifikationssprachen, die wir sämtlich in diesem Buch behandeln werden – definiert. XML ist dabei nicht im engeren Sinne dem Semantic Web zuzuordnen, sondern eher dem (klassischen) Web. RDF(S) und OWL hingegen sind

³<http://www.w3.org>

sogenannte Ontologiesprachen, die speziell für die Verwendung im Semantic Web entwickelt wurden. Der Begriff *Ontologie* ist in unserem Zusammenhang als äquivalent zum Begriff *Wissensbasis* zu verstehen und beschreibt schlicht ein in RDF(S) oder OWL erstelltes Dokument, welches Wissen einer Anwendungsdomäne modelliert.

Die zweite wichtige Zutat für das Semantic Web sind zweifellos Methoden zur Schlussfolgerung von „neuen“ Informationen aus gegebenen. Um größtmöglichen Nutzen aus spezifiziertem Wissen zu ziehen, muss auch implizite Information aus einer Spezifikation extrahiert werden können. Hier befinden wir uns natürlich im Fachgebiet der *formalen Logik* mit ihren vielen Ausprägungen. Wir werden in diesem Buch eine Reihe von logischen Formalismen und Schlussfolgerungsmethoden behandeln.

Ein weiterer Punkt im Hinblick auf die Idee des Semantic Web sei hier noch besprochen. Nicht selten findet man Äußerungen wie, die Grundidee des Semantic Web bestünde darin, Maschinen bzw. Computer in die Lage zu versetzen, die Bedeutung von Informationen zu „verstehen“. Davon abgesehen, dass sich damit ein nicht zu unterschätzendes philosophisches Problem auftäte, würde dies Aspekte der künstlichen Intelligenz hinter der Semantic-Web-Idee aus unserer Sicht zu sehr betonen. Die Ziele des Semantic Web sind eigentlich wesentlich moderater: Finde Wege und Methoden, Informationen so zu repräsentieren, dass Maschinen damit in einer Art und Weise umgehen können, die aus menschlicher Sicht nützlich und sinnvoll erscheint. Dabei mag das entstehende Verhalten von Systemen in den Augen mancher Nutzer durchaus als intelligent erscheinen.

1.4 1.4 Semantische Technologien

Das Semantic Web basiert auf Grundlagentechnologien, die wir *Semantische Technologien* nennen. Zu diesen gehören Wissensrepräsentationssprachen für Ontologien sowie Methoden und Werkzeuge zur Erstellung, Wartung und Anwendung von Ontologien. Anwendungen dieser Technologien sind mittlerweile nicht mehr nur auf das World Wide Web beschränkt, sondern erstrecken sich über viele Bereiche der Informatik wie Wissensmanagement, Medieninformatik, Datenintegration, Kognitive Systeme, Ambiente Intelligenz, Softwaretechnik, Maschinelles Lernen, eScience, Internetökonomie, Bioinformatik etc. Zurzeit erscheint es tatsächlich wahrscheinlich, dass kommerzielle Anwendungen Semantischer Technologien sich zuerst primär auf Bereiche beziehen werden, die nicht dem offenen World Wide Web zuzuordnen sind, sondern vielmehr industrielle Lösungen zum Wissensmanagement darstellen werden.

In diesem Buch behandeln wir das Fundament Semantischer Technologien und des Semantic Web, nämlich die Sprachen zur Repräsentation von Wissen in der Form von Ontologien.

Syntax vs. Semantik

Zwei Begriffe, die bei der Definition von Sprachen und in der Logik eine zentrale Rolle spielen und uns im Folgenden häufig begegnen werden, wollen wir hier kurz erklären.

Unter *Syntax* versteht man im Allgemeinen eine Menge von Regeln zur Strukturierung von Zeichen und Zeichenketten. In der Linguistik steht Syntax für eine Menge von Regeln, um aus Worten Sätze oder Phrasen zu bilden. In der Informatik, und auch im Sinne dieses Buches, steht Syntax für eine Menge von Regeln, um Programme oder Dokumente mit bestimmten Eigenschaften, z.B. gültige XML-Dokumente, zu erzeugen.

Semantik steht allgemein für „Bedeutung von Wörtern, Phrasen oder Symbolen“. In der Linguistik steht Semantik für die Wissenschaft der Bedeutung in verschiedenen Sprachformen. In der Informatik, und insbesondere im Bereich Semantic Web, versteht man unter Semantik die Bedeutung von Worten bzw. Zeichen(-ketten) und ihre Beziehungen untereinander. Auf den Semantikbegriff in der formalen Logik gehen wir im Kapitel 4 näher ein.

Allgemein steht im Sinne des Buches Syntax für die normative Struktur von Daten, welche erst durch Semantik eine Bedeutung erhält.

1.5 Zusammenfassung

1.5

Das Semantic Web bietet Lösungsansätze für die Probleme, welche sich aus der Fülle der im Web präsenten Informationen ergeben, beispielsweise das Auffinden relevanter Informationen, die Integration von Informationen aus verschiedenen Quellen und das Problem des impliziten Wissens. Es werden Standards zur Repräsentation von Information eingeführt, die deren Austausch ermöglichen sollen. Methoden zum automatischen Schlussfolgern sollen implizites Wissen erschließen.

1.6 Weiterführende Literatur

1.6

Der vielzitierte Originalartikel, in dem der Direktor des W3C, Tim Berners-Lee, die Vision des Semantic Web vorstellt, ist [BLHL01]. Ein neuerer Artikel von ihm dazu ist [SBLH06].

Die zentrale Webseite des W3C zu Semantic Web findet sich unter <http://www.w3.org/2001/sw/>.

Für aktuelle Anwendungen und zu Fragen der Realisierung von Semantic-Web-Technologien auf den in diesem Buch behandelten Grundlagen empfehlen wir [SS04].

Eine Anwendungsstudie findet sich in [SHES05] und ein Visionsartikel zu semantischen Technologien und künstlicher Intelligenz ist [SAHS06].