

Semantische Metadaten für den Webauftritt einer Bibliothek

Semantic metadata for a library website

Abstract

The semantic web has been developing for the last 10 years and with its Resource Description Framework (RDF) and vocabularies it seems to be usable today. Unfortunately RDF is a quite complex technique that comes with loads of standards and high IT know-how and is at least not designed for simple use in libraries or for normal web site operators. In contrast to this large-scale and complex solution the great and leading companies for search engine sites like Google, Microsoft, Yahoo und Yandex formed an alliance and defined 'structured data' to describe semantic information for web pages. Basing on the structured data definition 'Library' (<https://schema.org/Library>), the Library of the Medical Faculty Mannheim defines some of its basic data as machine-readable structured data on its web site. Furthermore, the website's starting page contains semantic metadata in Open Graph format (that Facebook crawls and analyses) and Dublin Core format.

Keywords: web technology, semantic web, metadata, schema.org, library

Zusammenfassung

Das Semantic Web ist schon seit über 10 Jahren viel beachtet und hat mit der Verfügbarkeit von Resource Description Framework (RDF) und den entsprechenden Ontologien einen großen Sprung in die Praxis gemacht. Vertreter kleiner Bibliotheken und Bibliothekare mit geringer Technik-Affinität stehen aber im Alltag vor großen Hürden, z.B. bei der Frage, wie man diese Technik konkret in den eigenen Webauftritt einbinden kann: man kommt sich vor wie Don Quijote, der versucht die Windmühlen zu bezwingen. RDF mit seinen Ontologien ist fast unverständlich komplex für Nicht-Informatiker und somit für den praktischen Einsatz auf Bibliotheksseiten in der Breite nicht direkt zu gebrauchen. Mit Schema.org wurde ursprünglich von den drei größten Suchmaschinen der Welt Google, Bing und Yahoo eine einfach und effektive semantische Beschreibung von Entitäten entwickelt. Aktuell wird Schema.org durch Google, Microsoft, Yahoo und Yandex weiter gesponsert und von vielen weiteren Suchmaschinen verstanden. Vor diesem Hintergrund hat die Bibliothek der Medizinischen Fakultät Mannheim auf ihrer Homepage (<http://www.umm.uni-heidelberg.de/bibl/>) verschiedene maschinenlesbare semantische Metadaten eingebettet. Sehr interessant und zukunftsweisend ist die neueste Entwicklung von Schema.org, bei der man eine 'Library' (<https://schema.org/Library>) mit Öffnungszeiten und vielem mehr modellieren kann. Ferner haben wir noch semantische Metadaten im Open Graph- und Dublin Core-Format eingebettet, um alte Standards und Facebook-konforme Informationen maschinenlesbar zur Verfügung zu stellen.

Schlüsselwörter: Web-Technologie, Semantic Web, Metadaten, Schema.org, Bibliothek

Andreas Bohne-Lang¹

¹ Medizinische Fakultät
Mannheim, Universität
Heidelberg, Mannheim,
Deutschland

Einleitung

Das World Wide Web wird auch als WWW, W3 oder Web abgekürzt. Häufig wird auch der Begriff ‚Internet‘ als Synonym verwendet, was nicht korrekt ist, da das WWW nur ein Teil dessen ist, was das Internet ausmacht. Eine der großen Herausforderungen des World Wide Web der Gegenwart ist es, die bereitgestellten Informationen so aufzubereiten, dass diese präzise bei einer Websuche gefunden werden können.

Informationen im WWW werden auf Webseiten (Web Pages) zur Verfügung gestellt, die in der Regel im Dateiformat html erstellt sind. Die Gesamtheit der Webseiten eines Webauftritts (z.B. einer Bibliothek) wird als Website bezeichnet, da sie zusammen an einem Speicherort unter einer logischen URL (uniform resource locator, symbolische Web-Adresse) bzw. einer physischen Adresse (IP-Nummer) abgelegt sind. Die ähnlich klingenden Begriffe „Website“ und „Webseite“ bezeichnen also verschiedene Sachverhalte und dürfen nicht verwechselt oder gar synonym benutzt werden.

Suchmaschinen und deren Betreiber spielen eine federführende Rolle bei der Suche von Information anhand von Suchvorgaben im Web; denn sie bestimmen die auf eine Anfrage gefundenen Suchtreffer und somit die Sichtbarkeit potentieller Funde im Web. Derzeit ist der Suchraum von Suchmaschinen eingeschränkt, da diese nur die Webseiten-Inhalte indexieren und in der Regel keine Information über deren Bedeutung haben. Komplexe Kontextinformationen über den Inhalt einer Webseite werden von den Webseiten-Erstellern meist nicht zur Verfügung gestellt. Suchanfragen von Benutzern basieren bei den großen Suchmaschinen für gewöhnlich auf einzelnen Stichworten, welche mit einem logischen ‚und‘ verknüpft werden. Anfragen komplexerer semantischer Natur wie ‚Welche Bibliothek im Umkreis hat heute bis 20 Uhr geöffnet und hat das Buch „Anatomische Prüfungsfragen“‘ sind derzeit nicht möglich. Einer der Gründe hierfür ist die fehlende semantische Kodierung der Inhalte der Webseite in Form von Metadaten innerhalb ihrer html-Datei, so dass diese maschinenlesbar und bedeutungstragend hinterlegt wären. Dieser Artikel soll kurz beschreiben, wie (Medizin-)Bibliotheken, Arztpraxen, Krankenhäuser und andere Webseitenbetreiber hier selber aktiv eingreifen und somit semantische Inhalte auf ihrem Webauftritt verankern können, um zukünftig komplexe Suchanfragen zu fördern und zu ermöglichen.

Metadaten in Bibliotheken

Bei der Definition von Metadaten in Bibliotheken sei hier die Deutsche Nationalbibliothek zitiert, welche auf ihrer Website folgendes schreibt: „Metadaten sind (strukturierte) Daten, die eine Ressource, eine Entität, ein Objekt oder andere Daten beschreiben. Sie können darüber hinaus dem Auffinden, der Verwendung sowie der Verwaltung einer Ressource, einer Entität etc. dienen. Der Begriff Metadaten wird in unterschiedlichen Kontexten verwen-

det. Im informationswissenschaftlichen und bibliothekarischen Kontext versteht man hierunter Daten, die der Beschreibung von elektronischen Ressourcen dienen. Der Trend geht jedoch dahin, den Begriff Metadaten auch für Daten und Kataloge in Datenbanken zu verwenden, die nicht-elektronische Ressourcen beschreiben.“ [1]. In Bibliotheken ist die Erfassung von Metadaten von Medien schon seit jeher eines der Hauptaufgabengebiete des Erwerbungsprozesses. Dabei unterscheidet man in der Erschließung zwei Teilbereiche: Die Formalerschließung und die Sacherschließung. Die Formalerschließung erfasst die bibliographische Beschreibung des Werkes wie Name des Autors, Erscheinungsjahr, Erscheinungsort und Verlag, aber auch Seitenanzahl, Größe und Kaufpreis. Dabei hält sich die Formalerschließung eng an bibliothekarische Regelwerke, um einen Austausch der Daten mit anderen Einrichtungen zu ermöglichen. Die aktuellen Regelwerke sind Regeln für die alphabetische Katalogisierung (RAK) und Resource Description and Access (RDA). Im Gegensatz zu der Formalerschließung steht die Sacherschließung oder auch Inhalterschließung. Hierzu wird der Inhalt des Mediums per Autopsie von einer geschulten Person wie z.B. einem Fachreferent erfasst und klassifiziert. Bei der Klassifikation wird das Medium bestimmten semantischen Kategorien eines meist hierarchischen Ordnungssystems zugeordnet. Als bekannteste und gebräuchlichste sind die Dewey Decimal Classification (DDC), die Regensburger Verbundklassifikation (RVK) und die Library of Congress Classification (LCC) zu nennen. Aber auch Klassifikationen von Verlagen und dem Buchhandel wie ONline Information eXchange (ONIX) oder Business Industry Classification (BIC) werden verwendet. Allen Metadaten gemein ist, dass sie das Auffinden von Medien bzw. Medieninhalten erleichtern und ermöglichen sollen. Betrachtet man die Aufstellungssystematik ‚numerus currens‘ (fortlaufende Nummer) in Bibliotheken, bei der die Medien in der zeitlichen Abfolge der Erwerbung in die Regale gestellt werden (was konkret einer zufälligen Aufstellung entspricht), so bedarf das Auffinden einzelner Medien umfangreiche Kenntnisse über das Werk. Dabei stammen die Ansätze und Lösungen zum Auffinden aus Zeiten ohne Computer und Suchmaschinen und konnten nur mit Mitteln der lokalen Bibliothek umgesetzt werden.

Semiotisches Dreieck

Das Semiotische Dreieck (siehe Abbildung 1) beschreibt das Problem, welches die Verarbeitung von Informationen mit sich bringt. Dieses lässt sich in drei Teilbereiche aufteilen:

- Zum einen gibt es in unserer Vorstellung die (abstrakte) Idee eines Objektes. Wenn wir uns einen Baum vorstellen, so assoziieren wir Gedanken wie: „oben Zweige, unten Wurzeln, betreibt Photosynthese, benötigt Wasser, ist im Sommer grün, der Stamm ist aus Holz“, etc. Dass alles zusammen bildet die Idee eines Baumes.

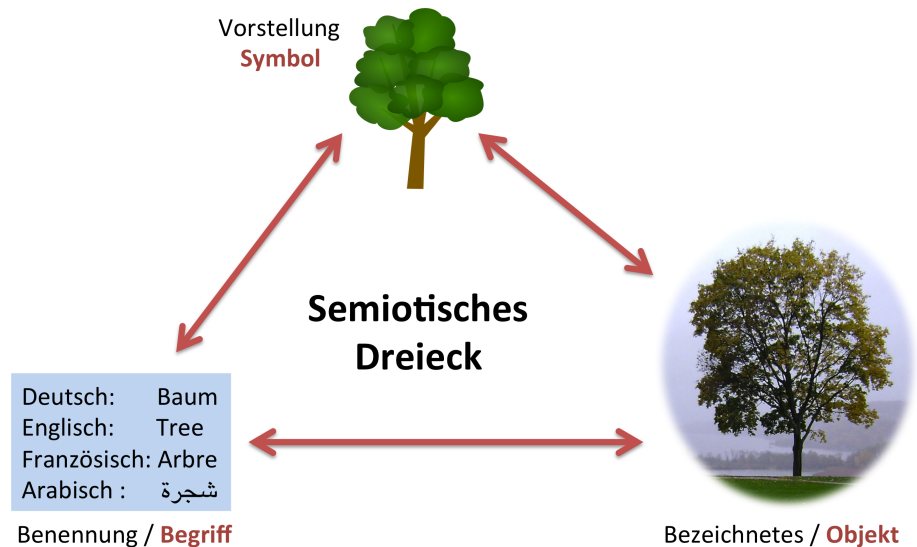


Abbildung 1: Semiotisches Dreieck

- Dann gibt es die konkrete Ausprägung (das Objekt) dessen, worauf sich unsere Vorstellung beruft, also einen echten Baum in der Landschaft wie eine Eiche, Fichte oder einen Apfelbaum.
- Der dritte Teil des Semiotischen Dreiecks ist die Benennung. Durch die Sprachvielfalt hat jede Sprache ihren eigenen Ausdruck für ‚Baum‘. Die Engländer bezeichnen es als ‚tree‘, die Franzosen als ‚arbre‘ und im Arabischen ist es ‚شجرة‘.

Bei der maschinellen Verarbeitung von Daten muss man die drei Teilbereiche berücksichtigen. Neben der eigentlich Information muss man Daten über die Information bereit halten. So ist z.B. die Zahl 45678 zunächst nur eine Zahl ohne Bedeutung. Erst mit der Zuordnung einer Verwendung, wie „Telefonnummer=45678“, wird der Kontext festgelegt, in dem diese Zahl Anwendung findet. Die gleiche Zahl kann in einem anderen Kontext eine andere Bedeutung haben. Daher bedarf es bei der Verarbeitung von Informationen semantischer Angaben darüber, in welchem Kontext die Daten anzuwenden sind. Wichtig ist in diesem Zusammenhang, dass man ein gemeinsames Wörterbuch mit Schlüsseln definiert und diese eindeutig festlegt. Die Verwendung verschiedener Schlüssel (Systeme) erschwert oder verhindert die maschinelle Verarbeitung, da man auf Konkordanzen angewiesen wäre. Für die maschinelle Verarbeitung muss das Wissen ferner strukturiert abgebildet werden. Hierfür gibt es verschiedene Möglichkeiten mit unterschiedlicher Mächtigkeit. Neben dem reinen Festlegen von Begriffen gibt es die Möglichkeiten, Über- und Unterordnungen, Wenn-Dann-Beziehungen, Begriffs-Räume, Synonyme und Ähnlichkeiten etc. zu definieren.

Im Bereich der Wissensrepräsentation kann man die semantische Reichhaltigkeit nach Pellegrini und Blumauer [2] in ihrer Mächtigkeit bewerten:

Semantische Reichhaltigkeit

1. *Glossar*: Ist eine einfache, meist alphabetisch sortierte Liste von Wörtern und ihren Erklärungen.
2. *Folksonomy*: Ist ein Kofferwort, das sich aus Folk und Taxonomie zusammensetzt. Es beschreibt von Benutzern (folks) vergebene Schlagworte (Tags) zu einem Begriff und entspringt dem Web 2.0-Umfeld.
3. *Taxonomie*: Die Taxonomie ist ein hierarchisches System, das mit dem Bildungsprinzip Über-/Unterordnung Elemente strukturiert. Taxonomien spielen z.B. in der Biologie mit der Unterteilung der Lebewesen in Art, Gattung oder Familie eine bedeutende Rolle.
4. *Thesaurus*: Der Thesaurus erweitert die Taxonomie um die zwei fest definierten Relationen der Objekte untereinander: die Ähnlichkeits- und die Synonym-Relation.
5. *Topic Map*: Sie besteht aus abstrakten Dingen, Assoziationen, Gültigkeitsbereichen für abstrakte Dinge und zugeordneten Dokumenten außerhalb der Topic Map. Es lassen sich Assoziationen zwischen den Objekten selbst definieren.
6. *Ontologie*: Die Königsdisziplin – sie besteht aus Begriffen, Typen, Instanzen, Relationen, Vererbung und Axiomen, welche Zusammenhänge zwischen den einzelnen Objekten der Ontologie als auch mit anderen Ontologien mittels „wenn-dann“-Beziehungen, Zuweisungen, logischen Verknüpfungen und weiteren Funktionen ausdrücken kann.

Austausch von Informationen im Computerzeitalter

Nachdem ab den 80er Jahren immer mehr Rechner von großen Einrichtungen über fest definierte Adressen, Schnittstellen und Protokolle miteinander vernetzt wurden, entstand das „Internet“. Das World Wide Web im engeren Sinne entstand 1989 als Projekt an der For-

Tabelle 1: Diese 15 Elemente wurden 1998 offiziell als Dublin Core Metadata Element Set, Version 1.0 [30] veröffentlicht

1. Title	9. Format
2. Author or Creator	10. Resource Identifier
3. Subject and Keywords	11. Source
4. Description	12. Language
5. Publisher	13. Relation
6. Other Contributor	14. Coverage
7. Date	15. Rights Management
8. Resource Type	

Tabelle 2: Implementierung von Metadaten in HTML 5. Die genaue Formatierung der Dublin Core Metadata hat sich im Laufe der veröffentlichten HTML-Versionen verändert.

```
<head profile="http://dublincore.org/documents/dcq-html/">
<title>Dublin Core</title>
<link rel="schema.DC" href="http://purl.org/dc/elements/1.1/" />
<link rel="schema.DCTERMS" href="http://purl.org/dc/terms/" />
<meta name="DC.format" scheme="DCTERMS.IMT" content="text/html" />
<meta name="DC.type" scheme="DCTERMS.DCMItype" content="Text" />
<meta name="DC.subject" content="Dublin Core Metadaten-Elemente, Beispiel" />
<meta name="DC.creator" content="Andreas Bohne-Lang" />
<meta name="DCTERMS.license" scheme="DCTERMS.URI"
content="http://www.gnu.org/copyleft/fdl.html" />
<meta name="DCTERMS.rightsHolder" content="Universität Heidelberg" />
<meta name="DCTERMS.modified" scheme="DCTERMS.W3CDTF" content="2016-11-11" /> </head>
```

schungseinrichtung CERN und wurde von Tim Berners-Lee entwickelt. Dazu programmierte er einen Server und einen Client, der in der Lage war, Hypertexte, also Texte mit Verweisungen auf andere Adressen, zu verarbeiten. Am 6. August 1991 veröffentlichte er das ‚World Wide Web-Projekt‘ in einem Beitrag zur Newsgroup alt.hypertext und lud dazu ein, diese Software zu benutzen [3].

Das zugrunde liegende Hypertext Transfer Protocol (HTTP) war als offener Standard konzipiert und hat bis heute Bestand. Seit dieser Zeit spielt Tim Berners-Lee eine maßgebliche Rolle bei der Weiterentwicklung des World Wide Web, da er 2004 das World Wide Web Consortium (W3C) gründete, diesem als Direktor vorsteht und Standards, Spezifikationen, Richtlinien, Software und Hilfsprogramme (Tools) (mit-)entwickelt.

Metadaten in Webseiten

Für die Kennzeichnung von Metadaten innerhalb von Webseiten griff Berners-Lee auf das Konzept der Auszeichnungssprachen (Markup Languages) zurück, das schon in Form der Standard Generalized Markup Language (SGML) bekannt war. Bereits im ersten Entwurf von 1993 „Hypertext Markup Language (HTML) – A Representation of Textual Information and Meta Information for Retrieval and Interchange“ [4] sah Tim Berners-Lee vor, dass Webseiten einen Kopfbereich <head> haben, in dem Metadaten über die Webseite untergebracht werden sollen, und einen Inhaltsbereich <body>. Der damalige Entwurf definierte im Kopfbereich nur die Tags <title>, <isindex>, <nextid> und <link>. In der zweiten Version von HTML, die im November 1995 unter Request for Comments ‚rfc1866‘ [5] erschien, war der Metadatenbereich erweitert worden. Diese Version von HTML fügte

dem HEAD-Bereich ein weiteres Element hinzu und sah hier nun unter Punkt 5.2.5 den Bereich ‚Associated Meta-information‘ vor. Dieses neu eingeführte Meta-Element sah vor, dass über dessen Verwendung sowohl der Zugriff auf das Dokument geregelt werden sollte als auch, dass es Information über den Inhalt, die Eigenschaft und Dienlichkeit des Inhalts bereit hielt. Diese zweite Version von HTML stellte jedoch für das META-Tag nur ein Attribut das HTTP-EQUIV zur Verfügung. Dieses Attribut war so definiert, dass es universell eingesetzt werden konnte. So war es hiermit möglich, die Gültigkeit des Webseiten-Inhaltes zu definieren – Beispiel <META HTTP-EQUIV="Expires" CONTENT="Tue, 04 Dec 1993 21:29:02 GMT">. Aber auch einfach Schlagwörter wie <Meta Http-equiv="Keywords" CONTENT="Beany"> konnten so dem Server und damit auch dem Web-Browser übermittelt werden.

Dublin Core

Die 1995 in der Planung [6] vorliegende Version 3.0 von HTML wurde 1996 als W3C-Empfehlung herausgegeben und im gleichen Jahr wurden auch Überlegungen zu der Einbettung von Metadaten in HTML angestellt [7] (siehe Tabelle 1). Im Jahr 1997 wurde die HTML-Version 3 durch die Version 4.0 ersetzt, welche 1999 die Einbindung von Dublin-Core-Elementen in die technischen Spezifikationen der Requests For Comments (RFC) mit ‚Encoding Dublin Core Metadata in HTML‘ festlegte [7], [8]. Ab 2010 war die Einbindung von Metadaten keine RFC-Empfehlung mehr, sondern wanderte in die Obhut der Dublin Core Metadata Initiative DCMI (<http://dublincore.org/>). In Tabelle 2 wird gezeigt, wie aktuelle Dublin-Core-Metadaten

in Webseiten mittels HTML in der Version 5 zu integrieren sind.

Suchanfragen

Betrachtet man, wie die heutigen Benutzer im Web suchen, so wird man feststellen, dass die Suchanfragen meist aus einem oder mehreren Stichwörtern bestehen, welche die Benutzer in den Suchschlitz der Suchmaschinen eingeben. Dabei wird in der Regel intuitiv von einer ‚und‘-Verknüpfung ausgegangen, obschon z.B. die Suchmaschine von Google weitaus mehr Suchoperatoren unterstützt [9]. Durch die Schnittmengenbildung via ‚und‘ versuchen Benutzer, die Trefferseiten in Bezug auf die Fragestellung einzugrenzen. Unabhängig von der Schwierigkeit, Suchanfragen in natürlicher Sprache zu analysieren und zu beantworten, ist es derzeit nicht möglich, Suchanfragen einen einschränkenden Suchraum mitzugeben oder unkonkrete örtliche (location-based) Angaben wie ‚in meiner Umgebung‘ zu machen, welche in einem ersten Schritt die Position des Gerätes bestimmen und diese in einem zweiten Schritt auf die Trefferliste anwenden müssten. Von großen Nutzen wäre es, wenn man Suchanfragen mit komplexen Verknüpfungen und semantischen Gültigkeitsbereichen stellen könnte. Eine solche Suchanfrage könnte wie folgt aussehen: ‚Suche eine Bibliothek in meiner Umgebung, die heute bis 22 Uhr geöffnet hat und das Buch „Anatomische Prüfungsfragen“ hat‘. Um diese Anfrage bearbeiten zu können, muss man diese in die einzelnen bedeutungstragenden Teile zerlegen und diese für die Anfrage neu verknüpfen. Dabei ist es wichtig, die Bedeutung mit in die Suchanfrage einzubeziehen: ‚Ort=Bibliothek und Distanz von aktueller Position <10km und Öffnungszeiten <=22 Uhr und Wochentag=Freitag(heute) und Buch=Anatomische Prüfungsfragen ist im Bestand‘. Damit Suchmaschinen Anfragen dieser Art bearbeiten können, benötigen sie nicht nur die reine Webseite mit deren Inhalt, welche sie indexieren, sondern sie benötigen auch Informationen in einer maschinenlesbaren und festgelegten Art über die Bedeutung der Daten. Hierbei lassen sich zum Beispiel feste Elemente wie Adresse, Telefonnummer, Öffnungszeiten gut modellieren.

Das Semantic Web

Tim Berners-Lee hat nicht nur das World Wide Web, wie wir es heute kennen, mitbegründet, sondern die Entwicklung über die Jahre weiter voran getrieben. Mit der Zeit wurden die Server im World Wide Web immer zahlreicher, und somit auch die zur Verfügung stehenden Daten. Jedoch waren diese Datensammlungen meist isoliert und proprietär. 2001 publizierte Tim Berners-Lee Artikel in Nature [10] und Scientific American [11], in denen er seine Idee zu einer neuen Stufe des WWW skizzierte. Grundidee des ‚Semantic Web‘ (inzwischen auch als Web 3.0 bezeichnet) ist zum einen eine starke Formalisierung und Standardisierung aller beteiligten und eingesetzten

Komponenten. Dazu schreibt Tim Berners-Lee „The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings“ [11]. Ferner fordert er in diesem Kontext, dass für das Semantic Web die ‚Datensilos‘ geöffnet werden müssen, und er forderte, dass bei dem Semantic Web die Inhalte miteinander verknüpft werden müssen. Dies wird unter Linked Open Data (LOD) verstanden. Damit eine serverübergreifende Suchanfrage (auch als föderierte Suche bezeichnet) der Daten möglich wird, müssen alle verwendeten Komponenten klar definiert werden. Als Güteklassifikation für Daten im Internet führte Tim Berners-Lee 2006 ein Fünf-Sterne-System [12] ein, in dem er die Güte der zur Verfügung gestellten Daten klassifiziert:

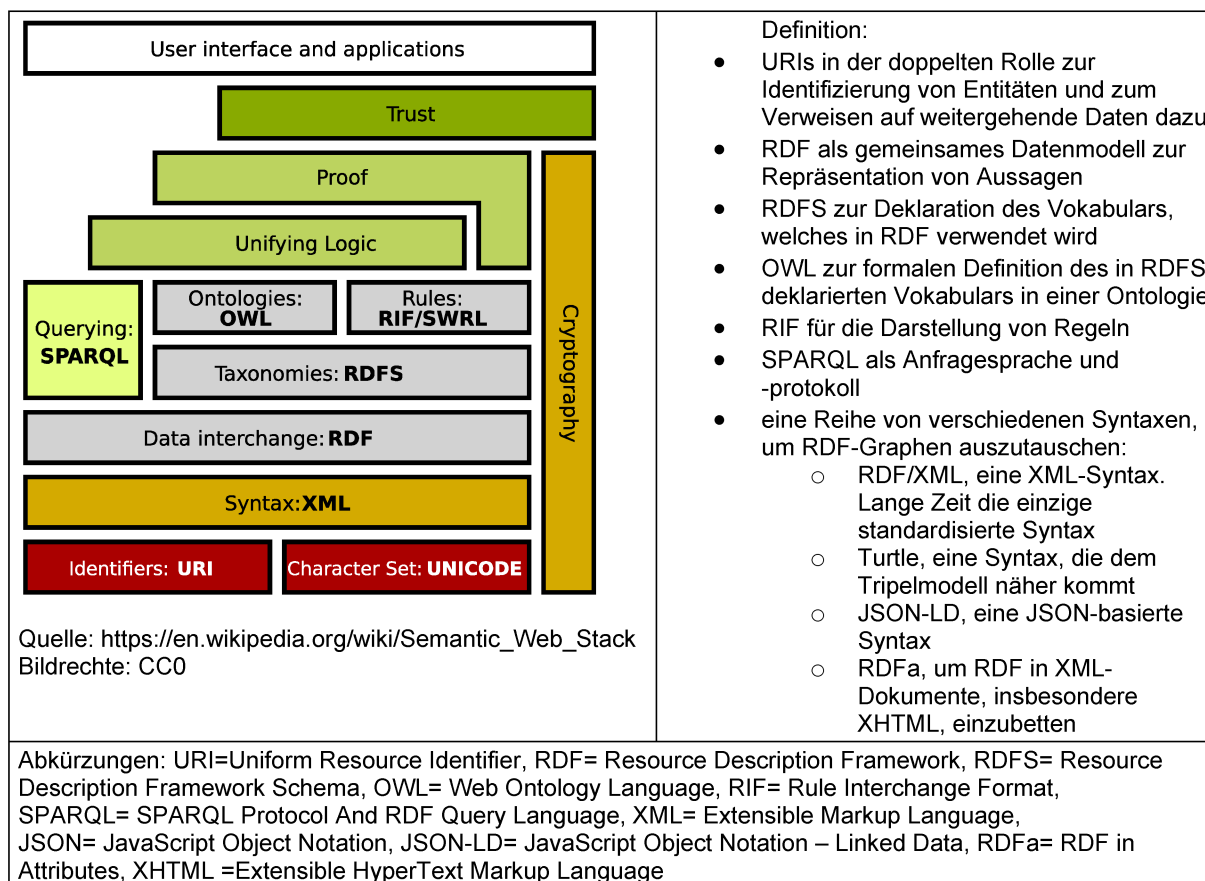
- 1 Stern: Die Daten sind frei im Netz verfügbar.
- 2 Sterne: Die Daten sind strukturiert und maschinenlesbar (also keine eingescannten Bilder oder Tabellen, sondern z.B. HTML-Tabellen).
- 3 Sterne: Die Daten liegen in einem offenen Format vor (also CSV-Dateien und keine Excel-Dateien).
- 4 Sterne: Alle drei ersten Punkte und die Daten liegen im offenen Standard RDF und SPARQL vom W3C vor, um Objekte zu identifizieren.
- 5 Sterne: Alle vier ersten Punkte und eine Verlinkung der Objekte zueinander.

Die Grundeinheit der Wissensrepräsentation im Semantic Web ist das Triple: eine Informationseinheit gibt den Verweis auf die Beziehung zu einer anderen Informationseinheit an.

Die Definition eines derartigen Wissens-Triple im Semantic Web ist minimal einfach: Subjekt, Prädikat, Objekt. Dabei kann ein Objekt wieder das Subjekt eines neuen Triple sein, wodurch sich ein Wissens-Netz aufspannt. Die Komplexität ergibt sich aus der Definition und Verwendung der eingesetzten Ontologien, in denen die drei Bestandteile des Triple definiert werden.

Die Elemente aus Tabelle 3 stellen beim Semantic Web einen Kernbereich dar. Für den praktischen Einsatz sind weitere fundierte technische Kenntnisse notwendig, wie z.B. für das Betreiben eines Triple-Store, einer Datenbank welche die Daten-Tripel speichert oder beim Erstellen oder Erweitern einer Ontologie. Einer der wichtigsten Punkte, die zum Erfolg oder auch zum Scheitern des Semantic Web beitragen, ist die Definition der Objekte und deren Beziehungen – die Ontologie (im Englischen auch als vocabulary bezeichnet). Die Initiative Linked Open Vocabularies [13] verzeichnet derzeit 577 Ontologien (Stand: 20.10.2016). Da es jedem freigestellt ist, eine Ontologie für seinen Themenbereich zu entwickeln und zu veröffentlichen, existiert ein gewisser Wildwuchs. Nicht selten kommt es daher vor, dass neu entwickelte Ontologien das Projektende, in dessen Rahmen sie erstellt wurden, nicht lange überleben. Im Bereich Bibliothek existieren derzeit einige Ontologien, wie zum Beispiel: Bibliographic Ontology Specification [14], FRBR-aligned Bibliographic Ontology [15], The Service Ontology [16], Holding Ontology [17] und Document Service Ontology

Tabelle 3: Der sogenannte Semantic Web Layer Cake (dt. „Schichtenkuchen“)



[18]. Die gesamte Komplexität des Projektes ‚Semantic Web‘ wird sehr gut in dem Buch ‚(Open) Linked Data in Bibliotheken‘ [19] wiedergegeben.

Aufwand und Nutzen

Obschon die Grundidee der Tripel sehr einfach ist, ist das für die Verarbeitung aufgespannte Rahmenwerk um so komplexer. Der Einstieg in die Welt von Linked Open Data und RDF setzt sowohl ein hohes Fachwissen der zu verwendenden Standards als auch profunde IT-Kenntnisse voraus. Dieser Aufwand steht in keinem Verhältnis zum Ergebnis, wenn eine Bibliothek oder eine andere Einrichtung lediglich ihre Eckdaten wie ihre Adresse, Erreichbarkeit oder Service-Bereiche maschinenlesbar und bedeutungstragend auf der Startseite ihrer Website anbieten möchte. Gerade die maschinell lesbare Kodierung der Öffnungszeiten ist allerdings für viele Bibliotheken durchaus von Interesse.

GoodRelations

Im Jahre 2008 entwickelte Martin Hepp von der E-Business and Web Science Research Group der Universität der Bundeswehr München ein Vokabular, welches die wichtigsten Attribute und Vorgänge von Geschäften und Geschäftsvorgängen beschreibt [20], [21]. Dieses Vokabular umfasst einen weiten Bereich, der von Flugpreisen (Airfare) bis zu Ferienhäusern (Vacation homes) und Vi-

deos reicht [21]. Dabei setzt GoodRelations auf RDF auf und ist somit in der Anwendung sehr komplex. Auf der Homepage von GoodRelations existieren einige online zugängliche Generatoren [22], um einfache Kontextdaten in syntaktisch korrekten Code umzusetzen.

Alternative Schema.org

Im Gegensatz zu GoodRelations (basierend auf RDF) bietet Schema.org eine einheitliche Ontologie für die Strukturierung von Daten auf Websites auf der Basis von bereits bestehenden Auszeichnungssprachen an und macht somit eine Einbindung in Webseiten sehr einfach. Ein Großteil der Klassen und Attribute von Schema.org wurde aus führenden Ontologien wie FOAF (FOAF ist die Abkürzung von ‚Friend of a Friend‘. Es ist ein Projekt zur maschinenlesbaren Modellierung von Personen und deren Beziehungen zu- und untereinander), GoodRelations und OpenCyc (Cyc stammt vom englischen *encyclopedia* und beschreibt eine maschinenlesbare Wissensdatenbank) übernommen. Verliert man auf der einen Seite die Flexibilität einer frei definierten Ontologie, so gewinnt man eine einfache Auszeichnungssprache, die jeder, der mit HTML umgeht, sich aneignen kann. Der Aufbau von Schema.org ist hierarchisch (siehe Abbildung 2). Der oberste Knoten unter schema ist das Objekt ‚Thing‘ mit den vier Attributen: name, description, url und image.

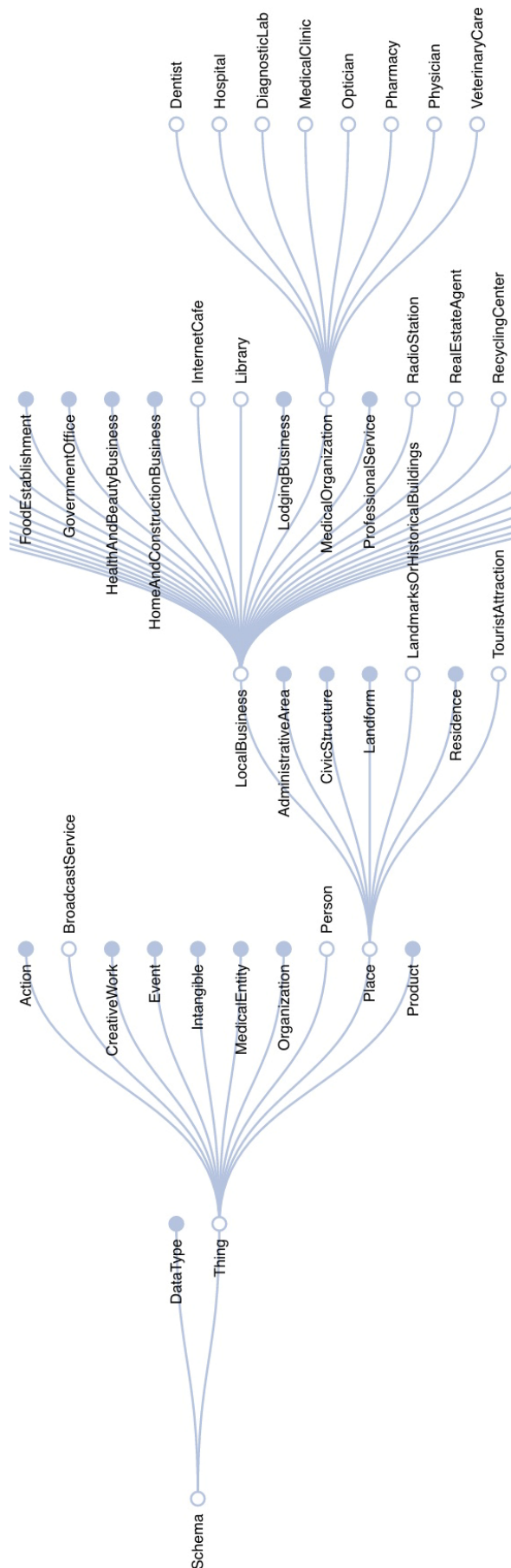


Abbildung 2: Ausschnitt des Schema.org-Pfades zur ‚Library‘-Definition und den medizinischen Einrichtungen

Darunter folgen weitere Objekte wie:

- CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
- Eingebettete nicht textuelle Objekte: AudioObject, ImageObject, VideoObject
- Event
- Organization
- Person
- Place, LocalBusiness, Restaurant ...
- Product, Offer, AggregateOffer
- Review, AggregateRating

Die gesamte Liste der Objekte steht auf der Web-Seite von Schema.org zur Verfügung [23].

Die Modellierung der Objekte, Attribute und Strukturen ist das Abbilden der realen Welt in das abstrakte Gerüst des Schema.org-Modells. Die Integration der so modellierten Daten in die Web-Seite ist auf drei Arten möglich (siehe Tabelle 4, Tabelle 5 und Tabelle 6). Hierbei ist eine Einbettung in den HTML-Code Seite mittels HTML Microdata und RDFa möglich, wobei Layout-Anweisungen und semantische Daten gemischt werden, wie Tabelle 4 und Tabelle 5 zeigen. Der Ansatz, die semantischen Daten im JSON-LD-Format in die Webseite zu integrieren (siehe Tabelle 6), trennt die semantischen Daten von dem Seiten-Körper (<body>) und platziert die Metadaten über die Web-Seite in dem eigentlich dafür vorgesehenen Kopf-Teil (<head>) der Webseite.

Ergebnis

Für die Startseite der Medizinischen Bibliothek Mannheim der Universität Heidelberg wurden die wichtigsten Eckdaten wie Anschrift, Telefonnummer, Öffnungszeiten etc. mit der von Schema.org zur Verfügung gestellten Library-Definition modelliert. Die Abbildung der Daten von der Webseite auf die strukturierten Daten von Schema.org ist in Abbildung 3 zu sehen. Der Erfolg und das praktische Ergebnis der Codierung semantischer Metadaten auf der Homepage lässt sich derzeit schlecht evaluieren, da sich die großen Suchmaschinenanbieter nicht direkt in die Karten schauen lassen. Bei der Analyse des Google-Egebnisses (siehe Abbildung 4) war auffällig, dass die Positionsangabe des Gebäudes auf dem Campus korrekt auf Google-Maps angegeben wurde. Als Positionsangabe wurden Geo-Daten in Form von Längen- und Breitengraden hinterlegt. Eine Bestimmung des Gebäudes anhand der Adresse ist auszuschließen, da der gesamte Campus unter einer Adresse firmiert. Bei der Modellierung der Daten hat sich gezeigt, dass es mitunter mehrere Möglichkeiten gibt, Inhalte anzugeben. Dieses resultiert aus der Polymorphie des Konzeptes von Schema.org, wo ein Objekt von mehreren Eltern-Knoten deren Attribute erben kann. Konkret ist an dieser Stelle die Modellierung der Öffnungszeiten anzuführen. So ist eine Bibliothek unter den Objekten ‚Local Business‘ als auch unter ‚Place‘ aufgehängt, welche beide unabhängig die Öffnungszeiten kodieren. Bei der Entwicklung und Validierung der Daten

Tabelle 4: HTML Microdata Format

```
<body>[...]
  <div itemscope itemtype="http://schema.org/Library">
    <h1 itemprop="name">MEDMA Library</h1>
    <p itemprop="description">A superb collection of fine books and services. </p>
    <p>Open: <span itemprop="openingHours" content="[Mo-Fr 08:00-00:00","Sa-Su 09:00-22:00]">
      Monday-Friday 8am-evening and on Weekend from 9am-evening </span></p>
    <p>Phone: <span itemprop="telephone" content="+6213833700">(0621)383-3700</span></p>
  </div>
</body>[...]
```

Tabelle 5: RDFa Format

```
<body>[...]
  <div vocab="http://schema.org/" typeof="Library">
    <h1 property="name"> MEDMA Library </h1>
    <p property="description"> A superb collection of fine books and services.</p>
    <p>Open: <span property="openingHours" content="[Mo-Fr 08:00-00:00", "Sa-Su 09:00-22:00]">
      Monday-Friday 8am-evening and on Weekend from 9am-evening </span></p>
    <p>Phone: <span property="telephone" content="+6213833700">(0621)383-3700</span></p>
  </div>
</body>[...]
```

Tabelle 6: LD-JSON Format

```
<head> [...]
<script type="application/ld+json">
{
  "@context": "http://schema.org",
  "@type": "Library",
  "name": "MEDMA Library",
  "description": "A superb collection of fine books and services.",
  "openingHours": [ "Mo-Fr 08:00-00:00", "Sa-Su 09:00-22:00" ],
  "telephone": "+6213833700"
}
</script>
</head>[...]
```

ist dem Autor ein Fehler in dem Google-Checker [24] für strukturierte Daten aufgefallen, welchen er an Google über ein dafür vorgesehenes Forum gemeldet hat. Eine Reaktion auf den Fehler hat lange gedauert und war nicht lösungsorientiert.

Das Modellieren von Metadaten für Facebook im Open Graph-Format im Kopfbereich der Web-Seite war für die Bibliothek der Medizinischen Fakultät Mannheim von großer Bedeutung, da Facebook bei der Auswahl eines Vorschaubildes nach eigenem Ermessen ein Bild von der Seite auswählt, wenn es keine Vorgaben gibt. Hier wurde vor dem Einbau semantischer Metadaten als Vorschaubild ein Buch-Cover aus der Kollektion ‚Neueste e-Books‘ ausgewählt, welches je nach Fachgebiet (z.B. Urologie, Gynäkologie, Suchtmedizin, etc.) des Buches nicht unbedingt die Bibliothek als solche repräsentierte. Mit den richtig kodierten Metadaten konnte das Vorschaubild auf die gewünschte Einstellung (ein Bild von der Außenansicht der Bibliothek) gebracht werden.

Diskussion und Ausblick

Der Einfluss von Metadaten beim Ranking der Suchergebnisse hat sich im Laufe der Jahre geändert. Die erste

Suchmaschine Altavista hat noch Daten aus dem Kopfbereich einer Webseite für das Positionieren des Treffers in der Trefferliste herangezogen. Im Kampf um die Platzierung der eigenen Web-Präsenz auf der ersten Trefferseite haben Webseitenbetreiber im Rahmen der Suchmaschinenoptimierung (Search Engine Optimization, SEO) vorsätzlich falsche Metadaten (wie z.B. Produktnamen der Konkurrenz) auf ihren Webseiten platziert. Als Konsequenz daraus entschieden sich um 2009 die großen Suchmaschinenanbieter, die Metadaten nicht mehr für das Ranking heranzuziehen. Dabei unterschieden sich die verschiedenen großen Anbieter ein wenig im Umfang [25], [26]. Bei Google werden die Daten jedoch ausgewertet [27].

Die Verwendung semantischer Metadaten zielt allerdings nicht auf die Platzierung im Ranking ab, sondern auf die Möglichkeit, bei Anfragen die Auswahl von Inhalten und Bedeutungen mit zu erfassen und somit semantisch komplexe Suchen zu ermöglichen. Mit der Modellierung von Daten nach den Schemata von Schema.org ist es auch Mitarbeitern einer Bibliothek (und anderen Einrichtungen wie Krankenhäusern, Arzt- und Zahnarztpraxen) mit geringer IT-Erfahrung möglich, Metadaten für ihren Webauftritt zu hinterlegen. Auch wenn man nicht die Freiheit in der Modellierung und verteilten Suchabfragen


```
<script type="application/ld+json"> {
  "@context": "http://schema.org",
  "@type": "Library",
  "openingHours": [ "Mo-Fr 08:00-00:00", "Sa-Su 09:00-22:00" ],
  "contactPoint": [{
    "@type": "ContactPoint",
    "contactType": "technical support",
    "telephone": "+49-621-383-3700",
    "availableLanguage": [ "German", "English" ],
    "hoursAvailable": [ "Mo-Fr 09:00-17:00" ]
  }, {
    "@type": "ContactPoint",
    "contactType": "customer support",
    "telephone": "+49-621-383-3700",
    "availableLanguage": [ "German", "English" ],
    "hoursAvailable": [ "Mo-Fr 08:00-20:00" ]
  } ]
  "address": {
    "@type": "PostalAddress",
    "addressCountry": "DE",
    "addressRegion": "BW",
    "addressLocality": "Mannheim",
    "postalCode": "68167",
    "streetAddress": "Theodor-Kutzer-Ufer 1-3 / Haus 42"
  },
  "faxNumber": "+49-621-383-2006",
  "telephone": "+49-621-383-3700",
  "geo": { "@type": "GeoCoordinates", "latitude": "49.491552",
    "longitude": "8.488676" },
  "hasMap": "http://intra4x.umm.de/index.php?id=373",
  "email": "bibliothek@medma.uni-heidelberg.de",
  "url": "http://www.umm.uni-heidelberg.de/bibl/"
}
```

ÖFFNUNGSZEITEN [mehr](#)

Mo - Fr 8 - 24 Uhr
Sa - So 9 - 22 Uhr

AUSKUNFT

Fachpersonal Auskunft
Mo - Fr 9 - 17 Uhr

Fachpersonal Ausleihe
Mo - Fr 8 - 20 Uhr
(jede 4. Woche 8 - 17 Uhr)

AKTUELLES

[Newsblog / Newsletter](#)

MITARBEITER

[AnsprechpartnerInnen](#)


ADRESSE

Bibliothek der Medizinischen Fakultät Mannheim
Universitätsmedizin Mannheim
Haus 42
Theodor-Kutzer-Ufer 1-3
68167 Mannheim

[Lieferadresse](#)

[für Navigationssysteme](#)

Telefon: +49 (0) 621/383 3700
Fax: +49 (0) 621/383 2006

 **Campus-Plan**

ÖFFNUNGSZEITEN [mehr](#)

Mo - Fr 8 - 24 Uhr
Sa - So 9 - 22 Uhr

AUSKUNFT

Fachpersonal Auskunft
Mo - Fr 9 - 17 Uhr

Fachpersonal Ausleihe
Mo - Fr 8 - 20 Uhr
(jede 4. Woche 8 - 17 Uhr)

AKTUELLES

[Newsblog / Newsletter](#)

MITARBEITER

[AnsprechpartnerInnen](#)

ADRESSE

Bibliothek der Medizinischen Fakultät Mannheim
Universitätsmedizin Mannheim
Haus 42
Theodor-Kutzer-Ufer 1-3
68167 Mannheim

[Lieferadresse](#)

[für Navigationssysteme](#)

Telefon: +49 (0) 621/383 3700
Fax: +49 (0) 621/383 2006


 **Campus-Plan**

Abbildung 3: Modellierung der Eckdaten der Bibliothek der Medizinischen Fakultät Mannheim. (Anmerkung: Das Attribut ‚hoursAvailable‘ wurde so modelliert, dass das Test-Tool für strukturierte Daten [24] dieses ohne Fehler akzeptiert, obschon Schema.org den Wertebereich hierfür anders definiert. Siehe hierzu auch den Abschnitt Ergebnis.)



Fotos anzeigen





Bibliothek der Medizinischen Fakultät Mannheim der Universität Heidelberg

Website
Routenplaner

Bibliothek

Befindet sich in: [Universität Heidelberg](#)

Adresse: Universität Heidelberg, Theodor-Kutzer-Ufer 1-3, 68167 Mannheim

Telefon: 0621 3833700

Öffnungszeiten: Heute geöffnet · 08:00–00:00Uhr ▾

Abbildung 4: Ergebnis einer Google-Suche. Das ‚Befindet sich in‘ ist eine Übersetzungsungenauigkeit, denn das englische Pendant ist ‚is part of‘ also ‚ist Teil von‘.

wie bei RDF hat, hat sich das Library-Schema (und andere Schemata von Schema.org) für Bibliotheken als einsetzbar und verständlich erwiesen. Inzwischen gibt es Erweiterungen des Stamm-Vokabulars von Schema.org wie MedicalWebPage [28], so dass auch bestimmte Bereiche, wie in diesem Fall die Webseiten von medizinischen Einrichtungen, detaillierter modelliert werden können.

Anmerkung

Diesem Artikel vorausgegangen ist ein Vortrag mit einem Abstract [29]. Das Abstract dieses Artikels ist identisch mit dem Abstract des Vortrags im Proceedings-Band. Die Folien des Vortrages finden sich unter Bohne-Lang auf <http://www.umm.uni-heidelberg.de/bibl/ueberuns/veroeffentlichungen.html>.

Interessenkonflikte

Der Autor erklärt, dass er keine Interessenkonflikte in Zusammenhang mit diesem Artikel hat.

Literatur

1. Deutsche Nationalbibliothek. Metadaten. 2016. Available from: http://www.dnb.de/DE/Standardisierung/Metadaten/metadaten_node.html
2. Pellegrini T, Blumauer A. Semantic Web und semantische Technologien: Zentrale Begriffe und Unterscheidungen. In: Pellegrini T, Blumauer A, editors. *Semantic Web – Wege zur vernetzten Wissensgesellschaft*. Berlin Heidelberg: Springer; 2006. p. 9-25.
3. Berners-Lee T. WorldWideWeb: Summary. 1991. Available from: <http://groups.google.com/forum/#!msg/alt.hypertext/eCTkk0oWTAY/bJGhZyooXzkj>
4. Berners-Lee T, Connolly D. Hypertext Markup Language (HTML). A Representation of Textual Information and MetaInformation for Retrieval and Interchange. 1993. Available from: <https://www.w3.org/Markup/draft-ietf-iiir-html-01>
5. Berners-Lee T, Connolly D. Hypertext Markup Language – 2.0. 1995. Available from: <https://tools.ietf.org/html/rfc1866>
6. Raggett D. HyperText Markup Language Specification Version 3.0. 1995. Available from: <https://www.w3.org/Markup/html3/html3.txt>
7. Arnett N, Bowman M, Christian E, Connolly D, Koster M, John K, Lagoze C, Mauldin M, Mogensen C, Nichols W, Niesen T, Weibel S, Wood A. A Proposed Convention for Embedding Metadata in HTML. 1996. Available from: <https://www.w3.org/Search/9605-Indexing-Workshop/ReportOutcomes/S6Group2>
8. Johnston P, Powell A. Expressing Dublin Core metadata using HTML/XHTML meta and link elements. 2008. Available from: <http://dublincore.org/documents/2008/08/04/dc-html/>
9. Suchoperatoren – Google Websuche-Hilfe. 2016. Available from: <https://support.google.com/websearch/answer/2466433?hl=de>
10. Berners-Lee T, Hendler J. Publishing on the semantic web. *Nature*. 2001 Apr;410(6832):1023-4. DOI: 10.1038/35074206
11. Berners-Lee T, Hendler J, Lassila O. The Semantic Web. *Scientific American*. 2001;284(5):34-43. DOI: 10.1038/scientificamerican0501-34
12. Berners-Lee T. Linked Data – Design Issues. 2006 [updated 18.06.2009]. Available from: <https://www.w3.org/DesignIssues/LinkedData.html>
13. Linked Open Vocabularies (LOV). 2016. Available from: <https://lov.okfn.org/dataset/lov/>
14. D’Arcus B, Giasson F. Bibliographic Ontology Specification – The Bibliographic Ontology. 2009. Available from: <http://purl.org/ontology/bibo/>
15. Shotton D, Peroni S. FaBiO, the FRBR-aligned Bibliographic Ontology. 2016. Available from: <http://purl.org/spar/fabio/>
16. Voß J. The Service Ontology. 2013. Available from: <http://purl.org/ontology/service>
17. Klee C, Jakob V. Holding Ontology. 2015. Available from: <http://purl.org/ontology/holding>
18. Voß J. Document Service Ontology (DSO). 2013. Available from: <http://purl.org/ontology/dso>
19. Danowski P, Pohl A, editors. (Open) Linked Data in Bibliotheken. Berlin, Boston: de Gruyter, Saur; 2013. DOI: 10.1515/9783110278736
20. Hepp M, editor. An Ontology for Describing Products and Services Offers on the Web. 16th International Conference on Knowledge Engineering and Knowledge Management (EKAW2008). Acitrezza, Italy: Springer LNCS; 2008.
21. Cookbook – GoodRelations Wiki 2016 [The GoodRelations Cookbook is a growing collection of recipes for developers]. Available from: <http://wiki.goodrelations-vocabulary.org/Cookbook>
22. GoodRelations Snippet Generator 2016. Available from: <http://www.ebusiness-unibw.org/tools/grsnippetgen/>
23. Full Hierarchy – schema.org 2016. Available from: <http://schema.org/docs/full.html>
24. Google Test-Tool für strukturierte Daten. 2016. Available from: <https://search.google.com/structured-data/testing-tool>
25. Official Google Webmaster Central Blog: Google does not use the keywords meta tag in web ranking. 2009. Available from: <https://webmasters.googleblog.com/2009/09/google-does-not-use-keywords-meta-tag.html>
26. Edwards C. Is The Meta Keyword Tag Used By Google, Bing or Yahoo? 2014. Available from: <https://chrisedwards.me/seo/keyword-meta-tag-google/>
27. Meta-Tags, die Google versteht – Search Console-Hilfe. 2016. Available from: <https://support.google.com/webmasters/answer/79812?hl=de>
28. MedicalWebPage – health-lifesci.schema.org. 2016. Available from: <https://health-lifesci.schema.org/MedicalWebPage>
29. Bohne-Lang A. Semantische Daten für den Webauftritt einer Bibliothek. In: Arbeitsgemeinschaft für medizinisches Bibliothekswesen (AGMB). Jahrestagung der Arbeitsgemeinschaft für medizinisches Bibliothekswesen (AGMB). Göttingen, 26.-28.09.2016. Düsseldorf: German Medical Science GMS Publishing House; 2016. Doc16agmb01. DOI: 10.3205/16agmb01
30. Dublin Core Metadata Element Set, Version 1.0: Reference Description. 1998. Available from: <http://dublincore.org/documents/1998/09/dces/>

Korrespondenzadresse:

Andreas Bohne-Lang
Medizinische Fakultät Mannheim, Universität Heidelberg,
Ludolf-Krehl-Str. 13–17, 68167 Mannheim, Deutschland
andreas.bohne-lang@medma.uni-heidelberg.de

Bitte zitieren als

Bohne-Lang A. Semantische Metadaten für den Webauftritt einer
Bibliothek. *GMS Med Bibl Inf.* 2016;16(3):Doc17.
DOI: 10.3205/mbi000372, URN: urn:nbn:de:0183-mbi0003721

Artikel online frei zugänglich unter

<http://www.egms.de/en/journals/mbi/2017-16/mbi000372.shtml>

Veröffentlicht: 04.01.2017

Copyright

©2017 Bohne-Lang. Dieser Artikel ist ein Open-Access-Artikel und steht unter den Lizenzbedingungen der Creative Commons Attribution 4.0 License (Namensnennung). Lizenz-Angaben siehe <http://creativecommons.org/licenses/by/4.0/>.