

Die Semantic Web Ernüchterung

oder wie ein CMS Semantik lernt

von Eike Diestelkamp und Benjamin Birkenhake

Hypertext ist mit dem Versprechen angetreten, menschliches Denken und Wissen angemessener abzubilden. Nicht-lineare Inhalte sollten nichtlinear organisierbar sein. Das funktioniert auch in übersichtlichen Systemen, ist aber für das größte aller Hypertextsysteme – das WWW – weit weniger erfolgreich. Im Moment organisiert das menschliche Wissen im Web vor allem die größte aller Suchmaschinen: Google. Und das mit wachsenden Problemen.

Suchmaschinen arbeiten in erster Linie mit statistischen Methoden und Zeichenketten. Das kann keine gute Abbildung von Wissen sein und führt bei wachsenden Datenmengen zwangsläufig zu schlechteren Suchergebnissen. Das „Semantic Web“ tritt an, mehr Ordnung in das Wissen im Web zu bringen um Inhalte und Wissen nicht nur menschen-, sondern auch maschinenlesbar und damit besser verwaltbar zu machen. Dabei ist der eigentliche Ansatz gar nicht neu: Es geht um Metadaten. Was man zuvor in Metatags und Keywords unterbringen wollte, wurde schon 1998 von Sir Tim Berners-Lee persönlich durch die Vision des großen semantischen Netzes ersetzt.

Der Stoff aus dem die Träume sind

Um die heutigen Inhaltsrepräsentationen in Wissensrepräsentationen zu verwandeln, auf denen Maschinen sinnvoll operieren können, bedarf es eines formalen Modells. Das Semantic Web stellt dazu Werkzeuge und Standards zur Verfügung. Dabei sind die Grundkonzepte recht übersichtlich.



Die kleinste Einheit des Semantic Web ist eine Entität. Je nachdem wen man fragt ist eine Entität eine URL, eine Zeichenkette oder eine andere Form von Information. Zwischen Entitäten können getypte Relationen bestehen. Eine Summe von Entitäten und Relationen bezeichnet man auch als Ontologie. Mit diesem Konzept ist es möglich, viele kleine Wissenseinheiten zu einem großen Wissen zu vereinen und zentral verfügbar zu

machen. So wie heute unzählige HTML-Seiten das sicht- und surfbare WWW bilden, sollen in Zukunft XML-Standards wie RDF, RDFS und OWL ein Wissensnetz bilden.

Der Morgen danach

Einer der zentralen Ansprüche des Semantic Web liegt in der Erschließung implizit vorhandenen Wissens. Die dabei einfachste Operation ist eine logische Schlussfolgerung, ein Syllogismus. Man kann Entitäten und Relationen als Aussagen verstehen. Beispielsweise ist es formal korrekt aus den beiden Aussagen (Prämissen) „Jeder Mensch ist sterblich“ und „Aristoteles ist ein Mensch“ die neue Aussage (Schlussfolgerung) „Aristoteles ist sterblich“ zu ziehen. Die Technologien des Semantic Web zielen darauf ab solche Schlüsse (Inferenzen) automatisch zu ziehen. Im Web könnte man folgende Situation vorfinden: Website A definiert „Nirvana“ als eine Rock-Band und Website B definiert „Nevermind“ als ein Album der Band Nirvana. Dem aufmerksamen Leser erschließt sich die triviale und logische Schlussfolgerung intuitiv: „Nevermind“ ist ein „Rock-Album“. Bedenkt man das im Web herrschende Chaos und erweitert das Beispiel in diesem Sinne, offenbaren sich sofort die Schwachstellen des Konzepts: Wenn Website C „Nirvana“ als religiösen Erlösungszustand definiert, dann folgt daraus, dass „Nevermind“ das Werk eines religiösen Erlösungszustands ist, der es in Deutschland auf zwölf Chart-Platzierungen gebracht hat.

Die Autoren von Webseiten sind für verwertbare Metadaten verantwortlich. Mit dem Grad an Detailtreue steigt der Nutzwert, aber auch die Komplexität. Die konsequente Einhaltung von Konventionen zur Formulierung von Metadaten spielt eine zentrale Rolle für die Qualität späterer Operationsergebnisse. Die latente Angst eines jeden Webmasters, trotz oder gerade wegen zu detaillierter Metadaten vielleicht nicht mehr oft genug gefunden zu werden, macht es offenbar schwierig den Motor des Semantic Web zum Laufen zu bringen. Wenn wir ehrlich sind werden Meta-Tags heute eher unter dem Gesichtspunkt einer optimalen Suchmaschinenplatzierung formuliert als durch altruistische Tugenden motiviert. Es wird sich zeigen in welcher Form Technologien des Semantic Web einen tatsächlichen Mehrwert für das gesamte Web leisten werden.

Ein weiteres Problem für das Semantic Web stellt der hochgradig generische Ansatz dar. In HTML ist noch klar, dass das Tag „h1“ eine Überschrift ersten Grades auszeichnet. Die Interpretation der Semantik einer RDF-Aussage findet hingegen in der Implementierung statt und wird nur geringfügig von den Standards des W3C-Konsortiums vorgeschrieben. Engere Konventionen zur Interpretation schränken die universelle Einsetzbarkeit zur Abbildung vielfältiger Wissensbereiche deutlich ein. Als Konsequenz daraus kann es zum einen zu mangelnder Austauschbarkeit der Informationen zwischen unterschiedlichen Wissensinseln kommen. Zum anderen besteht die Gefahr, dass ein Monopolist mit einer miserablen Implementierung die Ausschöpfung des Semantic-Web-Potenzials vereitelt.

Dem hehren Konzept drohen allerdings weit profanere Probleme als die bisher beschriebenen. Wenn man sich erinnert aus welchen Gründen das Meta-Tag den Tod der Bedeutungslosigkeit gestorben ist wird klar, dass dem Semantic Web ein ganz ähnliches Schicksal droht. Dem Missbrauch semantischer Information zur Förderung des Umsatzes von zweifel-

haften Waren hat auch das Semantic Web nur wenig entgegengesetzt. Aus dem Tag-Abuse in HTML wird RDF-Abuse.

„Unfortunately, much of the Web is like an anthill built by ants on LSD.“

- Jakob Nielsen -

Semantic Web und Content Management

Das Semantic Web als großes Ganzes, als die zweite Ausbaustufe des Internet, ist noch in weiter Ferne. Wir können heute nur schwer abschätzen ob es das Semantic Web, wie es sich Tim Berners-Lee wünscht, geben wird. Das heißt aber nicht, dass die Konzepte dahinter schlecht oder sinnlos wären. Im Gegenteil: Content Management Systeme (CMS) sind bereits auf einem guten Weg, nur eben nicht zu einem Semantic Web, sondern zu vielen semantischen Netzen. Die Konzepte hinter dem Semantic Web lassen sich in kleineren, geschlossenen Systemen – wie sie häufig von Content Management Systemen verwaltet werden – sogar deutlich besser etablieren als im Großen. Gängige CMS-Lösungen bringen bereits ein paar Tugenden für zukünftige Semantic-Web-Systeme mit. Für einen sinnvollen Einsatz von Semantic-Web-Technologien (SW-Technologien) gibt es aber Grundvoraussetzungen.

Eigentlich könnte man in einem CMS fast alle Daten mit SW-Technologien implementieren: die Benutzer-Verwaltung, die Rechteverwaltung, die Workflow-Verwaltung und selbstverständlich die Content-Verwaltung. Das ist allerdings nur sehr bedingt sinnvoll. SW-Technologien einzusetzen bedeutet in der Regel höhere Anforderungen an Personal und Software. Daher sollte sich der erhöhte Einsatz gerade in geschlossenen Systemen lohnen. Die internen Verwaltungsprozesse eines CMS mit SW-Technologien zu realisieren führt zwar zu einer stark generischen System-Architektur, die auch ihre Vorteile hat. Aber am Ende des Tages sind semantische Netze eher etwas das mit Inhalten zu tun hat.

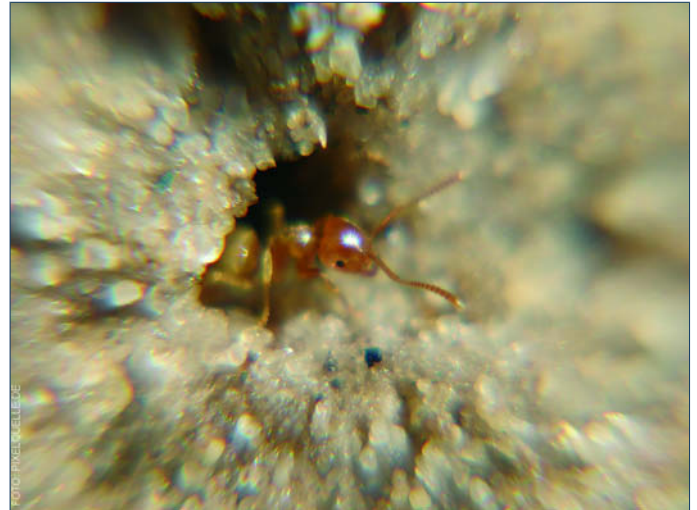
Vom Stichwort zum Wissensnetz

Im Bereich der Inhaltsverwaltung gibt es in bestehenden Content Management Systemen bereits Semantic-Web-Konzepte, fast ohne Kenntnis der Entwickler: Schlag- und Stichwortkataloge mit denen man einzelne Inhalte auszeichnen kann finden sich in praktisch jedem besseren CMS. Ein solcher Katalog kann unter bestimmten Umständen als kontrolliertes Vokabular bezeichnet werden und ist dann eine sehr simple Form einer Ontologie. In einem Musikportal könnte es einen Schlagwortkatalog mit dem Titel „Genre“ geben, der Begriffe wie „Jazz“, „Rock“, „Pop“, „Elektro“, „Hip Hop“ und „Klassik“ enthält.

Damit endet in der Regel die Funktionalität der meisten CMS, nicht aber die Wünsche der Benutzer. Schnell würde sich eine Beispielredaktion genötigt fühlen, den Schlagwortkatalog um „Heavy Metal“, „Stoner Rock“, „Grunge“, „House“, „Techno“ oder „Drum and Bass“ zu ergänzen. Die Einordnung dieser Begriffe neben die bereits bestehenden Begriffe ist ebenso naheliegend wie falsch, denn „Grunge“ ist offensichtlich ein Unterbegriff von „Rock“. Ist ein Artikel über ein Album als „Grunge“ gekennzeichnet so wird er nicht gefunden, wenn ein Besucher nach „Rock“ sucht, obwohl es sich doch dabei um Rock-Musik handelt. Für dieses Problem gibt es eine richtige und eine falsche Lösung. Die Falsche wäre, den Artikel zusätzlich als „Rock“ auszuzeichnen. Das würde bereits kurzfristig in

einen erheblichen Mehraufwand für die Redakteure ausarten, was praktisch der Untergang jedes Semantic-Web-Konzepts ist. Der Einsatz von SW-Technologien steht und fällt mit der Disziplin der Redakteure und Autoren, auch in geschlossenen Systemen.

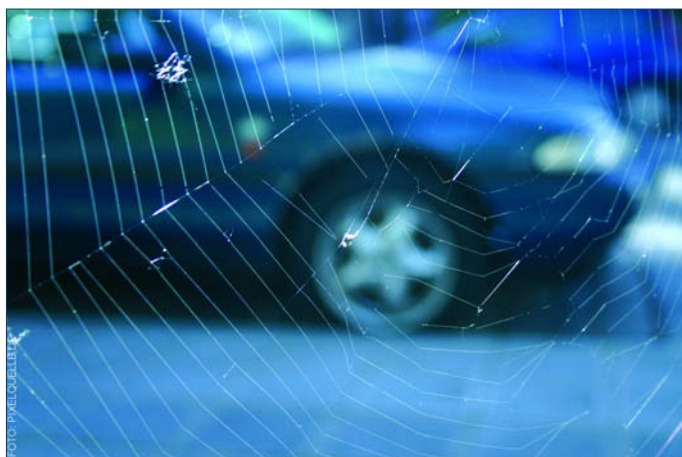
Die richtige Lösung wäre also einen Baum aufzubauen, in dem „Grunge“ als Unterbegriff von „Rock“ eingeordnet ist. Die Baumstruktur erlaubt dem CMS alle als „Grunge“ ausgezeichneten Artikel zu präsentieren wenn nach „Rock“ gesucht wurde. Diese baumartige Taxonomie ist schon eine etwas reichere Ontologie. Die Entitäten dieser Ontologie sind einzelne Genre und die Äste stellen eine Oberbegriffs-Unterbegriffs-Relation dar.



Der nächste Schritt ist ebenso logisch wie einfach: Aus dem Baum wird ein Netz. Hat man, wie unsere Beispielmusikredaktion, ein sehr differenziertes Verständnis von einem Themenbereich, empfindet man einen Baum zur Strukturierung dieses Bereichs nach einiger Zeit nicht mehr als angemessen. Zum einen kann es mitunter schwer sein einen neuen Begriff eindeutig in dem Baum einzuhängen. Das Genre „Crossover“ verrät bereits im Namen, dass es mehrere Ursprünge hat. Man kann „Crossover“ nicht einfach entweder unter „Rock“ oder unter „Hip Hop“ einhängen. Zum anderen verbirgt die Hierarchie andere, diffizilere Zusammenhänge, wie etwa die chronologische Ordnung. Die Lösung in diesem Fall ist ein semantisches Netz mit unterschiedlichen, getypten Relationen. Die Typen der Relationen sind wichtig damit das System weiterhin Schlüsse ziehen kann, wie: Wer nach „Rock“ sucht sollte auch „Crossover“ finden.

Wo Semantic Web drauf steht muss nicht immer XML drin sein

Verwendet man Semantic-Web-Konzepte in einem geschlossenen System, spielt es kaum eine Rolle für welches Speicherformat man sich entscheidet. In einem solchen Rahmen spricht in der Tat einiges für Datenbanklösungen: Das generische Konzept des Semantic Web lässt sich sauber und deutlich performanter in eine Datenbank bringen. Und da der Mehrwert semantischer Arbeit nur an den Besucher und die Redakteure unseres Portals weitergegeben wird, entfällt der Bedarf die Daten über XML zu syndizieren. So kann man Semantic-Web-Funktionalität in bestehende CMS-Architekturen integrieren, die in der Regel auf relationalen Datenbanken aufbauen.



Content Management Systeme haben aber neben den eben beschriebenen Metadatenkonzepten eine weitere „natürliche“ Verbindung zum Semantic Web. Legt der Redakteur einer Musikredaktion eine neue Plattenkritik an, so wird er sich in der Regel einem Formular gegenübersehen, das unter anderem Felder wie „Interpret“, „Titel“ und „Plattenfirma“ enthält. Bei einem guten CMS sollte man wenigstens bei den Feldern „Interpret“ und „Plattenfirma“ die bereits von anderen Redakteuren eingetragenen Interpreten und Label zur Auswahl bekommen. Der hier ursprüngliche Zweck – die Vermeidung von Redundanzen – führt dazu, dass die Daten bereits in Bezug auf ihre Semantik im System gehalten werden. Content Management Systeme fangen semantische Entitäten auf und verwalten diese in bestimmten Kontexten. In unserem Beispielportal hätte so ein Besucher die Möglichkeit, durch alle Artikel nach einem Interpreten zu suchen.

Damit hat man bereits ziemlich viele Entitäten für eine Ontologie beisammen. Jetzt fehlen noch die Relationen. Auch diese werden teilweise bereits durch die vorgegebenen Formulare implizit erzeugt und lassen sich explizit ausdrücken: Schreibt ein Redakteur eine Plattenkritik zu Nirvanas Album „Nevermind“, ließen sich daraus die Entitäten „Nirvana“ und „Nevermind“ generieren. Außerdem die Relationen „Nirvana ist ein Interpret“, „Nevermind ist ein Werk“ und „Nirvana ist der Interpret von Nevermind“. Diese Entitäten und Relationen können zusätzlich zu der neuen Plattenkritik und automatisch in das semantische Netz des Portals übernommen werden. Dort können dann die Entitäten von Hand über weitere Relationen mit anderen Entitäten in Verbindung gesetzt werden, wie z.B. „Foo Fighters sind Nachfolgebänd von Nirvana“.

Da es sich hier um eine kontrollierte Umgebung handelt, ist keine Verwechslung mit religiösen Erlösungszuständen zu befürchten.

Eine weitere Anwendungsmöglichkeit von semantischen Netzen in geschlossenen Systemen ist ein semantisches Glossar. Jede Entität wird dabei automatisch als ein Glossar-eintrag exportiert. Die mit dieser Entität über Relationen verknüpften Entitäten werden ebenfalls und mit der Art ihrer Relation zum eigentlichen Eintrag präsentiert. Dabei unterstützt die graphische Präsentation durch die Position der assoziierten Entitäten die Interpretation des Benutzers. So steht der zentrale Begriff hier in der Mitte. Oberbegriffe finden sich darüber, Unterbegriffe darunter. Aus dem Glossar heraus kann dann wiederum automatisch zu einzelnen Inhaltsobjekten verlinkt werden die mit dem Begriff verbunden sind. Die Erzeu-

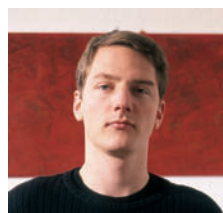
gung des gesamten semantischen Glossars ist vollständig automatisierbar.

Die Verhältnismäßigkeit der Mittel

Unser Beispiel zeigt einen Aspekt von Semantic-Web-Lösungen im CMS-Kontext: Es macht in der Regel erst dann Sinn wenn man es mit verhältnismäßig viel und verhältnismäßig wertvollem Content zu tun hat. Für Malermeister Müller kann ein CMS zwar eine gute Lösung sein, weil er so seine Inhalte selber pflegen kann, aber er wird wohl kaum ein Schlagwort-Netz von dem Umfang brauchen wie eine Musikredaktion. Die Frage, ab wann der Einsatz von SW-Technologien in Content Management Systemen sinnvoll ist, bleibt aus Ermangelung einer weiten Verbreitung dieser Technologie bisher offen. Sicher ist, dass Verlage, die große Mengen Wissen und vielfach wiederverwendbare Inhalte verwalten, bereits in der einen oder anderen Form auf SW-Technologien setzen.

Sicher ist auch, dass die Konzepte hinter dem Semantic Web in erster Linie Metadaten-Konzepte sind. Und eins muss man dem W3C-Konsortium lassen: Über Metadaten haben sie sehr gründlich nachgedacht. Hat man also die Anforderung, komplexe Metadaten zu bestimmten Content-Einheiten für ein geschlossenes System zu verwalten, wie es etwa im Knowledge-Management oder im E-Learning der Fall ist, drängt es sich geradezu auf, Konzepte und Know-how aus der SW-Technologie zu verwenden. Ob man dafür auf die Formate des W3C zurückgreift oder eigene Implementierungen verwendet ist am Ende des Tages fast irrelevant. Wichtig ist in erster Linie die saubere Implementierung der richtigen Konzepte innerhalb des eigenen Systems. Ist dies gewährleistet, ist die Distribution der semantischen Inhalte in vielfältige Semantic-Web-Formate jederzeit mit geringem Aufwand möglich. So kann dann die geschlossene Wissensinsel doch noch Teil des großen Semantic-Web-Ozeans werden.

DER AUTOR



Eike Diestelkamp hat Texttechnologie an der Uni Bielefeld studiert. Als Geschäftsführer von HDNET berät er seit sechs Jahren Kunden bei nationalen und internationalen Webprojekten. Zu seinen Interessensgebieten zählen Markup-Sprachen und Verfahren zur Hypertextanalyse.

DER AUTOR



Benjamin Birkenhake M.A. hat als Dozent im Bereich Texttechnologie der Uni Bielefeld gearbeitet und ist Gründer des Digitalkombinats. Derzeit promoviert er über den Einsatz von XML-Technologien in den Geisteswissenschaften.