UNIVERSITÄT
DUISBURG
ESSEN

Taxonomien & Ontologien

Seminararbeit

vorgelegt von

Andreas Tacke

Soziales Retrieval im Web 2.0

Sommersemester 2008

Arbeitsgruppe Informationssysteme

Datum: 05. Oktober 2008

Betreuung:
Prof. Dr.–Ing. Norbert Fuhr

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einführung

Der ständig wachsende individuelle Informationsbedarf und die täglich größer werdende Menge an Informationen im Web stellen eine der großen kulturellen Herausforderungen unserer Zeit dar. Erschwerend ist dabei die Tatsache, dass der überwiegende Teil heutiger Web–Inhalte lediglich darauf ausgerichtet ist, vom Menschen gelesen und interpretiert zu werden. Gegenwärtige Suchmaschinen können aus diesem Grunde nur auf syntaktische Informationen zurückgreifen. Eine einfache und übergreifende Möglichkeit, semantische Angaben zu ergänzen beziehungsweise zu manipulieren, fehlt. So ist es beispielsweise nahezu unmöglich, gezielt nach Adressen, Personen, Datumsangaben etc. zu suchen. Man bräuchte eine Möglichkeit, diese Konzepte einheitlich darzustellen, damit sie von Suchmaschinen berücksichtigt werden können.
Ein weiteres Problem, das mit dem Aufkommen des so genannten *Web 2.0* einhergeht, ist, dass immer mehr Informationen in Form von Mutlimediadaten vorliegen, auf die außerhalb ihrer jeweiligen Domäne kein Zugriff besteht (vgl. *YouTube*, *MySpace* und *Yahoo! Flickr*). ähnlich verhält es sich mit dem so genannten *Invisible Web*, das die Web–Inhalte beschreibt, die beispielsweise dynamisch generiert werden oder sich hinter Webservices verbergen.
Im Rahmen dieser Arbeit sollen zwei verschiedene Ansätze behandelt werden, die sich mit den oben genannten Problemstellungen befassen und Lösungsansätze zur Verfügung stellen.
In Kapitel 2 sollen zunächst *Taxonomien* als simple, aber dafür einfach handhabbare Methode zur inhaltlichen Strukturierung von Informationen diskutiert und an zwei Beispielen aus der Praxis näher erläutert werden.
In Kapitel 3 werden dann *Ontologien* vorgestellt. Diese besitzen zwar Merkmale, die weit über die Möglichkeiten von Taxonomien hinausgehen, sind dafür jedoch komplexer und aufwändiger zu implementieren. Um die Rolle von Ontologien im Web zu umreißen, soll in diesem Kapitel vor allem das *Semantic Web* mit seinen Kerntechnologien beschrieben werden, bevor mit *FOAF* die Brücke zu sozialen Netzwerken geschlagen werden soll.
Das nachfolgende Kapitel befasst sich mit der Übertragbarkeit der vorgestellten Ansätze auf das Web 2.0 und im letzten Kapitel folgt dann eine kritische Einschätzung der Thematik.

2 Taxonomien

Der Begriff *Taxonomie* wird heute — teils missbräuchlich — für viele verschiedene Konzepte verwendet, meist bezeichnet er jedoch eine Art abstrakte, hierarchische Struktur. Ursprünglich geht er zurück auf den schwedischen Na-

turwissenschaftler Carl von Linné, der im 18. Jahrhundert eine hierarchische Klassifikation für Lebewesen entwarf, die in weiterentwickelter Form nach wie vor in der Biologie eingesetzt wird [1].
Taxonomie leitet sich ab von den griechischen Wörtern *táxis* und *nomos* und bedeutet frei übersetzt so viel wie *Ordnungsgesetz*. Im Laufe der Jahre hat sich der Begriff neben der Biologie vor allem in den Bibliotheks– und Sprachwissenschaften etabliert, in denen er für Begriffsklassifikationen verwendet wird. Die Bedeutung in der Informationswissenschaft, die dieser Arbeit zu Grunde liegt, ist direkt aus der Bibliothekswissenschaft entlehnt.
Bei dieser Interpretation einer Taxonomie wird ein *geschlossenes* (oder auch *kontrolliertes*) *Vokabular*, das aus einer Menge von *Termen* zu einem bestimmten Thema besteht, in einer hierarchischen Struktur angeordnet [1]. Wie es sich mit nicht geschlossenen Vokabularen verhält, soll im Kapitel über Ontologien näher erläutert werden.
Mathematisch gesehen handelt es sich bei einer Taxonomie um eine Baumstruktur. Es gibt ein eindeutiges Wurzelelement, von dem alle weiteren Knoten ausgehen. Zwischen dem Wurzelelement, den nachfolgenden Knoten und deren Kindern besteht also jeweils eine überbegriff–Unterbegriff–Relation. Die meisten Taxonomien, die heute im Web zu finden sind, verfügen darüber hinaus über Referenzen von Begriffen zu verwandten Begriffen, anschaulich also andere Zweige der Baumstruktur, womit die Grenzen zur Ontologie verschwimmen. Ontologien besitzen jedoch noch einige Alleinstellungsmerkmale, die eine klare Trennung der beiden Konzepte rechtfertigen. Auf diese soll an anderer Stelle genauer eingegangen werden.

2.1 Beispielimplementierungen

Taxonomien eignen sich sehr gut, um große Mengen von Informationen anhand gemeinsamer Charakteristika zu strukturieren. Um zu veranschaulichen, wie sie praktisch im Web Anwendung finden, sollen im Folgenden zwei Beispielimplementierungen vorgestellt werden: zum einen das *Open Directory Project* und zum anderen das *Drupal Taxonomy Module*.

2.1.1 Open Directory Project

Das *Open Directory Project*, im Jahre 1998 von zwei Mitarbeitern der Firma Sun gegründet, ist ein so genanntes Web–Verzeichnis. Dabei handelt es sich um eine Taxonomie zur inhaltlichen Klassifikation von Webseiten. Mittlerweile wird das Projekt von Netscape betrieben, das seinerseits zum Time–Warner–Konzern gehört. Gestützt wird das Projekt — ähnlich wie bei Wiki-

pedia — von einer Community aus Freiwilligen, die entsprechende Einträge vornehmen und bearbeiten, womit man es zu den *Web–2.0*–Anwendungen zählen kann [2].
Die Bezeichnung *Web–Verzeichnis* ist darauf zurückzuführen, dass die Kategorien zur Klassifikation der Webseiten in Form von Verzeichnissen und Unterverzeichnissen angeordnet sind. Unter Wurzelverzeichnis *Top* befinden sich derzeit die Unterverzeichnisse *Arts*, *Business*, *Computers*, *Games*, *Health*, *Home*, *News*, *Recreation*, *Reference*, *Science*, *Shopping*, *Society* und *Sports*. Neben der inhaltlichen hat sich im Laufe der Zeit auch eine regionale Klassifikation herauskristallisiert, so dass man von zwei parallelen Klassifikationen sprechen kann.
Die Informationsgewinnung geschieht durch browsen der Taxonomie (s. Abb. 1), wobei die gewünschten Themen sukzessive eingegrenzt werden. Dadurch eignet sie sich besonders dann, wenn das eigene Informationsbedürfnis nicht genau spezifiziert werden kann. Browsing ist jedoch nicht die einzige Möglich-

Abbildung 1: Open Directory Project — Taxonomie–Browser

keit, nach Informationen zu suchen. Hat man schon genauere Informationen über das gesuchte Thema gefunden, kann man über ein Suchfeld Anfragen formulieren, die auf der Baumstruktur arbeiten und gegebenenfalls entsprechende Verzeichnisse zurückliefern. Des Weiteren steht die komplette Taxonomie als so genannter *RDF–Dump* zur Verfügung. Bei *RDF* handelt es sich um das verwendete Metadatenformat, welches später im Bezug auf das *Semantic Web* noch Erwähnung finden wird. Diese Daten dienen unter anderem

als Grundlage für die Verzeichnisdienste vieler Internetportale, wie beispielsweise denen von Google, AOL oder Netscape.

2.1.2 Drupal Taxonomy Module

Bei *Drupal* handelt es sich um ein Projekt für ein freies so genanntes *Content Management System*, eine Softwareplattform zur Erstellung und Verwaltung von Webseiten. Drupal wird insbesondere zum Aufbau von Web–Communities eingesetzt und bietet in seiner Standardinstallation Funktionen wie Blogs, Foren und eine Benutzerarchitektur. Zusätzliche Funktionalität wird über eine Vielzahl von Software–Modulen bereitgestellt.
Ein besonderes Merkmal von Drupal ist das *Taxonomy Module* zur Organisa-

Abbildung 2: Drupal Taxonomy Module

tion von Inhalten einer Webseite. Das Taxonomy Module erlaubt die Erstellung von *Vokabularen* zur Klassifikation von Blog–Einträgen, Kommentaren, Foren–Beiträgen etc. Ein Vokabular ist hierbei zuerst einfach eine (unendliche) Menge von Begriffen, die den erwähnten Objekten zugeordnet werden können. Insofern unterscheidet es sich zunächst nicht vom so genannten *Tagging* (Verschlagwortung), wie man es aus anderen Blog–Systemen kennt. Beim Taxonomy Module können die Begriffe darüber hinaus jedoch in einer Hierarchie angeordnet werden. Inhalte, die mit entsprechenden Begriffen aus dieser Hierarchie versehen werden, werden dann automatisch in diese eingeordnet. Des Weiteren kann eine Webseite mehrere solcher Begriffshierarchien definieren. Dies kann, wie schon beim Open Directory Project gezeigt, von

Nutzen sein, wenn die Inhalte nach verschiedenen Kriterien klassifiziert werden sollen.
Bei der Erstellung von Vokabularen werden zwei Modi unterschieden. Zum einen kann ein moderierter Ansatz gewählt werden, bei dem nur Nutzer, die zuvor mit den entsprechenden Rechten ausgestattet wurden, Begriffe definieren und Inhalte damit versehen können. Zum anderen kann dies auch durch alle Nutzer geschehen (vgl. *Folksonomy*) [3]. Ist ein Objekt mit einem Begriff versehen worden, erscheint dieser in Form eines Hyperlinks unter der jeweiligen Überschrift (s. Abb. 2). Folgt man dem Link, werden nicht nur die Inhalte aufgelistet, die ebenfalls mit diesem Begriff versehen wurden, sondern auch all jene, denen ein Unterbegriff zugeordnet wurde. Zusammen mit der Möglichkeit, direkt auf die Begriffe der Hierarchie zuzugreifen, ergibt sich somit eine innovative und umfassende Lösung zur Inhaltsorganisation in einer Web–2.0–Anwendung.

2.2 Stärken und Schwächen

Die Beispiele haben gezeigt, dass Taxonomien ein einfaches und zuverlässiges Konzept zur Organisation von Informationen im Web darstellen können. Für den Benutzer sind sie insofern einfach nachvollziehbar, als dass die hierarchische Organisation von Daten im Allgemeinen ein vertrautes Prinzip ist, wie man es beispielsweise von der Ordermetapher bei Dateisystemen moderner Betriebssysteme gewohnt ist. Ein weiterer Vorteil von Taxonomien ist die vergleichsweise einfache Implementierbarkeit, da auf simple Datenstrukturen wie Bäume zurückgegriffen werden kann.
Auf der anderen Seite muss man für diese Vorzüge Abstriche der Flexibilität und Beschreibungsfähigkeit hinnehmen. Natürliche Denkprozesse werden mitunter schlecht nachgebildet. Wie bei der in Abbildung 3 gezeigten Literaturtaxonomie zu sehen ist, lässt sich nur nach einem Hauptkriterium anordnen, in diesem Fall nach Epochen. Dies bedeutet, dass jede Gattung für jede Epoche gegebenenfalls erneut aufgeführt werden muss. Dabei werden nicht nur Redundanzen erzeugt, es gibt auch keine direkte Verbindung zwischen Autoren, die derselben Gattung angehören, aber in unterschiedlichen Epochen geschrieben haben. Des Weiteren können äquivalente Begriffe wie etwa „Doktorand“ und „PhD Student“ nicht erfasst werden. Die oben beschriebene Möglichkeit, andere Zweige zu referenzieren, kann dies nur unzureichend kompensieren. Zwar existieren mit *Thesauri* und *Topic Maps* noch zwei Technologien, die weitere Beziehungen zwischen Begriffen einer Taxonomie einführen, diese sollen hier aber aufgrund des begrenzten Umfanges dieser Arbeit nicht behandelt werden.
Ontologien bieten weiterhin alle Möglichkeiten von Taxonomien und darüber

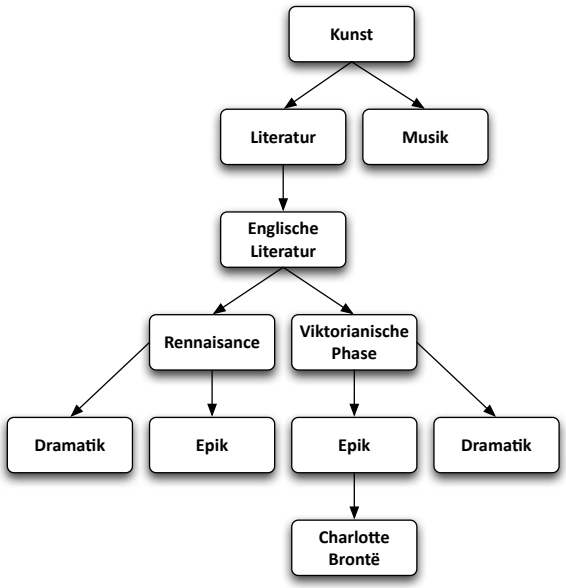

Abbildung 3: Grafische Darstellung einer Literaturtaxonomie

hinaus Ansätze, die angesprochenen Beschränkungen aufzuheben.

3 Ontologien

Der Begriff *Ontologie* stammt ebenfalls aus dem Griechischen und setzt sich aus den Wörtern *on*, dem Genitiv von „Sein“ und *logos*, „Lehre von“, zusammen. Ihren Ursprung hat die Ontologie in der Metaphysik, einem Teilgebiet der Philosophie, und befasst sich in diesem Zusammenhang mit der Frage, warum etwas existiert. In der Informationstechnologie wurde der Begriff in den Achtziger Jahren des letzten Jahrhunderts von Forschern auf dem Gebiet der künstlichen Intelligenz für die Modellierung von Wissen aufgegriffen. Allgemein sind Ontologien im informationswissenschaftlichen Sinne seitdem als formale Beschreibungen von Konzepten innerhalb einer Wissensdomäne definiert [4]. Ontologien sorgen also für ein gemeinsames Verständnis dieser Konzepte.
Die Intention war es, Wissensdomänen in maschinenlesbarer Form zu modellieren, um so eine Schnittstelle für agentenbasierte Softwaresyteme zu haben, die auf dieses Wissen zurückgreifen, um Aufgaben mit Hilfe von automatisiertem Schließen zu lösen. Mittlerweile haben sich Ontologien als zentrale Komponente in Wissenssystemen herauskristallisiert. Ein Grund hierfür ist, dass sie einen Abstraktionsgrad besitzen, der sie agnostisch im Bezug auf die

zu Grunde liegende Technologie macht. Um dies zu gewährleisten, werden Ontologien mit Hilfe so genannter *Wissensrepräsentationssprachen* ausgedrückt. Es gibt eine Vielzahl solcher Wissensrepräsentationssprachen, oder auch *Ontologiesprachen*, da der Fokus dieser Arbeit aber auf Webanwendungen liegt, sollen diese nur am Beispiel der *Web Ontology Language* (OWL) behandelt werden.
Der grundlegende Aufbau einer Ontologie ist unabhängig von der verwendeten Sprache. Ontologien setzen sich aus den folgenden drei Hauptelementen zusammen [5]:

- *Classes* sind die Konzepte der Wissensdomäne. In Anlehnung an Kapitel 1 könnte zum Beispiel eine Person ein solches Konzept sein. *Classes* können dabei entweder vom Typ *concrete* oder *abstract* sein, je nachdem ob sie instanziert werden dürfen oder nicht. Wie im Kapitel über Taxonomien bereits erwähnt, besitzen auch Ontologien eine hierarchische Struktur. Man spricht in diesem Zusammenhang allerdings von Vererbung, da eine Unterklasse automatisch alle Eigenschaften der Oberklasse „erbt“. Hier unterscheidet sich die Ontologie grundlegend von einer Taxonomie. Welche Auswirkungen dies im einzelnen hat, soll im Laufe des Kapitels anhand einiger Beispiele näher erörtert werden.

- *Slots* sind die Attribute einer *Class*. Dies können sowohl andere *Classes*, wie zum Beispiel die Person des Vaters oder der Mutter, als auch beschreibende Attribute, etwa der Name der Person sein. Für jeden *Slot* wird überdies dessen Kardinalität festgelegt.

- *Individuals* sind die eigentlichen Instanzen von *Classes*. Hierzu werden die *Slots* mit Werten belegt. Ein *Individual* wäre im oben genannten Beispiel also eine Person, die durch einen bestimmten Namen und zwei andere Instanzen von Person, nämlich Vater und Mutter, identifiziert ist.

Anwendungsbeispiele sollen in den Kapiteln über die *Web Ontology Language* und die Ontologie–Entwicklungsumgebung *Protégé–2000* präsentiert werden. Wenn man Ontologien im Web diskutiert, ist es unvermeidlich, auch über das *Semantic Web* zu sprechen, in dem sie eine zentrale Rolle spielen.

3.1 Semantic Web

Wie in Kapitel 1 bereits erwähnt, mangelt es heutigen Webseiten an semantischen Informationen. Um die Nachteile dieses Umstandes genauer zu verdeutlichen, sei folgender Anwendungsfall gegeben: Gesucht werden Telefonnummer und Anschrift aller Arztpraxen von Orthopäden in Essen, NRW, die

auch am Samstag mindestens von 10 Uhr bis 12 Uhr Sprechstunden anbieten. Auch ohne besondere Kenntnis der Materie zu besitzen wird dem Anwender heutiger Suchmaschinen schnell deutlich, dass die vorliegende Aufgabe nicht ohne Weiteres mit den gegeben technischen Möglichkeiten zu lösen ist. Eine aktuelle Google-Suche mit der Anfrage „orthopäde essen sprechstunde samstags“ liefert beispielsweise an erster Stelle bereits einen Orthopäden aus Hanau. Zeitraumangaben können ebenfalls nicht sinnvoll berücksichtigt werden. Der Grund hierfür ist, dass Suchmaschinen die Informationen über so genannte *Crawler*–Programme direkt aus dem Text einer Webseite beziehen.
Um auch die gezielte Suche nach semantischen Informationen zu ermöglichen, schlug Tim Berners-Lee im Jahre 1999 mit dem Semantic Web eine Erweiterung des World Wide Web vor, die die formale Repräsentation dieser Informationen gewährleisten sollte. Oberstes Ziel der Semantic–Web–Initiative ist es, die Suche nach Informationen mit Hilfe von *Software–Agentensystemen* zu ermöglichen. Diese werden dann mit Informationen wie den oben genannten programmiert und suchen im Web nach Ressourcen, die diesen entsprechen.
Man entschied sich dazu, die semantischen Informationen in Form von Ontologien zu beschreiben, da diese sich wegen ihrer in Kapitel 3 beschriebenen Technologieunabhängigkeit besonders für die Verwendung im Web–Kontext eignen [6].
Damit die Software-Agenten auf diese Informationen zugreifen können, wurde vom World Wide Web Consortium mit RDF ein universelles Metadatenformat entwickelt, das diese Aufgabe erfüllen soll.

3.1.1 RDF und OWL

Das *Resource Description Framework* (RDF) wurde ursprünglich als reines Metadatenformat für das Semantic Web konzipiert, hat sich im Laufe seiner Lebensdauer jedoch zu einer universellen Architektur zur Beschreibung von Informationen mit Hilfe diverser Syntax–Formate weiterentwickelt. Da sich diese Arbeit auf das Web konzentriert, soll hier nur die für diesen Kontext relevante *RDF/XML*–Variante vorgestellt werden. Die Verwendung der Syntax–Elemente von RDF/XML speziell für die Wissensrepräsentation ist in der *RDF Schema Vocabulary Description Language* festgehalten.
Allgemein werden Ressourcen in RDF mit Subjekt–Prädikat–Objekt–Ausdrücken, so genannten *Triplets*, beschrieben [7]. Wie dies im konkreten Fall aussieht, soll in dieser Arbeit am Beispiel der *Web Ontology Language*, kurz *OWL*, dargelegt werden.
Bei OWL handelt es sich syntaktisch um eine Untermenge von RDF/XML und semantisch von RDF Schema, die speziell für die Erstellung maschinenlesbarer Ontologien für das Semantic Web definiert wurde [8].

Abbildung 4 verdeutlicht anhand eines Ausschnitts aus einer in OWL kodier-

```
<owl:Ontology rdf:about="Example">
  <rdfs:label rdf:datatype="http://www.w3.org/2001/
     XMLSchema#string">
    Genre Classification Layer
  </rdfs:label>
  <rdfs:comment rdf:datatype="http://www.w3.org/2001/
     XMLSchema#string">
    Represents the genre classification
  </rdfs:comment>
</owl:Ontology>

<owl:Class rdf:ID="Bebop">
  <rdfs:subClassOf rdf:resource="#Contemporary_Jazz"/>
</owl:Class>
<owl:Class rdf:ID="New_Orleans_Jazz">
  <rdfs:subClassOf rdf:resource="#Classic_Jazz"/>
</owl:Class>
<owl:Class rdf:ID="Jazz_Pop">
  <rdfs:subClassOf rdf:resource="#Rock"/>
  <rdfs:subClassOf rdf:resource="#Contemporary_Jazz"/>
</owl:Class>
```

Abbildung 4: OWL–Codebeispiel in RDF Schema

ten Ontologie über Musikgenres, wie die in Kapitel 3 eingeführten Elemente in der OWL–Syntax dargestellt werden.

3.1.2 FOAF–Framework

Das Akronym *FOAF* steht für *Friend of a Friend* und ist ein Projekt mit dem Ziel, Semantic-Web-Technologie mit sozialen Netzwerken zu verbinden. Dazu entwarf man eine Ontologie zur Beschreibung von Personen und deren Beziehungen zu anderen Personen und (Web–)Ressourcen. In einem Artikel aus dem Jahr 2007 griff Tim Berners–Lee die Idee auf und prägte die Sichtweise vom Web als so genannten *Giant Global Graph*, also einem weltumspannenden Netzwerk aus Personen und Inhalten [9]. Mit FOAF wird versucht, ein standardisiertes, maschinenlesbares Format zur Beschreibung dieses Graphen bereitzustellen. Abbildung 5 zeigt einen vereinfachten Aus-

schnitt aus einem solchen Graphen, die einzelnen Elemente sollen im Folgenden erläutert werden. Dazu wird eine Ontologie verwendet, die in RDF

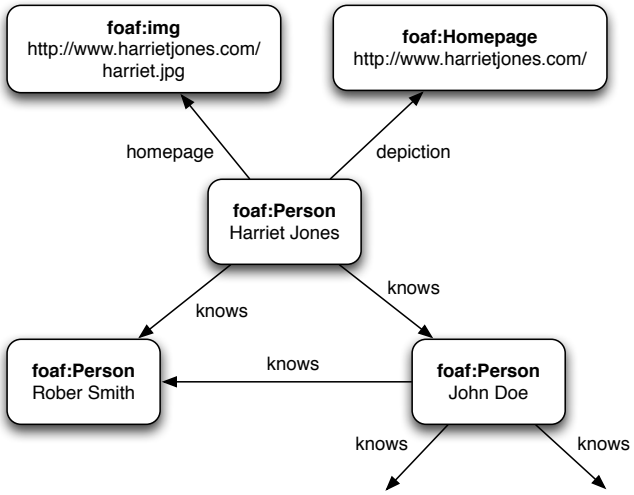

Abbildung 5: Vereinfachte Darstellung eines FOAF–Graphen

Schema dargestellt wird. FOAF stellt ein Vokabular in einem eigenen Namensraum zur Verfügung, mit dem die eigene Identität und die Beziehungen zu anderen Personen ausgedrückt werden kann. Ein wichtiger Grundgedanke bei der Entwicklung von von FOAF war, die Anwendung einfach und dezentral zu gestalten. Dazu erstellt eine Person ein eigenes FOAF-Profil und veröffentlicht dieses auf einer persönlichen Webseite oder einer vergleichbaren Webressource. Ein solches Profil identifiziert eine Person eindeutig und ist damit vergleichbar mit einer *Jabber–ID*[1] oder einer *ICQ-UIN*[2]. In diesem Profil beschreibt man zunächst die eigene Identität und gibt dann eine Liste von Verweisen auf Profile von Personen an, die man kennt [10]. Abbildung 6 zeigt ein beispielhaftes FOAF–Profil in RDF Schema, in dem einige grundlegende Begriffe aus dem FOAF–Vokabular vorgestellt werden. FOAF bietet noch eine Reihe weiterer Begriffe wie etwa *foaf:workplacehomepage*, so dass sich eine Person nahezu beliebig genau selbst beschreiben kann.

3.2 Protégé–2000

Mit der Entwicklung von Ontologien befasst sich in der Informatik das *Ontology Engineering*. Da es sich bei Ontologien mitunter um sehr große und

[1]Wikipedia–Definition einer Jabber–ID: http://tinyurl.com/43mjj6
[2]Wikipedia–Eintrag über ICQ–Nummern: http://tinyurl.com/3g39hp

```
<foaf:Person>
  <foaf:surname>Tacke</foaf:surname>
  <foaf:name>Andreas Tacke</foaf:name>
  <foaf:firstName>Andreas</foaf:firstName>
  <foaf:gender>male</foaf:gender>
  <foaf:img rdf:resource="http://www.is.inf.uni-due.de
     /staff/images/tacke.jpg" />
  <foaf:homepage rdf:resource="http://www.is.inf.uni-
     due.de/staff/tacke.html.en" />
  <foaf:interest dc:title=Antoine Dufour" rdf:resource
     ="http://www.candyrat.com/artists/AntoineDufour/"
      />
  <foaf:based_near geo:lat="41.378665" geo:long="
     2.164598" />

  <foaf:knows>
    <foaf:Person>
      <foaf:name>Stefan Tomanek</foaf:name>
      <rdfs:seeAlso rdf:resource="http://www.is.inf.
         uni-due.de/staff/tomanek.rdf" />
    </foaf:Person>
  </foaf:knows>
</foaf:Person>
```

Abbildung 6: FOAF–Codebeispiel in RDF Schema

komplexe Strukturen handelt, wäre es sehr zeitaufwändig und fehleranfällig, diese direkt in einer Ontologiesprache zu schreiben.
An der Stanford University wurde aus diesem Grunde das Projekt *Protégé–2000* ins Leben gerufen, mit dem Ziel, den Entwicklungsprozess übersichtlicher und allgemein handhabbarer zu gestalten. Protégé–2000 unterstützt diverse Ontologiesprachen über eine Plug–In–Architektur, ist im Rahmen dieser Arbeit aber insbesondere durch die Unterstützung von OWL von Bedeutung. Dadurch eignet es sich besonders für die Entwicklung von Ontologien für das Semantic Web [11].
Der Arbeitsablauf von Protégé–2000 sieht zunächst die Erstellung der Klassenhierarchie vor (s. Abb. 7). Im nächsten Schritt werden für jede Klasse die entsprechenden Slots definiert. Im letzten Arbeitsschritt können schließlich Instanzen von Klassen angelegt werden. Der komplette Ablauf ist dabei völlig unabhängig von der Sprache, in der die Ontologie letztlich abgespei-

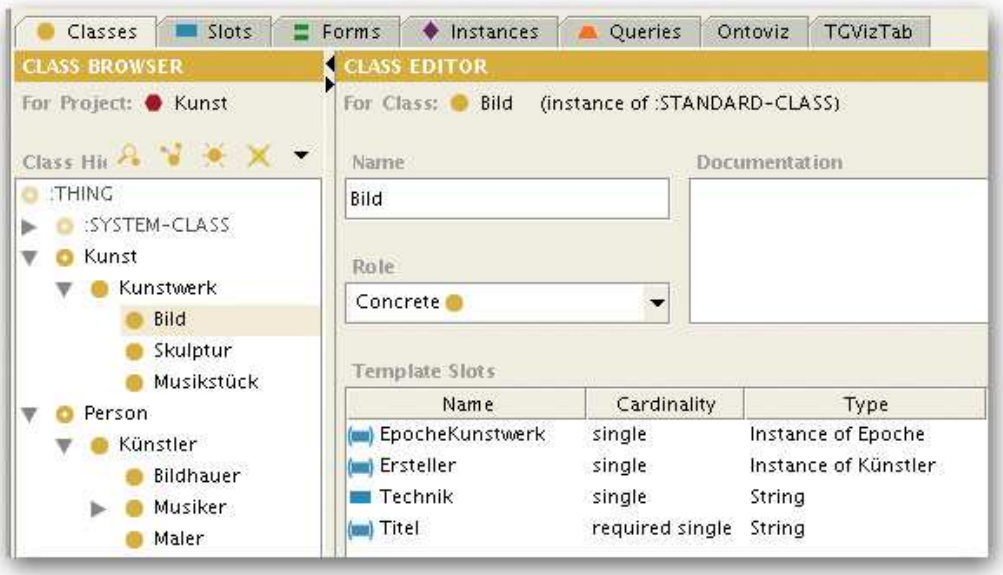

Abbildung 7: Protégé–2000 — Class–Browser und –Editor

chert werden soll.
Des Weiteren bietet Protégé–2000 noch Werkzeuge zur Visualisierung und Validierung. So können etwa Ontologien als Graph dargestellt oder Anfragen an diese formuliert werden.

3.3 Ontologiebasiertes Retrieval

Wenn Informationen erst einmal in Form einer Ontologie kodiert sind, gibt es verschiedene Möglichkeiten, diese wieder zu extrahieren. Die einfachste Möglichkeit ist das Browsen mit Hilfe eines Werkzeuges, wie etwa dem in Kapitel 3.2 vorgestellten Protégé–2000. Dies würde jedoch keinerlei Vorteil gegenüber Taxonomien einbringen. Das in Kapitel 3.1 formulierte Ziel war die Informationsgewinnung auf Basis von Software–Agentensystemen. Dazu bedarf es einer Möglichkeit, Anfragen an eine Ontologie zu formulieren. Vom *World Wide Web Consortium* wurde für diesen Zweck die *Simple Protocol and RDF Query Language*, kurz *SPARQL*, entwickelt.

3.3.1 Simple Protocol and RDF Query Language

SPARQL erlaubt es, Anfragen an RDF–Graphen, wie etwa eine in OWL kodierte Ontologie, zu formulieren. Eine typische SPARQL–Anfrage besteht aus einem *Select*–Statement, in dem Variablen deklariert werden, an die dann

Werte gebunden werden und einem *Where*–Statement, in dem Bedingungen für die zurückzuliefernden Werte deklariert werden können. Des weiteren gibt es die Möglichkeit über das Schlüsselwort *Prefix* Kürzel für eine URI[3] zu vergeben, die dann in den oben genannten Statements verwendet werden können, um die Lesbarkeit der Anfrage zu verbessen.
Die Triplets des RDF–Graphen werden dann mit den Bedingungen in der Where–Klausel abgeglichen und an die Variablen gebunden. Die Typisierung ist dabei dynamisch, das heißt Klassen werden ausschließlich über ihre Attribute „gematcht“ [12].
Im vorliegenden Beispiel (s. Abb. 8) werden alle Klassen vom Typ „Genre“

```
PREFIX genres: <http://example.com/genreOntology#>
SELECT ?genre
WHERE {
  ?x genres:ID ?genre;
     genres:subClassOf ?y .
  ?y genres:ID "Rock".
}
```

Abbildung 8: Beispiel für eine SPARQL–Anfrage

zurückgeliefert, die Unterklassen vom Genre mit der ID „Rock“ sind. Bezug nehmend auf die in OWL kodierte Beispielontologie aus Kapitel 3.1.1 würde also `rdf:ID="Jazz_Pop"` zurückgeliefert.
SPARQL ist bereits in mehreren Programmiersprachen implementiert[4] und ist damit für die Entwicklung der in Kapitel 3.1 beschriebenen Software–Agenten prädestiniert.

3.4 Vor– und Nachteile

Ontologien bieten gegenüber Taxonomien einige Verbesserungen. So ist es beispielsweise möglich, die in Kapitel 2.2 erwähnten Strukturen darzustellen. Abbildung 9 veranschaulicht, wie man die Taxonomie aus Abbildung 3 in eine Ontologie überführen würde. Anstatt für jede Epoche wiederholt alle Gattung auflisten zu müssen, wurde hier einfach eine neue Klasse Gattung eingeführt, von der die Klasse Autor nun „erbt“. Dabei könnte ein Autor auch von mehreren Gattungen erben, was in der vorliegenden Abbildung lediglich nicht berücksichtigt wird.

[3] W3C-Definition einer URI: http://www.w3.org/Addressing/
[4] Liste der SPARQL–Implementierungen: http://esw.w3.org/topic/SparqlImplementations

Ein weiterer wichtiger Vorteil von Ontologien ist, dass sie ein einheitliches

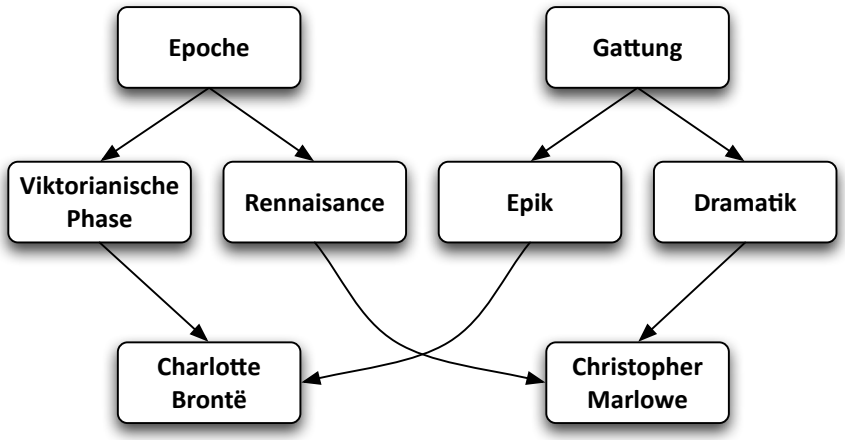

Abbildung 9: Grafische Darstellung einer Literaturontologie

Verständnis und eine einheitliche Darstellung von Konzepten innerhalb einer Wissensdomäne gewährleisten. Dies ist für eine spätere maschinelle Verarbeitung der Ontologiedaten unerlässlich.
Der Entwurf von Ontologien ist ein komplizierter und langwieriger Vorgang. Nicht umsonst gibt es mit Ontology Engineering eine eigene wissenschaftliche Disziplin, die sich mit diesem Thema auseinandersetzt. Insbesondere im Web–Kontext kommt erschwerend hinzu, dass mitunter viele Individuen an der Modellierung einer Wissensdomäne beteiligt sind und eine Einigung mit wachsendem Umfang immer problematischer wird.

4 Anwendbarkeit im Web 2.0

Wie aus den Beispielen in Kapitel 2 hervorgeht, sind Begriffshierarchien im Allgemeinen und Taxonomien im Besonderen im Web 2.0 bereits Realität. Aufgrund ihrer niedrigen Komplexität sind Taxonomien technisch einfach zu realisieren und werden bei der Organisation von Inhalten im Web 2.0 angesichts wachsender Informationsbestände vermutlich an Bedeutung gewinnen. Dabei stehen sie in direkter Konkurrenz zu Tags, die einerseits noch weniger Aufwand bei der Realisierung bedeuten, andererseits bei der klaren Strukturierung von Informationen noch Nachteile haben, die aber durch neue wissenschaftliche Erkenntnisse kompensiert werden könnten [13]. Aufgrund der vielen unterschiedlichen und erfolgreichen Implementierungen ist es auf lange Sicht fraglich, ob sich die Taxonomie im Web–Kontext standardisieren

lässt. Des Weiteren gibt es zu diesem Zeitpunkt keine umfassende Initiative, übergreifende, taxonomiebasierte Formate im Web zu etablieren, so dass sie wohl in Zukunft in diesem Zusammenhang keine wichtigere Rolle spielen werden, als sie es ohnehin schon tun. Dies liegt nicht zuletzt daran, dass es aufgrund der vielen verschiedenen Beziehungstypen schwierig ist, Netzwerke im Allgemeinen und soziale Netzwerke im Besonderen adäquat mit Taxonomien zu beschreiben. Hier bieten Ontologien klare Vorteile.
Zwar sind Ontologien im Web bislang eine weitestgehend unrealisierte Idee, werden aber vom World Wide Web Consortium im Rahmen der Semantic–Web–Initiative gestützt. Wie am Beispiel des FOAF–Frameworks verdeutlicht wurde, eignen sich Ontologien hervorragend, um der immer weiter voranschreitenden Konvergenz von Personen und (Multimedia–)Inhalten im Web Rechnung zu tragen. Erfahrungsgemäß ist es jedoch schwierig, derartige Standards im Web zu etablieren, wenn kein unmittelbarer Anreiz besteht. Vielmehr gibt es eine Vielzahl so genannter De–facto–Standards, die durch erfolgreiche Anwendungen entstanden sind. Im Web 2.0 haben sich einige wenige Plattformen für bestimmte Anwendungszwecke herauskristallisiert. Es läge nun also an Branchengrößen wie *Google*, *MySpace*, *Facebook* oder *Yahoo!*, sich auf ein gemeinsames Datenaustauschformat für Benutzerprofile zu einigen, das idealerweise auf einem offenen, ontologiebasierten Standard wie FOAF basiert. Dadurch würden Lösungen für die übergreifende, semantische Suche nach Informationen in sozialen Netzwerken ermöglicht. Dies hätte natürlich auch datenschutzrechtliche Implikationen. Der Schaden durch ein kompromittiertes Benutzerprofil würde sich beispielsweise auf alle sozialen Netzwerke ausdehnen, in denen man angemeldet ist.
Anders als Taxonomien und Tagging eignen sich Ontologien aufgrund ihrer Komplexität schlecht als Basis für die Strukturierung von Web–Inhalten durch Benutzer. Bei der Klassifikation von Inhalten, wie sie in Kapitel 2.1 vorgestellt wurde, müsste für den Nutzer weitestgehend transparent sein, dass eine Ontologie zu Grunde liegt. Das bedeutet wiederum, dass sich „Insellösungen“ bilden, da jeweils nur ein stark begrenztes Vokabular verwendet werden kann. Es bleibt abzuwarten, ob sich Ontologien angesichts der starken Konkurrenz durch Tagging–Systeme langfristig in diesem Bereich durchsetzen werden.

5 Zusammenfassung

In dieser Arbeit wurden zunächst die Probleme bei der Suche nach Informationen im Web und insbesondere im Web 2.0 beschrieben. Einer der wichtigsten Punkte dabei war das Fehlen einer Möglichkeit, semantische Informa-

tionen in maschinenlesbarer Form zu ergänzen beziehungsweise zu manipulieren.
Als Lösungsvorschläge wurden zunächst Taxonomien als in ihrer Funktionalität beschränkte aber einfach anwendbare Werkzeuge zur Strukturierung und Klassifikation von Informationen im Web und — mit Hilfe zweier Beispielimplementierungen — im Web 2.0 diskutiert.
Das nachfolgende Kapitel befasste sich mit Ontologien, die die Beschreibungsfähigkeiten von Taxonomien mit Hilfe von Techniken wie Mehrfachvererbung und Restriktionsregeln erweitern und sich deshalb besser dazu eignen, um komplexere Strukturen — wie etwa in sozialen Netzwerken — abzubilden. Funktionsweise und Bedeutung von Ontologien wurden dann im Zusammenhang mit dem Semantic Web erörtert werden, für das sie die zentrale technologische Komponente bilden. Des weiteren wurde mit FOAF eine Brücke vom Semantic Web zum Web 2.0 im Allgemeinen und sozialen Netzwerken im Besonderen geschlagen. Mit Protégé–2000 sollte dann ein praktischer Einblick in die Entwicklung von Ontologien gegeben werden, bevor mit SPARQL als Abschluss des Kapitels eine Technologie zur Extraktion von Informationen aus (Web–)Ontologien vorgestellt wurde.
In Kapitel 4 wurde dann unter Berücksichtigung zuvor beschriebener Probleme, wie der hohen Komplexität von Ontologien, diskutiert, inwieweit sich die in den vorangegangen Kapitel erarbeiteten Ansätze im Web 2.0 anwenden lassen.

Literatur

[1] Lars Marius Garshol. Metadata? Thesauri? Taxonomies? Topic Maps!, 2004. Available from World Wide Web: http://www.ontopia.net/topicmaps/materials/tm-vs-thesauri.html [cited 12.09.2008].

[2] Open Directory Project. Available from World Wide Web: http://www.dmoz.org/about.html [cited 12.09.2008].

[3] Drupal Taxonomy Module. Available from World Wide Web: http://drupal.org/handbook/modules/taxonomy [cited 12.09.2008].

[4] Tom Gruber. Ontology - Definition in Encyclopedia of Database Systems, September 2007. Available from World Wide Web: http://tomgruber.org/writing/ontology-definition-2007.htm.

[5] Natalya F. Noy and Deborah L. McGuinness. Ontology development 101: A Guide to Creating Your First Ontology. Technical report, Stanford University, Stanford, CA, 94305, 2001.

[6] Tim Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, page 18, May 2001.

[7] Resource Description Framework. Available from World Wide Web: http://www.w3.org/RDF/ [cited 14.09.2008].

[8] Deborah L. McGuinness and Frank van Harmelen. OWL Web Ontology Language Overview. Available from World Wide Web: http://www.w3.org/TR/owl-features/.

[9] Tim Berners-Lee. Giant Global Graph, November 2007. Available from World Wide Web: http://dig.csail.mit.edu/breadcrumbs/node/215 [cited 28.09.2008].

[10] Edd Dumbill. Finding Friends With XML and RDF, 2002. Available from World Wide Web: http://www.ibm.com/developerworks/xml/library/x-foaf.html [cited 11.09.2008].

[11] Protégé-2000 Ontology Editor. Available from World Wide Web: http://protege.stanford.edu/doc/users.html [cited 25.09.2008].

[12] Eric Prud'hommeaux and Andy Seaborne. SPARQL Query Language For RDF, January 2008. Available from World Wide Web: http://www.w3.org/TR/rdf-sparql-query/ [cited 23.09.2008].

[13] Paul Heymann and Hector Garcia-Molina. Collaborative Creation of Communal Hierarchical Taxonomies in Social Tagging Systems. Technical report, Stanford University, Stanford, CA, 94305, April 2006.