Topic Maps – Semantische Verknüpfungen für Sammlungen

Topic Maps – Semantic Links for Collections

Lars Bröcker
Fraunhofer-Institut für Medienkommunikation
Competence Center NetMedia
Schloss Birlinghoven
53754 Sankt Augustin
Tel.: 02241 14 1993, Fax: 02241 14 2597
E-Mail: Lars.Broecker@imk.fraunhofer.de, Internet: www.imk.fraunhofer.de

Zusammenfassung:

Immer mehr Bestände von Museen und Archiven werden online zur Verfügung gestellt. Mit dem Wechsel in das neue Medium können neue Benutzerkreise erreicht und für die Arbeit der Einrichtungen interessiert werden. Damit geht einher, die Navigation im Angebot zu erleichtern. Denn die Suche in digitalen Archiven orientiert sich oft am vorliegenden Katalog. Eingabemasken, die Detailkenntnis über die Sammlung erfordern, schrecken neue Nutzer ab oder vermitteln durch Fehleingaben ein falsches Bild der verfügbaren Inhalte. Trefferlisten enthalten keine Hinweise, wie ein gefundenes Dokument zu anderen der Sammlung in Beziehung steht, die Fülle des verfügbaren Wissens bleibt verborgen. Der ISO Standard *Topic Maps* verspricht Abhilfe. Inhalte können in Beziehung gesetzt werden, wodurch der Kontext der Sammlung fassbar und im WWW darstellbar wird. Nutzer können sich anhand der Semantik durch die Sammlung bewegen und sich so das angebotene Wissen besser erschließen.

Abstract:

Museums and archives keep on digitising their collections in order to make them available on the WWW. The move to this new medium offers the chance to attract new user groups and to interest them in the work of the institutions offering the service. This requires easier means of navigation in the collection. Search interfaces in digital archives tend to honour the bounds of existing catalogues. Search forms that require intimate knowledge of the makeup of the collection discourage new users and are prone for wrong entries. Result sets include no information regarding the connection between results, thus much of the knowledge inside the collection remains hidden. The ISO standard named *Topic Maps* comes to the rescue. It can define associations between pieces of a collection, which adds contextual information to the collection in the form of a semantic network. This network can be navigated and allows for easier understanding of the meaning of each piece of a collection.

1. Motivation

Im Verlauf der letzten Jahre hat sich das Internet - und darin vor allem das World Wide Web - zur größten Wissenssammlung der Welt entwickelt. Selbst kleinere Museen, Sammlungen oder Archive stellen zumindest Teile ihrer Daten in digitaler Form zur Verfügung. Darin liegt einerseits die Chance, dass die Vision von "'Information at your fingertip"' wahr wird, andererseits wird jedoch der Bedarf nach effizienten Suchmöglichkeiten zum Auffinden all dieser Informationen im Internet immer größer. Denn das World Wide Web wächst: Im Jahr 2000 gab es allein 2,5 Milliarden unterschiedlicher Webseiten und heute dürfte diese Zahl schon weit übertroffen sein[1]. Nach der jüngsten Studie der Internet Software Society, gab es im Januar 2003 rund 172 Millionen Websites, eine Steigerung von 16% gegenüber 2002 und sogar 57% gegenüber 2001 [2]. Heute gebräuchliche Suchmaschinen helfen bei der Menge an möglichen Fundorten nur bedingt weiter,

da die Qualität ihrer Ergebnislisten sowohl von den verwendeten Anfragetermen, als auch von der Internet-Abdeckung der verwendeten Suchmaschine abhängig ist. Damit kommen zwei Probleme zusammen: Heutige Suchmaschinen können die Masse an Webseiten nicht nur nicht mehr bewältigen, sie haben außerdem keinen Zugriff auf die Semantik, die den Webseiten inne wohnt. Dadurch verschlechtert sich die Qualität der Suchergebnisse immens, denn es zählt nur das Vorhandensein der Suchbegriffe, nicht der Kontext, in dem sie verwendet werden.
Was also eigentlich bei der Suche im Internet benötigt wird, ist eine Art Atlas, in dem sich Fundorte für Informationen themenorientiert nachschlagen und eingrenzen lassen. Dies setzt aber eine stärkere semantische Unterfütterung des WWW voraus.
Das World Wide Web Consortium koordiniert ein Projekt, das eben diese Unterfütterung zum Ziel hat und schon selbst zum Schlagwort geworden ist: Das *Semantic Web*. Dieses Projekt hat das ambitionierte Ziel, Software-Agenten in die Lage zu versetzen, die Semantik von Webinhalten zu erfassen und gemäß ihrem Auftrag auszuwerten. Dies erfordert noch umfangreiche Arbeiten, besonders auf dem Gebiet der Beschreibungssprachen. Planungen gehen davon aus, dass das Projekt 2010 zum Abschluss gebracht werden kann.
Daneben gibt es allerdings einen anderen Ansatz, mit dem bereits heute die Semantik eines Datenbestandes modelliert werden kann: Die so genannten *Topic Maps*, die im Jahr 2000 von der International Standards Organisation (ISO) standardisiert wurden. Mit Hilfe dieses Standards lassen sich Objekte, Konzepte und ihre semantischen Beziehungen zueinander modellieren. Dabei entsteht ein semantisches Netz, das den Datenbestand wesentlich besser beschreibt, als dies bisher der Fall ist. Dieses Netz lässt sich für die Realisierung von Navigations- oder Suchprogrammen verwenden, die von den modellierten Beziehungen profitieren können und somit nicht nur auf Volltextsuchen angewiesen sind.

Dieses Paper geht genauer auf die Konzepte ein, die hinter den Topic Maps stehen. Kapitel zwei beschreibt die Grundlagen des ISO-Standards zu Topic Maps, Kapitel drei skizziert mögliche Anwendungen. In Kapitel vier wird auf Forschungsthemen in Zusammenhang mit dem Standard eingegangen. Das Paper schließt mit einem Ausblick auf die weitere Entwicklung des Gebiets.

2. Grundlagen des Standards

Im Jahr 2000 hat die International Standards Organisation, die Dachorganisation der nationalen Standardisierungseinrichtungen, den Standard *13250:2000 Topic Maps*[1] verabschiedet. Darin werden Mechanismen für die Modellierung von Objekten und ihren Beziehungen untereinander definiert. Der dabei entstehende Graph bildet dann die Topic Map oder T*hemenkarte* bezeichnet.

Der Standard kommt mit wenigen Instrumenten aus, um seinen Zweck zu erfüllen. Drei Grundbausteine werden für die Erstellung von Topic Maps benötigt: Topics, Occurrences und Associations. Diese werden im Folgenden beschrieben.

Topics: Ein Topic ist ein Thema, das innerhalb einer Topic Map modelliert werden soll. Dabei kann es sich um Personen, Dokumente, Objekte der realen Welt oder auch abstrakte Konzepte handeln. Topics können beliebig viele Namen zugewiesen werden, so dass z.B. Varianten eines Namens abgebildet werden können. Sie können außerdem typisiert werden, so dass sich komplexe Hierarchien aufbauen und abbilden lassen. Die Typen sind selbst wiederum Topics und ein Topic kann mehrfach typisiert sein.
Für den Verweis auf Informationen oder die eigentlichen Inhalte eines Topics dienen Occurrences, die in beliebiger Anzahl einem Topic zugewiesen werden können.

Occurrences: Eine Occurrence ist ein Verweis auf ein Vorkommen von Informationen zu einem Topic, üblicherweise in der Form eines Hyperlinks auf im Web erreichbare Ressourcen. Prinzipiell kann eine Occurrence aber auch direkt Textinformationen aufnehmen, ein Datum oder eine Zusammenfassung eines Dokuments könnte also direkt in einer Occurrence abgelegt werden. Wie

[1] Mittlerweile aktualisiert als ISO 13250:2003 Topic Maps [3]

Topics auch, lassen sich Occurrences typisieren, so dass Kategorien von Vorkommen besser auseinander halten lassen.

Associations: Association sind Bindeglieder zwischen den Topics und bilden die zusammenhänge der modellierten Domäne ab. Associations können, ebenso wie Topics, beliebig viele Namen erhalten und typisiert werden. Sie verbinden 2 oder mehr Topics miteinander, wobei jedem teilnehmenden Topic eine Rolle zugewiesen wird. Eine Rolle ist eine weitere Form der Typisierung und erlaubt es, genau festzuhalten, in welcher Funktion ein Topic an einer Association teilnimmt, was unter anderem bei der Darstellung als Graph wichtig ist.

Hierzu ein Beispiel: Angenommen, man wollte die Aussage „Beethoven komponierte die 9. Sinfonie" modellieren. Dazu müssen zwei Topics definiert werden, nämlich **Beethoven** und die **9. Sinfonie**. Für die Association *hat komponiert* werden außerdem zwei Rollen benötigt, nämlich die der Person, und die des Musikstücks. In der Topic Map wird die Aussage dann so aussehen: „**Beethoven** <Person> *hat komponiert* **9. Sinfonie** <Musikstück>".

Mit diesen drei Konzepten können bereits sehr komplexe Sachverhalte abgebildet werden. Zusätzlich gibt es noch drei weitere Instrumente, die Topic Maps noch mehr Möglichkeiten geben:

Scope: Scope lässt sich mit Gültigkeitsbereich übersetzen. In der Welt der Topic Maps dienen diese Gültigkeitsbereiche für zwei sehr wichtige Aufgaben: So können sie zur Gruppierung gleichartiger Daten eingesetzt werden, wodurch die Übersichtlichkeit von großen Topic Maps gesteigert wird. Wichtiger jedoch ist ihr Nutzen bei der Auflösung von Homonymen. Das ist eine Fähigkeit, die praktisch allen aktuellen Suchmaschinen fehlt und deren Fehlen maßgeblich für die große Anzahl irrelevanter Suchergebnisse verantwortlich ist. In einer Topic Map lassen sich jedoch mit Hilfe der Scopes beliebig viele verschiedene Geltungsbereiche für Topics und Associations definieren, so dass genau unterschieden werden kann, wann in einer Topic Map das Topic *Oper* in einem lokalen und wann in einem musikalischen Kontext verwendet wird.

Published Subject Indicator (PSI): Ein PSI ist ein Verweis auf ein andernorts definiertes Thema, der einem Topic in einer Topic Map angehängt werden kann. Damit ist für Betrachter – und wichtiger, für Computer – klar gestellt, dass an der angegebenen Stelle weitere Informationen zu diesem Thema zu finden sind. Benutzt werden sie, um für Topics aus Topic Maps verschiedener Autoren feststellen zu können, ob sie das gleiche Thema behandeln, also zum Beispiel den Komponisten Ludwig van Beethoven, oder einen Namensvetter, der vielleicht etwas ganz anderes gemacht hat. Normdateien, wie sie für Personen von Bibliotheken entwickelt werden, bieten sich als Quelle solcher PSI ebenso an, wie Publikationen von Normungseinrichtungen wie DIN oder ISO.

Merge von Topic Maps: Als Merge bezeichnet man die Vereinigung verschiedener Topic Maps in eine neue, in der die Daten der anderen wieder zu finden sind. Dabei sind gleiche Topics zu identifizieren, da diese zu einem Topic in der resultierenden Map verschmolzen werden. Für diese Verschmelzung sind zwei Regeln definiert. Topics werden zusammengeführt, wenn sie mindestens einen Namen in einem Scope teilen, bzw. wenn sie den gleichen PSI enthalten. Die zweite Regel erlaubt die sichere Verschmelzung zweier Topics, bei der ersten kann es leider zu falschen Ergebnissen kommen.

Diese sechs Instrumente stehen für die Erzeugung von Topic Maps zur Verfügung. Die ISO hat zwei Beschreibungssprachen für Topic Maps standardisiert, HyTM und XTM. HyTM basiert auf SGML und ist damit mittlerweile eher von historischem Interesse. Das auf XML basierende XTM [4] hat sich durchgesetzt, nicht zuletzt wegen der Vielzahl von Software, die XML-Sprachen verarbeiten und für den Einsatz im Internet aufbereiten kann.

3. Anwendungen

Topic Maps enthalten die semantischen Verflechtungen, die innerhalb einer bestimmten Domäne bestehen. Daher eignen sie sich sehr für alle Anwendungen, in denen es darum geht, Benutzern Einblicke in eine Domäne zu verschaffen, die sie normalerweise nicht so einfach erlangen könnten. Für den Einsatz im Internet ist XTM besonders gut geeignet, denn XML-Beschreibungen lassen sich relativ einfach in ein Format bringen, das für Websites genutzt werden kann. So ist denn der Einsatz in Portalen und digitalen Archiven als Such- und Navigationshilfe als bester Anwendungsfall zu nennen. Gestützt durch eine geeignete Visualisierung können Benutzer sich schnell einen Überblick über einen Datenbestand verschaffen. Die verschiedenen Typisierungen und Scopes erlauben, im Gegensatz zu Suchformularen, einen thematischen Einstieg und die Verknüpfungen der einzelnen Inhalte erlauben eine Einordnung eines Stücks in seinen Kontext.

Aber auch außerhalb digitaler Archive sind Anwendungen für Topic Maps denkbar. Im Bildungssektor können Topic Maps z.B. zur Beschreibung eines Themengebiets genutzt werden. Die Lernenden würden damit in die Lage versetzt, sich ein Themengebiet explorativ zu erschließen; neben den Fakten ließen sich Verbindungen und Beweggründe mittels Associations ausdrücken und weiterführendes Material in den Occurrences ablegen. Statt einer Menge von Fakten ließe sich also Wissen vermitteln, was gerade in Umgebungen ohne Lehrperson, also beim e-Learning oder Fernunterricht sehr nützlich ist.

Unternehmen können diesen Aspekt der Wissensvermittlung aufgreifen und mit Hilfe von Topic Maps ein Wissenssystem aufbauen, in dem Abläufe und Zusammenhänge festgehalten werden können. Damit lässt sich verhindern, dass Wissen im Unternehmen verloren geht, wenn zum Beispiel jemand in Ruhestand geht.

4. Forschungsthemen

Im Zusammenhang mit Topic Maps gibt es noch eine Reihe offener Forschungsfragen. So ist der Prozess der Erstellung einer Topic Map nach heutigem Stand der Technik eine rein manuelle Arbeit. Für kleine und überschaubare Topic Maps stellt das noch kein Problem dar, wohl aber für die retrospektive Aufarbeitung eines großen Archivs. Für solche Datenbestände müssen Methoden entwickelt werden, die das Erzeugen von Topic Maps für große Datensammlungen mit vertretbarem Aufwand ermöglichen. Eine vollständige Automatisierung des Prozesses ist jedoch nicht zu erwarten.

Ein weiteres Forschungsfeld ist die Visualisierung von Topic Maps. Hier spielt vor allem die Benutzerfreundlichkeit eine große Rolle. Die Schnittstelle sollte so konzipiert sein, dass sowohl Experten als auch Laien Nutzen aus den dargestellten Daten ziehen können. Hier könnten unterschiedliche Detaillierungsgrade der Karte genutzt werden, um Laien einen groben Überblick zu gewähren, während Experten wesentlich mehr Details zuschalten könnten. Bei der Visualisierung spielen aber auch Fragen der Skalierbarkeit der Anzeige eine Rolle. Große Datensammlungen können so viele Topics und Associations enthalten, dass die Masse an Informationen auf dem Bildschirm so groß wird, dass das zu enthaltene Wissen dahinter verloren geht. Hier ist dann geschicktes Filtern von Informationen gefragt.

5. Ausblick

Die semantische Anreicherung von Webinhalten auf standardisierte, computer-lesbare Art und Weise wird in den nächsten Jahren immer mehr an Bedeutung gewinnen. Das W3C setzt sich intensiv für die Verwirklichung des Semantic Web ein und wird dieses Ziel gewiss auch erreichen. Offen bleibt im Moment, wann die dazu benötigten Sprachen und Protokolle entwickelt und standardisiert sein werden. Digitale Archive von Bibliotheken oder Museen können aber schon heute von einer bereits standardisierten Technik, den Topic Maps, profitieren. Der ISO-Standard

gibt ihnen die Möglichkeit, bereits heute ihre Erfahrungen mit semantischen Netzen für ihre Datenbestände zu sammeln. Falls eine Einrichtung dann in einigen Jahren die erweiterten Möglichkeiten des Semantic Web nutzen wollte, könnten die gesammelten Daten automatisch übertragen werden; Übersetzungsprogramme in vom W3C verwendete Sprachen gibt es bereits heute.

Noch gibt es offene Forschungsfragen auf dem Gebiet der semantischen Netze, diese werden jedoch aktiv angegangen. Das Fraunhofer IMK in der Forschung zu Topic Maps und dem Semantic Web aktiv, Diplomarbeiten und Dissertationen sind dazu im Gange.

6. Referenzen

[1] Peter Lyman, Hal R. Varian et al. "How much Info?", Project Report of the University of California at Berkeley, School of Information Management and Systems, 2000. Zu finden unter http://www.sims.berkeley.edu/research/projects/how-much-info/internet.html

[2] Internet Software Consortium, „Internet Domain Survey Jan 2003". Zu finden unter http://www.isc.org/ds/WWW-200301/index.html

[3] International Standards Organisation ISO, „ISO 13250:2003 Topic Maps". Zu beziehen bei www.iso.org in der aktuellen Fassung.

[4] TopicMaps.org, „XML Topic Maps (XTM) 1.0". Zu finden unter http://www.topicmaps.org/xtm/index.html