Recherchebericht

Übersicht

1. Begriffe:

1.1 Ressource

Eine Ressource ist eine Wissensquelle welche in der Regel über eine URI identifiziert wird. Diese kann beispielsweise auf Webseiten, E-Mail-Adressen oder auch generele Konzepte verweisen. Zusätzlich gibt es auch unbenannte Ressourcen, sogenannte leere Knoten.

1.2 URI - Uniform Resource Identifier

Ein Uniform Resource Identifier ist ein Identifikator und besteht aus einer Zeichenfolge, die zur Identifizierung einer abstrakten oder physischen Ressource dient. URIs werden zur Bezeichnung von Ressourcen wie Webseiten, Webservices, oder E-Mail-Adressen im Internet und dort vor allem im WWW eingesetzt.
Eine Erweiterung der nur aus druckbaren ASCII-Zeichen bestehenden URIs sind die Internationalized Resource Identifiers (IRIs).

1.3 Ontologie

Ontologien in der Informatik sind meist sprachlich gefasste und formal geordnete Darstellungen einer Menge von Begrifflichkeiten und der zwischen ihnen bestehenden Beziehungen in einem bestimmten Gegenstandsbereich. Sie werden genutzt um Wissen in digitalisierter und formaler Form zwischen Anwendungsprogrammen und Diensten auszutauschen. Wissen umfasst dabei sowohl Allgemeinwissen als auch Wissen über sehr spezielle Themengebiete und Vorgänge.
Ontologien enthalten Inferenz- und Integritätsregeln, also Regeln zu Schlussfolgerungen und zur Gewährleistung ihrer Gültigkeit. Sie sind ein wichtiger Bestandteil des semantischen Webs und der Wissensrepräsentation im Teilgebiet Künstliche Intelligenz. Im Unterschied zu einer Taxonomie, die nur eine hierarchische Untergliederung bildet, stellt eine Ontologie ein Netzwerk von Informationen mit logischen Relationen dar.

1.4 RDF - Resource Description Framework

Das Resource Description Framework bezeichnet eine technische Herangehensweise im Internet zur Formulierung logischer Aussagen über Ressourcen und ist damit ein grundlegender Baustein des Semantischen Webs. RDF ähnelt den klassischen Methoden zur Modellierung von Konzepten wie UML-Klassendiagramme und Entity-Relationship-Modell. Im RDF-Modell besteht jede Aussage aus den drei Einheiten Subjekt, Prädikat und Objekt, wobei eine Ressource als Subjekt mit einer anderen Ressource als Prädikat näher beschrieben wird. Mit einer weiteren Ressource oder lediglich einem Literal als Objekt bilden diese drei Einheiten ein Tripel. Um global eindeutige Bezeichner für Ressourcen zu haben, werden diese dafür nach Konvention wie URLs geformt. URLs für allgemein häufig benutzte Beschreibungen sind öffentlich bekannt und können so weltweit für den gleichen Zweck verwendet werden, was Programmen ermöglicht die Daten wiederum für den Menschen sinnvoll darzustellen.

1.5 RDFS - RDF Schema

Das Resource Description Framework Schema legt für RDF eine Syntax für den gemeinsamen Datenaustausch fest. Zur Interpretation von in RDF formulierten Aussagen bedarf es allerdings noch eines gemeinsamen Vokabulars. Ein solches Vokabular wird auch Ontologie genannt, wenn es gleichzeitig Regeln für die richtige Verwendung der in ihm definierten Ressourcen enthält. RDFS stellt ein Vokabular zur Verfügung, mit dessen Hilfe eine bestimmte Anwendungsdomäne modelliert werden kann. Außerdem können die in der Domäne vorkommenden Ressourcen, ihre Eigenschaften und Relationen untereinander durch RDFS repräsentiert werden. Man kann also mit RDFS einfache Ontologien formalisieren. RDFS liegt die Idee eines mengentheoretischen Klassenmodells zugrunde. Wichtig ist hierbei, dass Klassen und Eigenschaften separat voneinander modelliert werden. Das Klassenkonzept macht es möglich, eine formale Beschreibung der Semantik der verwendeten RDF-Elemente festzulegen.

1.6 OWL - Web Ontology Language

OWL ist eine Spezifikation um Ontologien anhand einer formalen Beschreibungssprache erstellen, publizieren und verteilen zu können. Es geht darum, Termini einer Domäne und deren Beziehungen formal so zu beschreiben, dass auch Software die Bedeutung verarbeiten kann. OWL basiert technisch auf der RDF-Syntax und historisch auf DAML+OIL und geht dabei über die Ausdrucksmächtigkeit von RDF-Schema weit hinaus. Zusätzlich zu RDF und RDF-Schema werden weitere Sprachkonstrukte eingeführt, die es erlauben, Ausdrücke ähnlich der Prädikatenlogik zu formulieren.

1.7 Data Cube

Ein Data Cube, auch OLAP Cube oder Cube-Operator genannt, ist ein in der Begriff zur logischen Darstellung von Daten. Die Daten werden dabei als Elemente eines mehrdimensionalen Würfels angeordnet. Die Dimensionen des Würfels beschreiben die Daten und

erlauben auf einfache Weise den Zugriff. Daten können über eine oder mehrere Achsen des Würfels ausgewählt werden.
Diese Art der Darstellung ist für die Analyse von Daten von Vorteil, da auf verschiedene Dimensionen der Daten auf gleiche Weise zugegriffen wird.

1.8 SPARQL - SPARQL Protocol and RDF Query Language

SPARQL ist eine graph-basierte Abfragesprache für RDF. Die Syntax weist viele Parallelen zu SQL auf. Mit SPARQL ist es möglich komplexere Anfragen zu stellen als es reine Textsuchen ermöglichen würden. Dadurch werden Inhalte nicht nur maschinenlesbar sondern auch maschinenverständlich.

1.9 CSV - Comma-Separated Values

Das Dateiformat CSV beschreibt den Aufbau einer Textdatei zur Speicherung oder zum Austausch einfach strukturierter Daten. Ein allgemeiner Standard für das Dateiformat CSV existiert nicht, die zu verwendende Zeichenkodierung ist ebenso wenig festgelegt. In CSV-Dateien können Tabellen oder eine Liste unterschiedlich langer Listen abgebildet werden. Kompliziertere, beispielsweise geschachtelte Datenstrukturen können durch zusätzliche Regeln oder in verketteten CSV-Dateien gespeichert werden.

1.10 JSON - JavaScript Object Notation

Die JavaScript Object Notation ist ein kompaktes Datenformat in für Mensch und Maschine einfach lesbarer Textform zum Zweck des Datenaustauschs zwischen Anwendungen. Jedes gültige JSON-Dokument soll ein gültiges JavaScript sein und per eval() interpretiert werden können. Davon abgesehen ist JSON aber unabhängig von der Programmiersprache.

1.11 XML - Extensible Markup Language

Die Extensible Markup Language ist eine Auszeichnungssprache zur Darstellung hierarchisch strukturierter Daten in Form von Textdateien. Die XML-Spezifikation definiert eine Metasprache, auf deren Basis durch strukturelle und inhaltliche Einschränkungen anwendungsspezifische Sprachen definiert werden. Diese Einschränkungen werden durch Schemasprachen wie DTD oder XML Schema ausgedrückt. Ein XML-Dokument besteht aus Textzeichen, im einfachsten Fall in ASCII-Kodierung, und ist damit menschenlesbar. Binärdaten enthält es per Definition nicht.

1.12 XSD - XML Schema Definition

XSD ist eine Definition von Strukturen für XML-Dokumente. Anders als bei den klassischen XML-DTDs wird die Struktur in Form eines XML-Dokuments beschrieben. Darüber hinaus wird eine große Anzahl von Datentypen unterstützt. XML Schema beschreibt in einer komplexen Schemasprache Datentypen, einzelne Dokumente und Gruppen solcher Instanzen. Im Gegensatz zu DTDs kann bei Verwendung von XML Schemata zwischen dem Namen des XML-Typs und dem in der Instanz verwendeten Namen des XML-Tags unterschieden werden.

1.13 Data Mapping

Data Mapping ist ein Prozess der zwischen zwei verschiedenen Datenmodellen eine Verbindung herstellet. Diese Datenmodelle können sowohl atomare als auch Metadaten enthalten. Mapping funktioniert wie ein abstraktes Modell um Beziehungen innerhalb einer bestimmten Domäne zu ermitteln und ist oft der erste Schritt zur Datenintegration.

1.14 SDMX - Statistical Data and Metadata Exchange

SDMX legt Standards für den austausch statistischer Informationen fest. Zwei grundlegende Formate sind SMDX-ML mit XML-Syntax und SMDX-EDI mit EDIFACT-Syntax.

1.15 Dataset

Ein Dataset bezeichnet eine größere, zusammenhängende Datenmenge und besteht aus Metadaten welche den Inhalt beschreiben und der dazugehörigen Sammlung von Einträgen.

1.16 Model

Ein Model beschreibt die Struktur eines Dataset indem es Dimensions und erlaubte Einträge definiert.

1.17 Dimension

Eine Dimension ist eine Eigenschaft eines Eintrags welche dessen Sinn festlegt. Hierbei wird auch der Datentyp definiert. Allgemein unterscheidet man zwischen Attributes welche nur einzelne Werte beinhalten und Compound Dimensions welche zusammengesetzt sind. Die wichtigsten Typen sind der Measure-Type für einfache numerische Werte und der Time-Type für Datums- und Zeitangaben.

2. Konzepte:

2.1 OpenSpending

OpenSpending ist eine offene Plattform zum Datenaustausch welche versucht alle Transaktionen zwischen Regierungen und Unternehmen aufzuzeichen und visuell zu präentieren. Der Fokus auf diese ist jedoch keine technische Einschränkung, tatsächlich werden jegliche Transaktionen unterstützt.
Der Großteil kann in zwei Arten unterteilt werden, Transactional Spending Data und Budgetary Data. Transactional Spending Data zeichnet individuelle Transaktionen auf während Budgetary Data in Kategorien zusammengefasst wird.

2.2 LOD - Linked Open Data

Linked Open Data bezeichnet im World Wide Web frei verfügbare Daten welche per URI identifiziert sind und darüber direkt per HTTP abgerufen werden können und ebenfalls per URI auf andere Daten verweisen. Idealerweise werden zur Kodierung und Verlinkung der Daten RDF und darauf aufbauende Standards wie SPARQL und OWL verwendet sodass Linked Open Data gleichzeitig Teil des Semantic Web ist. Die miteinander verknüpften Daten ergeben ein weltweites Netz, das auch als Linked Open Data Cloud oder Giant Global Graph bezeichnet wird.

2.3 Semantic Web

Das Semantic Web ist eine Instanz von semantischen Netzen und außerdem eine Erweiterung des WWW. Ziel ist es die Bedeutung von Informationen für Computer verwertbar zu machen und damit automatisch für die interessierten Nutzer im Zuge einer Abfrage zu ordnen. Die Informationen im Web sollen von Maschinen interpretiert und automatisch weiterverarbeitet werden können. Informationen über Orte, Personen und Dinge sollen mit Hilfe des Semantischen Webs auf der Basis der Inhalte miteinander in Beziehung gesetzt werden können.

2.4 Jena

Jena ist ein in Java geschriebenes Open Source Framework für Semantische Netze. Es bietet eine Programmierschnittstelle zum Laden und Speichern von Daten in RDF-Graphen. Jena repräsentiert RDF-Graphen als abstrakte Modelle im Speicher oder in Datenquellen wie Dateien oder Datenbanken welche auch auf OWL beruhen können. Die Modelle können mittels SPARQL abgefragt und mittels SPARUL verändert werden. Jena arbeitet intern mit verschiedenen Reasonern und kann auch von externen Reasonern bedient werden.

2.5 LIMES - Link Discovery Framework for Metric Spaces

LIMES ist ein Framework zum Auffinden von Links im Datennetz. Es implementiert zeiteffiziente Methoden für umfangreiche Linkermittlungen basierend auf der Charakteristik des metrischen Raumes. Das Framework besteht aus sieben erweiterbaren Hauptmodulen wovon die beiden wichtigsten das Controller Module und das Data Module sind.

2.6 SILK - Semantic Inferencing on Large Knowledge

SILK ist ein System zur Wissenspräsentation, welche eine Sprache, eine Programmlogik, eine Benutzerobefläche und Möglichkeiten zum Austausch beinhaltet. SILK setzt bei der grundlegenden Anforderung das Semantic Web zu sehr großen Wissensbasen für Wissenschaft und Business zu skalieren an. Es erweitert unter anderem die Möglichkeiten der Wissenspräsentation von SPARQL.

2.7 SAIM - Semi-Automatic Instance Matcher

SAIM ermöglicht die Verknüpfung von Wissensbasen im Semantic Web. Dabei lieg der Schwerpunkt beim Instant-Matching, also dem Vergleich bezüglich dem Grad der Ähnlichkeit zwischen zwei verschiedenen Beschreibungen realer Objekte um gleiche reale Objekte zu finden, sehr großer Wissensbasen., welche als SPARQL-Endpunkte erreichbar sind. SAIM nutzt Techniken des maschinellen Lernens und ist kompatibel mit LIMES und SILK.

2.8 REST - Representational State Transfer

REST bezeichnet ein Programmierparadigma für Webanwendungen. Es gibt keine explizite Norm, daher gehen die Vorstellungen, was REST ist, auseinander. Im Allgemeinen bezeichnet REST die Idee, dass eine URL genau einen Seiteninhalt als Ergebnis einer serverseitigen Aktion wie das Anzeigen einer Trefferliste nach einer Suche darstellt, wie es der Internetstandard HTTP für statische Inhalte bereits vorsieht.
Es gibt fünf Eigenschaften, die ein REST-Dienst haben muss, wobei es den einzelnen Diensten überlassen ist, wie es implementiert wird: Adressierbarkeit, unterschiedliche Repräsentation, Zustandslosigkeit, Operationen und Verwendung von Hypermedia.

2.9 Maven

Maven ist ein Build-Management-Tool der Apache Software Foundation und basiert auf Java. Mit ihm kann man insbesondere Java-Programme standardisiert erstellen und verwalten. Maven versucht, den Grundgedanken der Konvention vor Konfiguration konsequent für den gesamten Zyklus der Softwareerstellung abzubilden. Dabei sollen Software-Entwickler von der Anlage eines Softwareprojekts über das Kompilieren, Testen und Packen bis zum Verteilen der Software auf Anwendungsrechnern so unterstützt werden, dass möglichst viele Schritte automatisiert werden können. Folgt man dabei den von Maven vorgegebenen Standards, braucht man für die meisten Aufgaben des Build-Managements nur sehr wenige Konfigurationseinstellungen zu hinterlegen, um den Lebenszyklus eines Softwareprojekts abzubilden.

3. Aspekte:

3.1 DBPedia

DBPedia ist ein Teil des Semantic Webs, der strukturierte Informationen aus
Wikipedia gewinnt, um sie als Linked Open Data nutzbar zu machen. Als Datenstandard wird das Resource Description Framework (RDF) genutzt.
Die Daten werden auf HTML-Seiten aufbereitet, mittels SPARQL können außerdem komplexere Anfragen an die Daten gestellt werden.

3.2 LinkedGeoData

Das Ontowiki ist ein semantisches Programm für Wissensmanagement im Semantic Web Kontext. Zum Verwalten der „Wissens“ in Form von maschinenlesbaren Daten gibt es ein Web-Benutzerinterface, in denen Klassen, Ressourcen und Einstellungen verwaltet werden können. Weitere Funktionen sind ein integrierter LinkedData Server, unabhängigkeit vom genutzten Backend zur Datenspeicherung, Erstellen von Wiki-Seiten oder das Darstellen von Geodaten und Visualisierung von statistischen Daten mittels CubeViz.

3.3 OntoWiki

Das Ontowiki ist ein semantisches
Programm für Wissensmanagement im Semantic Web Kontext. Zum
Verwalten der „Wissens“ in Form von maschinenlesbaren Daten gibt
es ein Web-Benutzerinterface, in denen Klassen, Ressourcen und
Einstellungen verwaltet werden können. Weitere Funktionen sind ein
integrierter LinkedData Server, unabhängigkeit vom genutzten Backend

zur Datenspeicherung, Erstellen von Wiki-Seiten oder das Darstellen
von Geodaten und Visualisierung von statistischen Daten mittels

3.4 CubeViz

CubeViz ist ein Browser um statistische Daten über das RDF Data Cube Vokabular, kompatibel mit SDMX, interaktiv darzustellen. Hierfür werden Filter- und Visualisierungsoptionen angeboten.

4. Quellen

http://svn.aksw.org/papers/2013/openspending2rdf/public.pdfh
http://svn.aksw.org/papers/2013/openspending2rdf/openspending/public.pdf
http://slidewiki.org/deck/750_semantic-data-web-lecture-series
http://community.openspending.org/help/guide/
http://aksw.org/Projects/LIMES.html
http://aksw.org/Projects/SAIM.html
http://silk.semwebcentral.org/
http://jena.apache.org/
http://linkedgeodata.org/about
https://github.com/AKSW/OntoWiki/wiki
http://aksw.org/Projects/CubeViz.html