

mehr zum thema:  
www.w3.org/2001/sw/  
www-protege.stanford.edu

## SEMANTIC WEB

Die Filterung relevanter Information mit Hilfe einer der zurzeit verfügbaren Suchmaschinen und einem Browser gleicht der Suche nach der Nadel in einem täglich anwachsenden Heuhaufen. Auch unter Einsatz größerer Rechenleistung kann dieses Problem nicht gelöst werden, da weiterhin überwiegend Heu gefunden wird und keine Nadeln. Das Semantic Web soll dieser Herausforderung begegnen. Durch die Anreicherung der im Web verfügbaren Daten mit Meta-Information (Semantik), ist es einer Anwendung (Web Agent) möglich die relevante Information zu berechnen. Die mühsame Arbeit des Berechnens wird (wieder) an die Maschine delegiert. Das Denken und Verstehen verbleibt beim Menschen. Der Artikel beschreibt zentrale Konzepte des Semantic Webs und zeigt potenzielle Risiken, die mit der Automatisierung der Informationsbeschaffung verbunden sind, auf.

Durch das World Wide Web Consortium (W3C) wurde in der Vergangenheit eine Reihe von Empfehlungen (Spezifikationen) auf unterschiedlichem Abstraktionsniveau verabschiedet. So sind Spezifikationen wie XML (*Extensible Markup Language*) und SOAP sehr technisch orientiert. Im Unterschied dazu definieren Spezifikationen wie die *Web Services Architecture* oder das *Semantic Web* Konzepte, die konkretisierende, technische Spezifikationen nutzen und ihr Zusammenspiel arrangieren. Per se ist im *Semantic Web* keine Technik oder Spezifikation ausgeschlossen. Vielmehr werden pragmatisch aus der Vielzahl der Möglichkeiten diejenigen genutzt, die das Konzept bzw. die Philosophie bestmöglich unterstützen. Dies schließt auch Spezifikationen ein, die nicht unter dem Patronat des W3C stehen. Ziel des *Semantic Webs* ist es, Aussagen über die Bedeutung der im Web nutzbaren Ressourcen (Personen, Objekte, Konzepte usw.) darzustellen, unabhängig davon, ob diese Ressourcen physisch im Web repräsentiert sind (z. B. mittels einer URL). Als Wunschvorstellung verbirgt sich dahinter nichts weniger als die Weiterentwicklung des gegenwärtig präsentationsorientierten Webs zu einem universellen Kommunikationsmedium, in dem maschinelle Agenten im Auftrag von Menschen gewisse Tätigkeiten vollführen. Die Agenten sollen hierbei zu einem gewissen Grade regelbasiert selbstständig agieren, ohne jedoch eigenständig Entscheidungen auf intelligenter Basis treffen zu können.

In **Abbildung 1** sind verschiedene Spezifikationen aus dem Umfeld der

„Knowledge Technologies“ und ihre Abhängigkeiten untereinander dargestellt (siehe auch [Top]). Die Abgrenzung der verschiedenen Konzepte untereinander ist naturgemäß etwas unscharf, da die Spezifikationen nicht Mosaiksteinchen eines von langer Hand geplanten Bildes sind, sondern in großen Teilen unabhängig voneinander vorangetrieben werden.

Der Aufbau des *Semantic Webs* – d.h. dessen Komponenten – wird aktuell vornehmlich aus zwei Perspektiven diskutiert. Zum einen handelt es sich um eine eher technik-zentrierte Betrachtungsweise im Kontext der *Service-Orientierten Architektur (SOA)*, wie sie zurzeit im Zusammenhang mit den Web-Services im Fokus steht. Bedeutung in maschinenwertbarer Form darzustellen, ist eine notwendige Voraussetzung, um das Ziel der Automatisierung im Rahmen einer SOA realisieren zu können (vgl. [Dos04-a], [Dos04-b], [Dos04-c], [Dos04-d]).

Im Rahmen dieses Artikels werden wir jedoch der zweiten Perspektive, der mehr anthropozentrischen Betrachtungsweise, den Vorzug geben, zumal dies auch dem ursprünglichen Kontext der Initiatoren der *Semantic-Web-Idee* – Tim Berners-Lee, James Hendler und Ora Lassila – entspricht. Eine der zentralen Aussagen ihres visionären Artikels von 2001 lautet: „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“ (vgl. [Ber01]).

Daraus lassen sich drei zentrale Ziele des *Semantic Webs* ableiten

- *The Semantic Web is an extension of the current web ...* Das existierende

## die autoren



Dr. Wolfgang Dostal  
(E-Mail: wolfgang.dostal@de.ibm.com)  
ist für die IBM Global Services (AMS SI) als IT-Architekt tätig und Mitglied der IBM Community of Excellence for SOA and Web Services.



† Mario Jeckle  
war Professor für Software-Engineering an der FH Furtwangen. Kurz nach Fertigstellung dieses Artikels kam er als Ersthelfer bei einem Verkehrsunfall auf tragische Weise ums Leben.



Dr. Ingo Melzer  
(E-Mail: paper@ingo-melzer.de) leitet bei der DaimlerChrysler Forschung das Team Integrationsarchitekturen. Er beteiligt sich an der Standardisierung von Web-Services im W3C und hat einen Lehrauftrag an der Universität Ulm.



Barbara Zengler  
(E-Mail: bz@barbara-zengler.de) ist selbstständige Trainerin und Beraterin in den Bereichen Systemintegration, objektorientierte Modellierung und Web-Techniken.



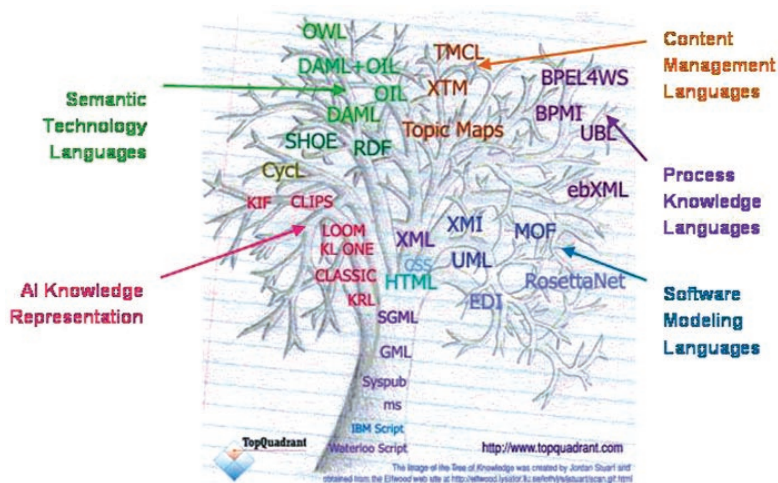


Abb. 1: The Tree of Knowledge Technologies (Quelle: www.topquadrant.com)

präsentationsorientierte Web (Browser-Web) wird nicht abgelöst oder neu definiert, sondern erweitert. Alle Anwender und Anwendungen, welche die Möglichkeiten des *Semantic Webs* nicht nutzen möchten oder können, werden auch zukünftig ihren vertrauten Browser weiter einsetzen können.

- ... in which information is given well-defined meaning, ... Die zusätzliche Bedeutungsdimension des *Semantic Webs* wird vom Menschen aufge-spannt und nicht durch die „Intelligenz“ von Maschinen. Die im Web verfügbaren Daten werden dabei explizit um Aussagen bezüglich ihrer Bedeutung (Semantik) angereichert.
- ... better enabling computers and people to work in cooperation. Das *Semantic Web* hat das Ziel, den Menschen bei der Nutzung des Webs zu unterstützen.

Im Folgenden werden wir zunächst motivieren, weshalb das Konzept des semantischen Webs benötigt wird. Anschließend sollen die Kernkomponenten, aus denen es „komponiert“ wurde, und ihr Zusammenspiel vorgestellt werden. Abschließend erfolgt ein kleiner Ausblick auf die kommenden Herausforderungen und auch Risiken des *Semantic Webs*.

## Das browser-orientierte Web

Der Kontext, in dem das Konzept – oder besser die Vision – des *Semantic Webs* entstanden ist, ist das WWW, erdacht aus der Notwendigkeit heraus, Wissenschaftlern den schnellen und effizienten Datenaustausch zu ermöglichen.

Lange Zeit wurde die heute anerkannt revolutionäre Idee des *Hyper Text Trans-*

*fer Protocols (HTTP)* und der *Hypertext Markup Language (HTML)* als das akademische Spielzeug einiger weniger betrachtet bzw. in seiner Leistungsfähigkeit aufgrund der zweifellos bestehenden Probleme (Stichwort „HTTP 404 – file not found“) als noch unreife Technik eingestuft. Mittlerweile prägt das WWW große Teile unserer Gesellschaft und der kommerziellen Geschäftsbeziehungen. Die Prägung ist sogar so weit vorgedrungen, dass die Begriffe *WWW* und *Internet* synonym verwendet werden.

Ablesen lässt sich der Siegeszug des WWW vor allem an der enormen Zuwachsrates verfügbarer Server. So stieg nach [Net] die Zahl der im Web verfügbaren Server in den letzten zwölf Monaten um 10 Millionen auf inzwischen etwa 50 Millionen Server an. Vordergründig bedeutet dies, dass soviel Daten wie noch nie in der Menschheitsgeschichte per Klick in Sekundenschnelle zur Verfügung stehen. Es gibt kaum eine Art von Inhalt, die im WWW nicht verfügbar wäre.

Eine der Schattenseiten dieser hohen Verfügbarkeit von immensen Informationsmengen ist die Schwierigkeit, die gewünschten oder relevanten Inhalte aus der Flut der angebotenen Daten herauszufiltern. Eine Ursache dieses Dilemmas ist, dass mit dem HTML-Formalismus semantische Anteile – also die Bedeutung des codierten Textes – nicht explizit notiert werden können. Auch wenn, wie in Listing 1, die Überschrift mit dem Element „h1“ markiert und damit von den restlichen Daten unterschieden wird, ist mit dieser Markierung keine Information über die inhaltliche Bedeutung des Textes der-gestalt verfügbar, dass eine Anwendung diese sicher auswerten könnte. Die Markierung besitzt im Wesentlichen syn-

taktischen Charakter. Ein Browser wertet das Element „h1“ als ein besonderes Formatierungsmerkmal aus. Streng genommen ist dies zwar auch schon etwas Semantik, allerdings nicht im Sinne des *Semantic Webs*, dass die Bedeutung des Textes (Element-Inhalt) verfügbar wäre. Diese verbleibt vielmehr nach dem Niederschreiben in der Vorstellungswelt des Autors der HTML-Seite. Die Interpretation der Daten im Browser (z.B. „Paris ist eine Stadt“) muss vom Leser der Seite vorgenommen werden.

Eine weitere Schwierigkeit besteht darin, dass Zeichen (Wörter, Begriffe) im Kontext ihres Auftretens interpretiert werden müssen. Veranschaulichen lässt sich dies anhand der Schlüsselwortsuche mit Hilfe einer Suchmaschine. Die ersten Volltext-Suchmaschinen im Web (etwa *Altavista*) führten eine reine Schlüsselwortsuche durch, d.h. die vom Benutzer eingegebenen Zeichen wurden Zeichen für Zeichen mit denen auf diversen HTML-Seiten verglichen. Daran werden gleich mehrere Probleme einer nicht-semantischen Suche deutlich:

- **Homonymie:** Dasselbe Wort kann in verschiedenen Kontexten mit unterschiedlichen Bedeutungen vorkommen.
- **Synonymie:** Dasselbe Konzept kann in verschiedenen Kontexten anders benannt sein).
- **Prioritäten:** Die Seiten mit den „wichtigen“ Informationen sollten in der Trefferliste weiter vorne erscheinen, wobei noch zu klären wäre, was aktuell das Wichtigere ist.

Heutzutage sind wir es gewohnt, dass Suchmaschinen á la *Google* diese Probleme zwar nicht vollständig lösen, aber die Situation des Nutzers schon verbessern. Wenn wir heute eine Menge von Stichwörtern bei einer Suchmaschine eingeben, dann erwarten wir mittlerweile, dass die Suchmaschine „errät“, was wir eigentlich meinen, und selbstständig eine Entscheidung trifft, was „eigentlich wichtig“ ist. Somit entsteht beim Nutzer der Eindruck, dass die Suchmaschine „versteh“, was wir meinen, und dass die Suche mehr ist als eine einfacher Abgleich von Stichwörtern.

## Implizit vs. explizit

Allerdings muss auch mit diesen Verbesserungen ein erheblicher Teil der ▶

**Transport von Semantik**

Bei der Erstellung von semantischen Systemen ist es wichtig zu wissen, dass Semantik selber nicht transportiert werden kann. Das, was transportiert wird, ist ein Symbol (Zeichen), welches das semantische Konzept repräsentiert. Die Erkenntnis der Bedeutung selber entsteht im Bewusstsein des Betrachters (eine etwas detailliertere Betrachtung findet sich in [Dos04-b]).

Für die Akteure im *Semantic Web* ergeben sich daraus einige Aspekte, die bei der Erstellung von semantischer Information berücksichtigt werden sollten.

- *Schärfe der Definition einer Ressource:* Zum einen sollte die Definition einer Ressource – beispielsweise durch ein Dokument – nicht zu einschränkend sein. Sollte in einem semantischen System der Begriff des Architekten definiert sein als jemand, der den Bauplan eines Gebäudes entwirft, dann wird die Analogie zu einem IT-Architekten nicht zu Stande kommen. Andererseits ist eine zu „offene“ Definition einer Ressource unter Umständen nicht hilfreich. Zum Beispiel ist nicht jede Frau mit einem Kind notwendigerweise auch eine Mutter.

- *Interpretationsspielraum einer Definition:* Im Grunde handelt es sich hier um ein Detail des vorangegangenen Aspekts. Allerdings ist es wichtig, dass bei der Erstellung von semantischen Systemen berücksichtigt wird, dass die meisten textuellen Definitionen durch andere unterschiedlich interpretiert werden können.

Diese Situation ist durchaus vergleichbar mit der eines Klassendesigners für eine objektorientierte Anwendung. Es geht darum, eine geeignete Festlegung zu finden. Was „geeignet“ bedeutet, ist eine hochgradig nicht-triviale Fragestellung, deren Komplexität stark mit der Mächtigkeit des Systems korreliert.

*Kasten 1: Vom Klassen- und Semantik-Design*

Selektion der Ergebnisse vom Nutzer selbst durchgeführt werden. Ein echter Fortschritt kann nur erreicht werden, wenn die Bedeutung der Seiten, die im Web zur Verfügung stehen, automatisch verarbeitet werden kann. Prinzipiell lassen sich zwei Ansätze unterscheiden:

- *Nachträgliche Erschließung implizit vorhandener Bedeutung:* Die einer Suchmaschine zu Grunde liegenden

– bzw. dessen Anwendung der Suchmaschine. Je mehr signifikante Stichwörter zur Suche vorgegeben werden, desto exakter kann das Suchergebnis ausfallen

- *Auswertung explizit vorhandener semantischer Information:* Dieser Ansatz ist dem vorherigem genau entgegengesetzt. In diesem Fall ist es die Aufgabe des Anbieters, die semantische Information in maschinen-verwertbarer

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 3.2 Final//EN">
<html>
<head>
</head>
<body>
<h1>Stadt: Paris </h1>
</body>
</html>
```

*Listing 1: Ein einfaches HTML-Beispiel mit impliziter Semantik*

Algorithmen gestatten es dabei, aus der Kombination von Wörtern den Bedeutungszusammenhang (Kontext) zu ermitteln. Der Vorteil dieser Vorgehensweise ist, dass sich der Bereiter der Daten (z. B. einer Web-Seite) keine Gedanken darüber machen muss, wie sich ein Suchender diese erschließen kann. Die gesamte Erkenntnisarbeit liegt beim Suchenden

Form aufzubereiten. Der Suchende muss seine Abfrage „nur noch“ in geeigneter Form formulieren und benötigt keine semantisch „intelligenten“ Algorithmen.

Der erste Ansatz ist in der Informatikdisziplin der Künstlichen Intelligenz (KI) beheimatet. Mit Hilfe von „intelligenten“ Anwendungen wird die Bedeutung erschlossen bzw. interpretiert. Dieses

Vorgehen ist zwar gut untersucht und kann gute Resultate vorweisen; allerdings bleibt das auch aus der zwischenmenschlichen Kommunikation bekannte prinzipielle Problem bestehen, dass bestehende Daten auf vielfältige Weise interpretiert werden können. Zudem wird die gesamte Last der Erkenntnisgewinnung auf die Datenverarbeiter abgewälzt.

Im Rahmen des *Semantic Webs* wird deshalb der zweite Ansatz – das Auswerten von explizit vorhandener semantischer Information – verfolgt. Die Information (hier Semantik) ist dabei in standardisierter Form zugänglich und kann bei Bedarf genutzt werden. Bedeutung ist gewissermaßen als Meta-Information vorhanden. Die Nutzung setzt voraus, dass der Anbieter von Informationen (z. B. Web-Seiten-Anbieter) die semantische Information selbst zur Verfügung stellt. D. h. der Anbieter im Web tritt in Vorleistung, wobei er jedoch gegenwärtig noch nicht sicher sein kann, dass sich dieser Aufwand lohnt. Aufgrund der Forderung der Kompatibilität des *Semantic Webs* zum *Browser Web* können die Anwender diese Information schlicht ignorieren, wenn sie für sich keinen Nutzwert in dieser Information sehen. Zwar muss auch in diesem Szenario der Anbieter seine Anwendung dergestalt erweitern, dass die angebotene semantische Information ausgewertet werden kann. Der Vorteil gegenüber der im ersten Ansatz implizit vorhandenen Semantik ist, dass die Algorithmen zur Auswertung „lediglich“ Regeln auswerten müssen, was im Vergleich zu den Algorithmen mit kognitiven Aspekten vergleichsweise einfach ist.

Um Semantik explizit in einer Form so beschreiben zu können, dass sie von beliebigen anderen Nutzern ausgewertet werden kann, bedarf es einer bzw. mehrerer Übereinkünfte. Sinnvollerweise sollten diese herstellerunabhängig und in Form allgemein akzeptierter Standards realisiert werden. Zurzeit konstituiert sich das *Semantic Web* aus den Spezifikationen XML, URI, RDF und OWL. Mit Ausnahme von XML werden wir diese im Folgenden näher betrachten.

**Identifizierung von Ressourcen**

Mit Hilfe eines *Uniform Resource Identifier (URI)* wird eine Ressource im Web

<sup>1)</sup> Darunter fallen auch Tippfehler und Dialekte.



eindeutig bezeichnet. Eine Ressource im Sinne des *Semantic Webs* ist alles, was in eindeutiger Weise benannt werden kann. Dies umfasst sowohl reale Objekte (Personen, HTML-Seiten etc.) als auch abstrakte Begriffe und Konzepte (Frieden, Freiheit etc.). Syntaktisch besteht jede URI aus drei Teilen:

```
<schema>:<schema-specific-part>#<fragment>
```

Mit Hilfe von schema kann ein Namensraum (d. h. Gültigkeitsbereich) definiert werden, für den die URI Wirksamkeit besitzen soll. Im Web finden gegenwärtig die Schemata „http“, „ftp“ und „mailto“ Anwendung. Innerhalb des Gültigkeitsbereiches eines Schemas kann durch den schema-specific-part ein beliebiger bestimmter Teil bezeichnet werden. Durch diesen Anteil kann der URI-Verwalter seinen Namensraum nahezu unbeschränkt unterstrukturieren.

Alle nach dem separierenden Nummernsymbol angegebenen Teile werden relativ zum schemaspezifischen Anteil interpretiert. Bekannteste Verwendung hierfür dürften die Fragment-Identifikatoren zur Adressierung von Teilen einer HTML-Seite sein. Aufgrund der Interpretation der URI als Name eines Konzepts – und nicht als eine physische Ressource – muss eine URI im Gegensatz zu einer URL nicht zwingend auflösbar sein. Allerdings werden URLs gerne zur Definition von URIs verwendet, da für die Internet-Domänen eine Infrastruktur existiert, mit deren Hilfe gewährleistet werden kann, dass eine URI eindeutig ist.

Die Einführung des URI-Konzepts war ein wichtiger Schritt zur Beschreibung von Bedeutung. Damit ist es möglich, eine beliebige Ressource (z. B. die Stadt Paris über die URI [www.paris.fr](http://www.paris.fr)) zu beschreiben. Die Wahl der URL als URI hat für uns Menschen den Vorteil, dass wir unter dieser Adresse nachlesen können, dass es sich um die Hauptstadt Frankreichs handelt.

### Beziehungen zwischen Ressourcen

Allerdings wird diese Information einer Anwendung nicht zugänglich sein, da diese als einfacher Text vorliegt – mithin nur implizit vorhanden ist. Gemäß dem Ansatz des *Semantic Webs* müssen jedoch die Bedeutungsaussagen explizit verfügbar sein.

Um dies zu ermöglichen, wird ein weiteres Konzept benötigt, das es gestattet, Ressourcen zueinander in Beziehung zu

setzen. Im Rahmen des *Semantic Webs* wird dazu das *Resource Description Framework (RDF)* genutzt (für eine detailliertere Beschreibung siehe [Dos04-b]). Mit dessen Hilfe können Aussagen der Art „Paris ist die Hauptstadt von Frankreich“ standardisiert ausgedrückt werden. Im Konzept von RDF lassen sich alle Aussagen als Tripel der Form *{Subjekt\Prädikat\Objekt}* darstellen. Die Bezeichnung einer Ressource als Subjekt bzw. Objekt bringt zum Ausdruck, dass es sich bei den in RDF formulierten Beziehungen (Prädikat) um gerichtete Graphen – von Subjekt zu Objekt – handelt. Die eingangs formulierte Aussage über Paris ließe sich somit als Tripel der Form

(1) *{Paris\ ist die Hauptstadt von\ Frankreich}*

darstellen. Dabei ist jedoch zu beachten, dass aus Sicht einer Anwendung statt der Bezeichner URIs zu verwenden sind. Somit ist

(2) *{[www.paris.fr](http://www.paris.fr)\ urn:city#capital\ <http://de.wikipedia.org/wiki/Frankreich>}*

das, was eine Anwendung „sehen“ könnte.

Die RDF-Spezifikation sieht vor, dass Ressourcen beliebig viele Verknüpfungen zu anderen Ressourcen haben können. Sogar Aussagen über Aussagen (*reification*) sind vorgesehen. In der Spezifikation zu RDF werden drei Darstellungsformen definiert, die für unterschiedliche Nutzungszwecke spezialisiert sind:

- **Graphische Darstellung:** In der graphischen Darstellung werden Ressourcen als Ellipsen dargestellt und Beziehungen zwischen Ressourcen als gerichtete Kanten (Pfeile). Die Aussage in (1) und (2) stellt sich somit wie in **Abbildung 2** gezeigt dar. Der Vorteil dieser Darstellung ist, dass sie für einen Menschen besonders bei komplexen Aussagen übersichtlicher ist als andere Darstellungen.
- **N-Tripel:** Dies ist eine Möglichkeit der Serialisierung der graphischen Darstellung. Vornehmlich wird diese Darstellung in der W3C-Gemeinde genutzt.
- **RDF/XML:** Für die Serialisierung in Form eines XML-Dokuments wird das Format RDF/XML angeboten. Die Aussage in **Abbildung 2** stellt sich in

RDF/XML wie in **Listing 2** gezeigt dar. Vorteil des RDF/XML-Formats ist, dass es sich um ein XML-Dokument handelt und damit gut in bereits bestehende Techniken des W3Cs einfügt, wie z. B. Web-Services.

Mit der Spezifikation von RDF wurde nicht der Versuch unternommen, alle möglichen Beziehungen, die zwei Ressourcen verknüpfen können, inhaltlich zu definieren. Die Erfahrung hat gezeigt, dass ein solcher zentralistischer, allumfassender Anspruch nicht durchführbar ist. Vielmehr wurde RDF so konzipiert, dass das vorgegebene Vokabular – welche Beziehungen und Ressourcen sind in meinem Kontext mit einer bestimmten Semantik vorbelegt – von RDF um eigene Vokabeln erweitert werden kann<sup>2)</sup>. Damit dieses Erweitern wiederum standardisiert vonstatten geht, wurde dazu eigens die Spezifikation „RDF-Schema“ erstellt. Zur Veranschaulichung haben wir in unserem Beispiel (**siehe Abb. 2**) die Beziehung „ist die Hauptstadt von“ eingeführt. Mit dieser Vokabel haben wir ein Prädikat definiert, das in unserem Kontext sinnvoll erscheint. Um noch mal zu verdeutlichen, dass URIs zur Anwendung kommen, haben wir keine URL zur Identifikation verwendet. D. h. wenn dem Leser oder der Maschine nicht aus der Wahl des Bezeichners intuitiv klar ist, um welche Art von Beziehung es sich hier handelt, dann muss er sich diese Information an anderer Stelle beschaffen.

Im Rahmen dieses Artikels werden wir nicht weiter auf die syntaktischen Aspekte von RDF eingehen. Das bisher Gesagte reicht aus, um zu verstehen, wie RDF und RDF-Schema dabei helfen, Beziehungen zwischen Ressourcen explizit darzustellen. Für eine weitere Vertiefung soll hier auf die entsprechende Literatur verwiesen werden (z. B. [Bri04]).

### Ontologien

Auch nach der Einführung von RDF treten noch Schwierigkeiten auf, wenn verschiedene, voneinander unabhängige Domänen auf semantischer Ebene miteinander kommunizieren wollen:

- Dieselbe Ressource kann über verschiedene URIs identifiziert sein (Synonyma).
- Ressourcen haben mitunter hierarchische Beziehungen oder Teilmengenbeziehungen. ▶



Abb. 2: Graphische Darstellung einer Aussage in RDF-Grammatik

■ Semantik muss aus vorhandener Information erschlossen werden (Interferenz).

Diese Schwierigkeiten können mit Hilfe von Ontologien gelöst werden. Der Begriff Ontologie ist aus der Philosophie entliehen und wurde von Informatikern, vornehmlich denen aus dem Forschungsgebiet der KI, mit neuer Bedeutung überladen.

Im Kern ist eine Ontologie ein hierarchisch geordnetes Beziehungssystem mit Regeln. In **Abbildung 1** sind die bekanntesten ontologischen Systeme wie *DAML (DARPA Agent Markup Language)* und *OIL (Ontology Inference Layer)* aufgeführt. Beide wurden inzwischen zu *DAML+OIL* vereinigt, das wiederum in den W3C-Standard der *Web Ontology Language (OWL)* überführt wurde. OWL (vgl. [McG04]) kann in diesem Sinne als Erweiterung von RDF betrachtet werden. Als Basis wurde die RDF-Grammatik genutzt.

In unserem obigen Beispiel haben wir Paris eindeutig mit Hilfe der URI `http://www.paris.fr` identifiziert. Für einen anderen Nutzer könnte es jedoch viel sinnvoller sein, Paris, die Hauptstadt Frankreichs, mit der URI `http://de.wikipedia.org/wiki/Paris` zu identifizieren. Damit eine Anwendung die Identität nachvollziehen könnte, muss eine Beziehung zwischen beiden Ressourcen hergestellt werden. Zu diesem Zweck könnte für unser Beispiel

eine Ontologie erstellt werden, die explizit die Identität der beiden Ressourcen darstellt. Die URIs bleiben nach wie vor verschieden und verbleiben unter der Kontrolle der lokalen Administratoren.

Die zweite Schwierigkeit, dass Ressourcen mitunter hierarchische Beziehung haben, ist eigentlich ein mehr syntaktisches Problem. Falls Sie in einer Suchmaschine nach Hauptstädten suchen, werden Sie Paris finden. Was ist aber, wenn sie nach Städten in Frankreich suchen? Ist eine Hauptstadt auch eine Stadt? Wir haben irgendwann gelernt, dass dies der Fall ist. Wenn der Ersteller eines Vokabulars die Beziehung „Eine Hauptstadt ist auch eine Stadt“ nicht explizit definiert hat, dann wird eine Anwendung *Paris* als Stadt niemals finden. Mit Hilfe einer Ontologie kann diese Beziehung separat zum eigentlichen Vokabular definiert werden und ist damit für eine Anwendung verfügbar.

Eine Ontologie, die sich auf die bisher genannten Aspekte beschränkt, wird in der Regel einfach als Taxonomie bezeichnet. Mit anderen Worten, Taxonomien sind Ontologien, die lediglich über *isSubclassOf*-Beziehungen verfügen. Durch die Einführung von Regeln wird aus einer Taxonomie eine Ontologie.

Auf diesem Wege können Ontologien es erleichtern, dass zwei Programme miteinander interagieren, ohne dass die Interpretationsvorschriften für die invol-

vierten Daten direkt in den Programmen codiert sein müssen. Sie nutzen dazu Metadaten, die beiden Seiten in Form von Ontologien zugänglich sind. Diese beinhalten Ableitungsregeln (Interferenz-Regeln), die für die entsprechende Domäne gelten und daher nicht jedes Mal übertragen werden müssen, sondern meist nur als Verweise angegeben werden. Auf dieser Basis ist es den Programmen möglich automatisch Schlüsse zu ziehen, wie zum Beispiel, dass die Hauptstadt Paris auch eine Stadt ist.

Allerdings ist der Wert einer Ontologie nicht nur durch seine Vollständigkeit oder Mächtigkeit zu definieren. Mindestens ebenso wichtig ist die Akzeptanz bei den Anwendern.

### Anreichern von Web-Seiten mit Semantik

Auch wenn noch keine allgemein akzeptierten und erfolgreichen semantischen Suchmaschinen am Markt präsent sind, lassen sich bereits heute semantische Techniken nutzen. Insbesondere kann die vorausschauende Anreicherungen bestehender präsentationsorientierter Web-Präsenzen um semantische Beschreibungen einen Ausweg aus dem gegenwärtig zu beobachtenden „Henne-Ei-Problem“ aufzeigen. Dieses Problem beschreibt die wechselseitige Abhängigkeit von Produzenten und Konsumenten, die beide auf die Nutzung bzw. das Angebot semantischer Beschreibungen verzichten, da diese noch nicht angeboten bzw. genutzt werden.

Im Grunde genügt bereits die Verfügbarkeit von durch *HTML-Meta-Tags* angereicherten Web-Seiten, um diese durch semantische Aussagen zu ergänzen. Stehen die Seitenquellen in einem maschinell verarbeitbaren Format (etwa XML) zur Verfügung oder sind diese in einem Content-Management-System abgelegt, so erlauben automatisierte Transformationsmechanismen (wie die *XSL-Transformationen*) die Ergänzung bestehender präsentationsorientierter Inhalte um einfache Semantikbeschreibungen.

Eine weitere Möglichkeit sind die *Simple HTML Ontology Extensions (SHOE)*, die selbst als XML-Sprache realisiert sind. Dabei handelt es sich um eine einfache Erweiterung von HTML, welche die Anreicherung von Dokumenten mit semantischer Information durch den

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:my="urn:city#">
  <rdf:Description rdf:about="http://www.paris.fr/">
    <my:capital rdf:resource="http://de.wikipedia.org/wiki/Frankreich/">
  </rdf:Description>
  <rdf:Description rdf:about="http://de.wikipedia.org/wiki/Frankreich/">
  </rdf:Description>
</rdf:RDF>
```

Listing 2: RDF/XML-Darstellung des Beispiels aus Abb. 2

<sup>2)</sup> Wichtig! Nicht RDF wird erweitert, sondern das eigene domänenspezifische Vokabular.



<b>DAML</b>	DARPA Agent Markup Language
<b>HTML</b>	Hypertext Markup Language
<b>HTTP</b>	Hyper Text Transfer Protocols
<b>OIL</b>	Ontology Inference Layer
<b>OWL</b>	Web Ontology Language
<b>RDF</b>	Resource Description Framework
<b>SHOE</b>	Simple HTML Ontology Extensions
<b>URI</b>	Uniform Resource Identifier
<b>URL</b>	Uniform Ressource Locator
<b>XHTML</b>	Extensible HTML
<b>XML</b>	Extensible Markup Language
<b>XSL</b>	Extensible Stylesheet Language

### Kasten 2: Abkürzungen

Autor in einer maschinenlesbaren Form erleichtert. Für die sinnvolle Verwendung ist zu empfehlen, die XML-basierte Fassung von HTML, XHTML, einzusetzen. Im Wesentlichen definiert SHOE lediglich einige Elemente, die es erlauben eine Ontologie zu konstruieren oder Verweise auf externe Ontologien anzugeben. Dabei ist es möglich Instanzen von lokalen und externen Ontologien zu erzeugen und über entsprechende Regeln Ausnahmen zu definieren. Dadurch unterstützt SHOE die Einbettung von Informationen aus einer Vielzahl von Quellen.

### Risiken

Das *Semantic Web* wirft jedoch nicht nur Licht voraus, sondern deutet auch bereits einige Probleme an, die sich aus seiner Einführung ergeben werden.

Schlaglichtartig sollen abschließend drei mögliche Risiken skizziert werden:

- **Sinkende Ergebnisbreite:** Ein nicht unerheblicher Anteil der Ergebnisse einer Web-Recherche wird durch Zufallsfunde zu Tage gefördert. Die qualitativen Eigenschaften des Webs gründen daher sowohl auf der Korrespondenz von „gesucht/gefunden“, aber auch auf „nicht-gesucht/aber dennoch interessant“. Durch automatische Nutzung semantischer

Techniken werden jedoch diese nebenbei aufgefundenen Daten gezielt eliminiert und der filternden menschlichen Interpretation entzogen.

- **Sinkende Vertraulichkeit:** Durch die eineindeutige Referenzierung einzelner Konzepte leidet tendenziell die existierende Anonymität, die sich aus der (in diesem Falle hilfreichen) Mehrdeutigkeit natürlicher Sprache ergibt. So können semantik-basierte Suchmaschinen anhand des in eindeutiger Weise URI-repräsentierten Konzepts zwar zuverlässig das Gewünschte auffinden und liefern, jedoch bedeutet dies im Umkehrschluss auch, dass die gegenwärtig noch verbleibende Unschärfe in der Auswertung des Anwenderverhaltens verloren geht.
- **Sinkende Verlässlichkeit:** Wird auf der Basis semantischer Techniken die Entscheidungsbefugnis der Filterung „relevanter“ oder „interessanter“ Inhalte an eine maschinelle Komponente verlagert, so ist diese – unter Kenntnis dieses Verlagerungsprozesses – angreifbar. In diesem Sinne lädt das Wissen über die Nutzung einer semantischen Suchkomponente beispielsweise nachgerade dazu ein, Inhalte mit falschen Semantikbeschreibungen (d.h. solchen, die nicht die tatsächlich angebotene Ressource beschreiben) zu versehen, um semantische Filterprozesse erfolgreich bestehen zu können.

### Zusammenfassung

Ob beziehungsweise wann die Vision des *Semantic Webs* Wirklichkeit wird, vermag heute niemand mit Bestimmtheit sagen. Dass wir ein wachsendes Problem mit der Informationsflut aufgrund der technischen Möglichkeiten haben, ist unbestritten. Die Ansätze des *Semantic Webs* zeigen eine Möglichkeit auf, einige der Probleme zu lösen. Dies spiegelt sich auch in den Empfehlungen von namhaften Analysten wie Gartner wider, bereits jetzt mit dem Aufbau des Know-hows zu beginnen, da Semantik eines der wichtigen Themen der IT sein wird (vgl. [Gar-02]). ■

### Literatur & Links

- [Ber01] T. Berners-Lee, J. Hendler, O. Lassila, *The Semantic Web*, Scientific America 2001 (siehe: [www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&sc=I100322](http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21&sc=I100322))
- [Bri04] D. Brickley, R.V. Guha (Hrsg.), *RDF Vocabulary Description Language 1.0: RDF Schema*, W3C Recommendation, 2004 (siehe: [www.w3.org/TR/2004/REC-rdf-schema-20040210/](http://www.w3.org/TR/2004/REC-rdf-schema-20040210/))
- [Dos04-a] W. Dostal, M. Jeckle, *Semantik, Odem einer Service-orientierten Architektur*, in: *JavaSpektrum* 01/2004
- [Dos04-b] W. Dostal, M. Jeckle, W. Kriebbaum, *Beschreibung von Semantik*, in: *JavaSpektrum* 02/2004
- [Dos04-c] W. Dostal, M. Jeckle, W. Kriebbaum, *Vokabulare und Ontologien*, in: *JavaSpektrum* 03/2004
- [Dos04-d] W. Dostal, M. Jeckle, *Web Services, der natürliche Träger von Semantik*, in: *JavaSpektrum* 04/2004
- [Gar-02] Gartner Research, *Note T-17-5338*, 20. August 2002
- [McG04] D.L. McGuinness, F. van Harmelen (Hrsg.), *OWL Web Ontology Language Overview*, W3C Recommendation, 2004 (siehe: [www.w3.org/TR/2004/REC-owl-features-20040210/](http://www.w3.org/TR/2004/REC-owl-features-20040210/))
- [Net] Netcraft, die Seite enthält Statistiken über wichtige Kennzahlen, siehe: [news.netcraft.com/archives/2004/04/01/april\\_2004\\_web\\_server\\_survey.html](http://news.netcraft.com/archives/2004/04/01/april_2004_web_server_survey.html)
- [Top] TopQuadrant, Inc., *Die Idee des „Tree of Knowledge Technologies“* stammt von R. Hodgson und J. Stuart von TopQuadrant, Inc (siehe auch: [www.topquadrant.com](http://www.topquadrant.com))