Freie Universität Berlin

Kundenprojekt Web Technologien

Ontology and Metadata Extraction
Metadata Schemas and Upper Ontologies

Dr. Lyndon J B Nixon

Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme

mailto: nixon@inf.fu-berlin.de
http://page.mi.fu-berlin.de/nixon

Überblick

Freie Universität Berlin

- Ontology Extraction
  - From XML to RDF
- Ontology Engineering
  - Best Practices
- Metadata Extraction
  - GRDDL
  - Concept Mapping
- Metadata Schemas
  - IMDB
- Upper Ontologies
  - Open Directory
  - DBPedia

Ontology Extraction

Freie Universität Berlin

- Erstes Problem: wie wird ein RDF Schema erstellt?
- Lösung → Es gibt schon ein XML Schema als „Grundlage"
- XML und RDF nicht gleich: Baum vs. Graph, Syntax vs. Semantik

- „Ontology Extraction" heißt
  - Erzeugung einer Ontologie (ihrer Klassen, Prädikaten, Axiomen) durch die Analyse von nicht-ontologischen Dateien, z.B.: Text, Datenbank, XML
- Semiautomatische Verfahren
  - Das Ergebnis braucht in der Regel einige Verbesserungen

Ontology Extraction aus XML

Freie Universität Berlin

- XML hat Elemente, Attribute, Elementinhalte (Character Data)
- RDF hat Klassen, Prädikaten, Axiomen (C1 subClass C2, P1 subProperty P2, P1 domain C1, P2 range C2)

- Mapping definieren: XMLS Struktur → RDFS Struktur

- Beispiel: Sergey Melnik
  http://www-db.stanford.edu/~melnik/rdf/fusion.html
  - XML Elemente werden hier als Prädikate interpretiert, außer wenn ein `rdf:instance` darauf hinweist, dass es als Klasse zu interpretieren ist.

Beispiel: XML → Klassen/Prädikaten

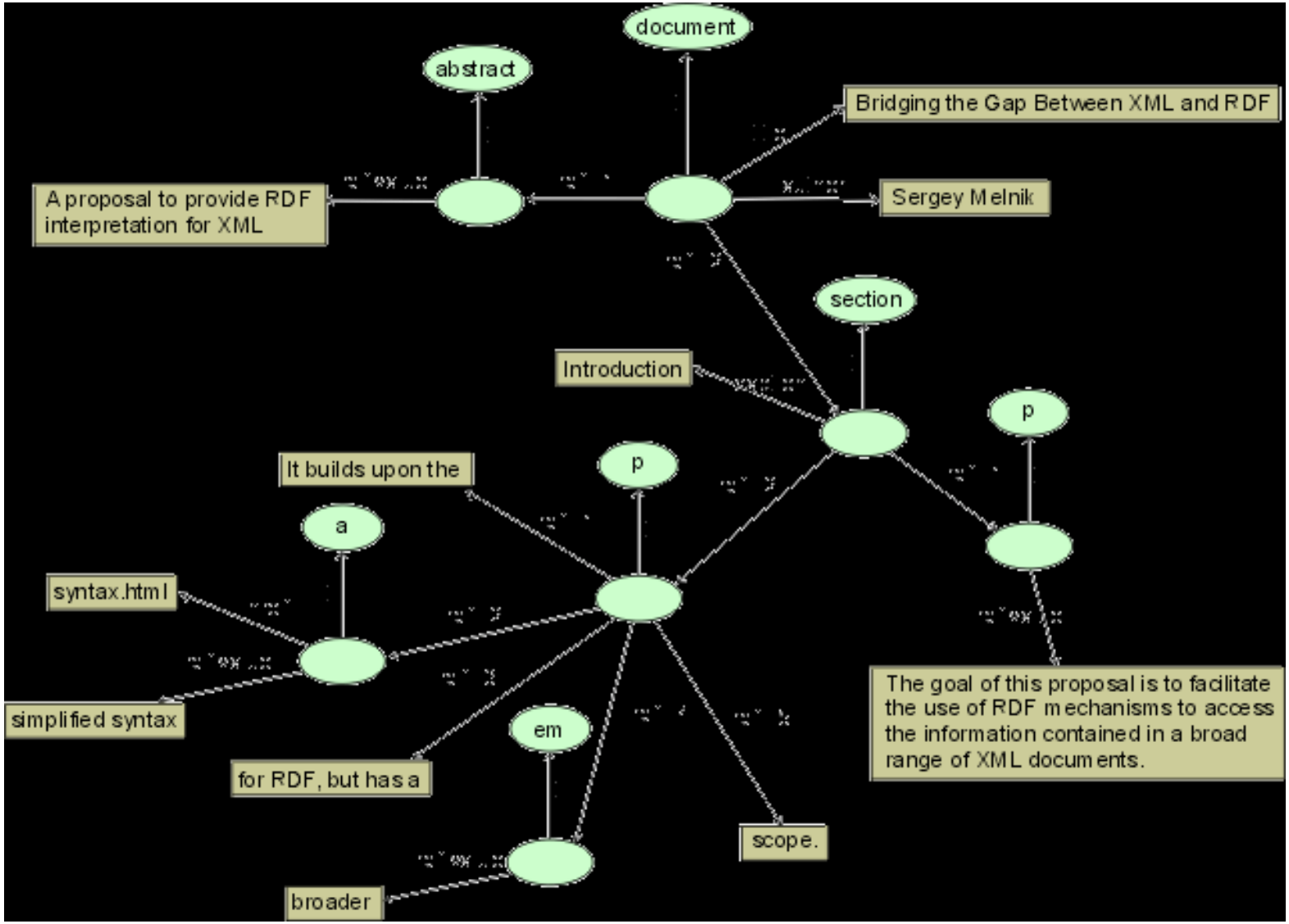
Freie Universität Berlin

```
<?xml version="1.0">
<document>
  <title>Bridging the Gap between RDF and XML</title>
  <author>Sergey Melnik</author>
  <abstract>A p
   XML</abstract>
  <section caption="Introduction">
    <p>The goal of this proposal is to facilitate the use of RDF
   mechanisms
          to access the information contained in a broad range of
   XML documents.</p>
    <p>It builds upon a <a href="syntax.html">simplified
   syntax</a> for RDF,
          but has a <em>broader</em> scope.</p>
  </section>
</document>
```

Document hat strukturierten Inhalt

Author hat unstrukturierten Inhalt

Beispiel: RDF (von Sergey Melnik)



Beispiel: XML -> Axiomen

```
<?xml version="1.0">
<document>
  <title>Bridging the Gap between RDF and XML</title>
  <au hasSection range Section
  <abstract>A proposal to provide RDF interpretation for
  XML</abstract>
  <section caption="Introduction">
    <p>The goal of this proposal is to facilitate the use of RDF
  mechanisms
        to access the information contained in a broad range of
  XML documents.</p>
    <p>It builds upon a <a href="syntax.html">simplified
  syntax</a> for RDF,
        but has a <em>broader</em> scope.</p>
  </section>
</document>
```

XML → RDF Mapping

Freie Universität Berlin

- Option 1: automatisiertes Mapping
  - z.B. XSLT, weil RDF eine XML Serialisierung hat
  - XPath wählt XML nodes
  - XSLT Templates erzeugen RDF Schema Statements

- Option 2: manuelles Mapping → Arbeitsweise
  - Entscheiden, ob Elemente/Attribute als Klassen oder Prädikate zu mappen sind
  - Entscheiden, welche Axiome hinzufügen sind
  - Entscheiden, ob es mehr Arbeit wäre, die Mappings in XSLT zu schreiben oder manuell RDF Schema zu formulieren

Ontology Engineering

Freie Universität Berlin

- Ontology Engineering ist ein Fachgebiet, in dem Methoden, Verfahren und Tools für die Erstellung von Ontologien erforscht werden.
- Siehe auch http://ontoworld.org/wiki/Ontology_Engineering

- Gutes Anfängerdokument „**Ontology Development 101: A Guide to Creating Your First Ontology**"

  http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html

Ontology Engineering 101

Freie Universität Berlin

1) There is no one correct way to model a domain there are always viable alternatives. The best solution almost always depends on the application that you have in mind and the extensions that you anticipate.

2) Ontology development is necessarily an iterative process.

3) Concepts in the ontology should be close to objects (physical or logical) and relationships in your domain of interest. These are most likely to be nouns (objects) or verbs (relationships) in sentences that describe your domain.

Ontology Best Practices

Freie Universität Berlin

- Ontology Best Practices heisst
  - „non-enforceable guidelines for ensuring quality of the ontology“
- Grundsätzliche Entscheidungen können später große Folgen haben
  - Klasse vs. Instanz
  - subClass/subProperty Verhältnisse
  - Klasse vs. Prädikat
- Komplexere Fragen zu Ontologieerstellung

Best Practices (1): Klasse oder Instanz

Freie Universität Berlin

- Ist ein Konzept als eine ***Klasse*** oder eine ***Instanz von einer Klasse*** zu modellieren?
  - Ist `Chardonnay` eine Instanz von `WhiteWine` oder ist es eine Klasse (subClassOf `WhiteWine`)?

- Entscheidung hängt vom Anwendungsgebiet und Ziel der Ontologie:
  - Welche Granularität wird gewünscht? Instanzen sind die **meist spezifischen** Konzepten in einer Ontologie.
  - Beispiel:
    - eine Ontologie für Weinempfehlungen hat Weinsorten als meist spezifische Konzept → `Chardonnay` hier wäre eine Instanz.
    - Eine Ontologie für ein Restaurant, wo einzelne Weinflaschen als Konzepte modelliert wird, hätte `Chardonnay` als Klasse.
  - Konzepte, die zusammen eine Hierarchie bilden, sind besser als Klassen zu modellieren.

Best Practices (2a): Hierarchien

Freie Universität Berlin

- Ontologien erlauben Klassen- oder Prädikathierarchie zu bilden und diese als Inferenz zu benutzen.

- subClassOf: *A subclass of a class represents a concept that is a "kind of" the concept that the superclass represents* (***Spezialisierung***)

- Wann eine neue Subklasse?
  - *„(1) have additional properties that the superclass does not have, or*
  - *(2) restrictions different from those of the superclass, or*
  - *(3) participate in different relationships than the superclasses*"

Best Practices (2b): Hierarchien

Freie Universität Berlin

- Siblings (Geschwister) - zwei Klassen mit der selben Superklasse: *All the siblings in the hierarchy must be at the same level of generality*

- Wieviele Siblings?
  - eine Klasse mit nur einer Subklasse ist vielleicht unnötig oder unvollständig;
  - eine Klasse mit zu vielen Subklassen benötigt vielleicht Zwischenklassen

Best Practices (3): Klasse oder Prädikat

Freie Universität Berlin

- Ist ein Konzept besser als eine Klasse oder durch hinzufügen eines Prädikats zu modellieren?
  - z.B. ist `WhiteWine` als neue Klasse zu modellieren oder führen wir auf `Wine` ein Prädikat `colour` mit Wert `white`?

- Entscheidung liegt an der Wichtigkeit des Konzepts in der Ontologie
  - Haben `WhiteWine`s andere Verhältnisse als `Wine`s?
    - Wenn wir Axiomen beschränkt auf `WhiteWine`s brauchen, dann muss sie eine Klasse sein.
  - Werden Instanzen von `WhiteWine` oft ihre Klasse ändern müssen?
    - Wenn wir Prädikaten nutzen, deren Wert sich ändern könnte, ist es nicht sinnvoll, das Konzept mit einer Klasse zu ersetzen, z.B. `AvailableWine`.
  - *„Usually numbers, colours, locations are properties and do not cause the creation of new classes*.

Best Practices: Komplexere Fragen

Freie Universität Berlin

- Semantic Web Best Practices and Deployment (SWBPD) Working Group sammelt komplexere Fragen von Ontologiee-Entwickler
- Siehe http://www.w3.org/2001/sw/BestPractices/
- Working Group Notes zu:
  - Defining N-ary Relations on the Semantic Web: Use With Individuals
  - Representing Classes As Property Values on the Semantic Web
  - XML Schema Datatypes in RDF and OWL
  - A Semantic Web Primer for Object-Oriented Software Developers
- Am besten: vermeide komplexere Fragen!

Metadata Extraction

Freie Universität Berlin

- Nachdem man ein RDF Schema erzeugt hat, kann man Metadaten nach diesem Schema erstellen
- Option 1: alles manuell schreiben
  - Langweilig und aufwendig!
- Option 2: RDF Statements aus anderen Daten extrahieren
  - Verfahren definieren und durchführen
  - Ergebnisse prüfen
- Metadata Extraction heisst
  - „Erzeugung von Metadata (RDF Statements, die gültig nach einem RDF Schema sind) durch die Analyse von nicht-ontologischen Dateien, z.B. Text, Datenbank, XML“

GRDDL

Freie Universität Berlin

- Spezifikation für RDF Metadata Erzeugung mit XSLT Stylesheets

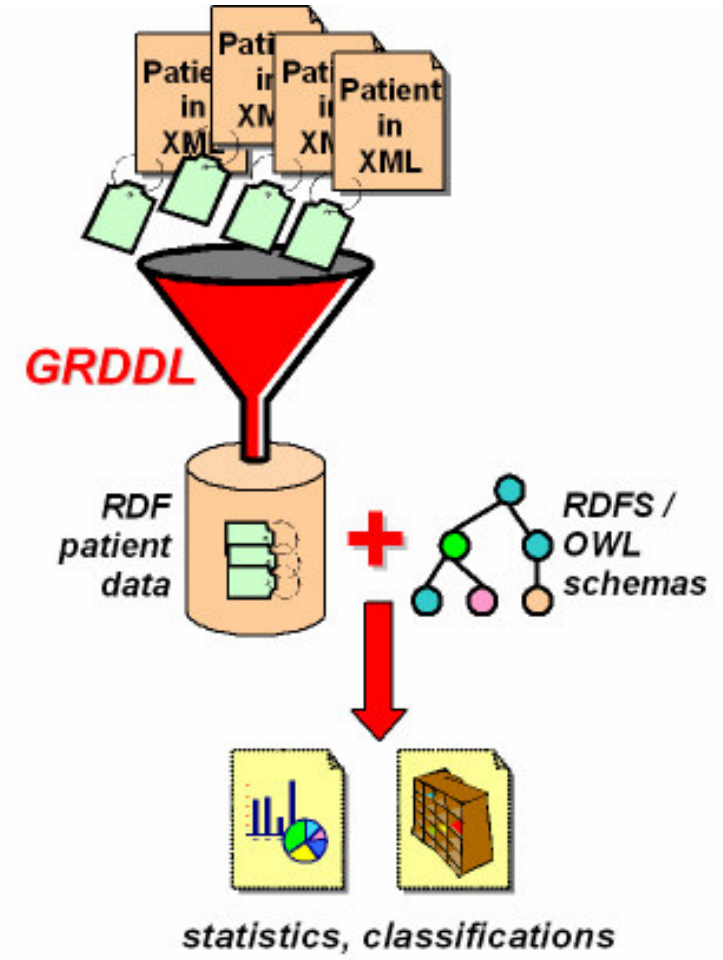
- Fokus ist eher für das Web (XHTML Seiten, Microformats) aber
- kann uns zeigen, wie RDF von XML erzeugt werden könnte

- Siehe http://www.grddl.org/

GRDDL Use Cases

- Siehe http://www.w3.org/TR/2007/NOTE-grddl-scenarios-20070406/

1. Eine Transformation von XML zu RDF wurde erstellt
2. GRDDL ermöglicht einem Client diese Transformation zu finden
3. Mit einem XSLT Processor wird RDF Data aus den XML Dateien extrahiert
4. Mit dem RDF Schema kann man diese RDF Metadata semantisch abfragen



GRDDL Transformation

Freie Universität Berlin

- XML → RDF mit XSLT: Grundlagen
  - 1. XML Elemente wählen, die Klassen in dem RDFS sind
  - 2. Instanz von dieser Klasse erzeugen (mit eindeutiger URI oder als Blank Node, d.h. ohne URI)

```
<_0> rdf:type s:Wine
```

  - 3. Kinderelemente wählen, die Prädikate in dem RDFS sind
  - 4. Statements mit dieser Instanz als Subject erzeugen

```
<_0> dc:title „This is a title“
```

  - 5. Kinderelemente wählen, die Klassen in dem RDFS sind, einem Instanz erzeugen (Schritt 2) und dann Statement mit diesem Instanz als Objekt erzeugen

    (Beachte schon erzeugte Instanzen!)

```
<_1> rdf:type s:Vineyard
<_0> s:hasVineyard <_1>
```

Concept Mapping

Freie Universität Berlin

- Quelldokumente: Konzepte sind Wörter in natürlicher Sprache
- RDF: Konzepte sind eindeutig durch ein URI zu identifizieren und können einer RDF Klasse zugeordnet werden
- Problem: wie findet man Konzepte in den Quelldokumenten und wie werden ihnen URIs zugeordnet

- XML: z.B. `<actor>Will Smith</actor>`
- RDF, könnte man das noch als „Literal" (simple datatype, hier String) interpretieren

```
<s:hasActor>Will Smith</s:hasActor>
```

- Was ist wenn wir auch über Will Smith reden wollen?

```
<s:hasActor
rdf:resource=„http://www.imdb.com/WillSmith"/>
```

Concept Mapping Verfahren

Freie Universität Berlin

- Man nutzt eine Heuristik – es ist nicht immer garantiert, dass es funktioniert

- Mapping: String → URI
- Mapping könnte klassenspezifisch sein
  - Actors kommen vielleicht aus einem anderen Namensraum als Tennisspieler
- Wenn möglich, kleine Unterschiede in der natürliche Sprache berücksichtigen
  - Will Smith und Will E. Smith sind vielleicht dieselbe Person
- Wenn die Quelldokumenten konsistent sind, umso besser für ein Mapping!
- Algorithm: String Parsing, dann URI Auswahl

Metadata Schemas

Freie Universität Berlin

- Anderer Vorschlag von Ontology Engineering 101:

  Ontologien wiederverwenden, dort wo möglich!

- Also, gibt es schon andere Ontologien, die für Sie relevant sein könnten?
  - IMDB in RDF
  - Open Directory Project
  - DBPedia

IMDB in RDF

Freie Universität Berlin

- IMDB Struktur als RDF Schema

  http://www.csd.abdn.ac.uk/~ggrimnes/dev/imdb/IMDB.rdfs

- Anderer Versuch

  http://www.cs.umbc.edu/~skallu1/IMDb.pdf (Leider ist die RDFS Datei nicht mehr erhältlich)

- Nicht vollständig, vielleicht gibt aber Ideen ☺
- Konzept URIs: IMDB gibt eine eindeutige URI zu jeder Instanz (Film, Schauspieler, Regisseur…)

Upper Ontologies

Freie Universität Berlin

- Was sind Upper Ontologies?
  - *„an ontology which describes very general concepts that are the same across all domains„*
- Nutzlich als Quelle für URIs, die Konzepte identifizieren
- Open Directory Project (ODP) als Quelle für Kategorien
- DBPedia (Wikipedia in RDF!) als Quelle für individuelle Konzepte, z.B. Personen, Orte

Open Directory Project

- Eine offene Directory von Web Links, organisiert unter Kategorien ähnlich zu Yahoo's Directory

Open Directory Project in RDF

Freie Universität Berlin

- Inhalt wird auch als RDF Dump angeboten
- Siehe http://rdf.dmoz.org/
- structure.rdf.u8.gz gibt den Kategorienstruktur in RDF
- Beispiel:

```
<Topic r:id="Top/Arts/Movies">
  <narrow r:resource="Top/Arts/Movies/Characters"/>
  <narrow r:resource="Top/Arts/Movies/Showtimes"/>
  <narrow r:resource="Top/Arts/Movies/Reviews"/>
</Topic>
```

- Mit Namensraum http://dmoz.org/RDF ergibt z.B. die URI für Movies:http://dmoz.org/RDF/Top/Arts/Movies
- Könnte Berechnung von semantischer „Ähnlichkeit" ermöglichen – Suche nach Movie Reviews bedeutet Showtimes auch relevant, oder Movies selber.

DBPedia: Wikipedia in RDF

Freie Universität Berlin

- The DBpedia project [extracts] structured information from Wikipedia and by making this information available on the Semantic Web.

- The DBpedia dataset currently provides information about more than 1.95 million “things”, including at least 80,000 persons, 70,000 places, 35,000 music albums, 12,000 films. Altogether, the DBpedia dataset consists of 103 million pieces of information (RDF triples).

- Siehe http://dbpedia.org/
- Beispielskonzept Woody Allen http://dbpedia.org/page/Woody_Allen

DBPedia - Nutzung

Freie Universität Berlin

- Vergibt jedem Wikipedia-Konzept eine eindeutige URI
  - Wie in Wikipedia, Konzepte wurde unterschieden
  - http://dbpedia.org/page/Apple ist das Obst
  - http://dbpedia.org/page/Apple_Inc ist die Firma
- Bietet ein SPARQL Endpoint, damit eine Anwendung Konzepte abfragen könnte
  - Möglichkeit um Konzepte zu finden, z.B. Konzepte mit Namen „Woody Allen“
- Durch dynamische Abfragen (u.a. mit JavaScript oder PHP im Web), Inhalt über ein Konzept immer aktuell halten

- Kurz: durch die Nutzung von DBPedia URIs entsteht die Möglichkeit, Ihre Metadata mit anderen Metadata im Web zu erweitern!

Das Ende

Freie Universität Berlin

- Fragen?