

# Ontologiebasierende Textklassifikation mittels mathematischer Verfahren

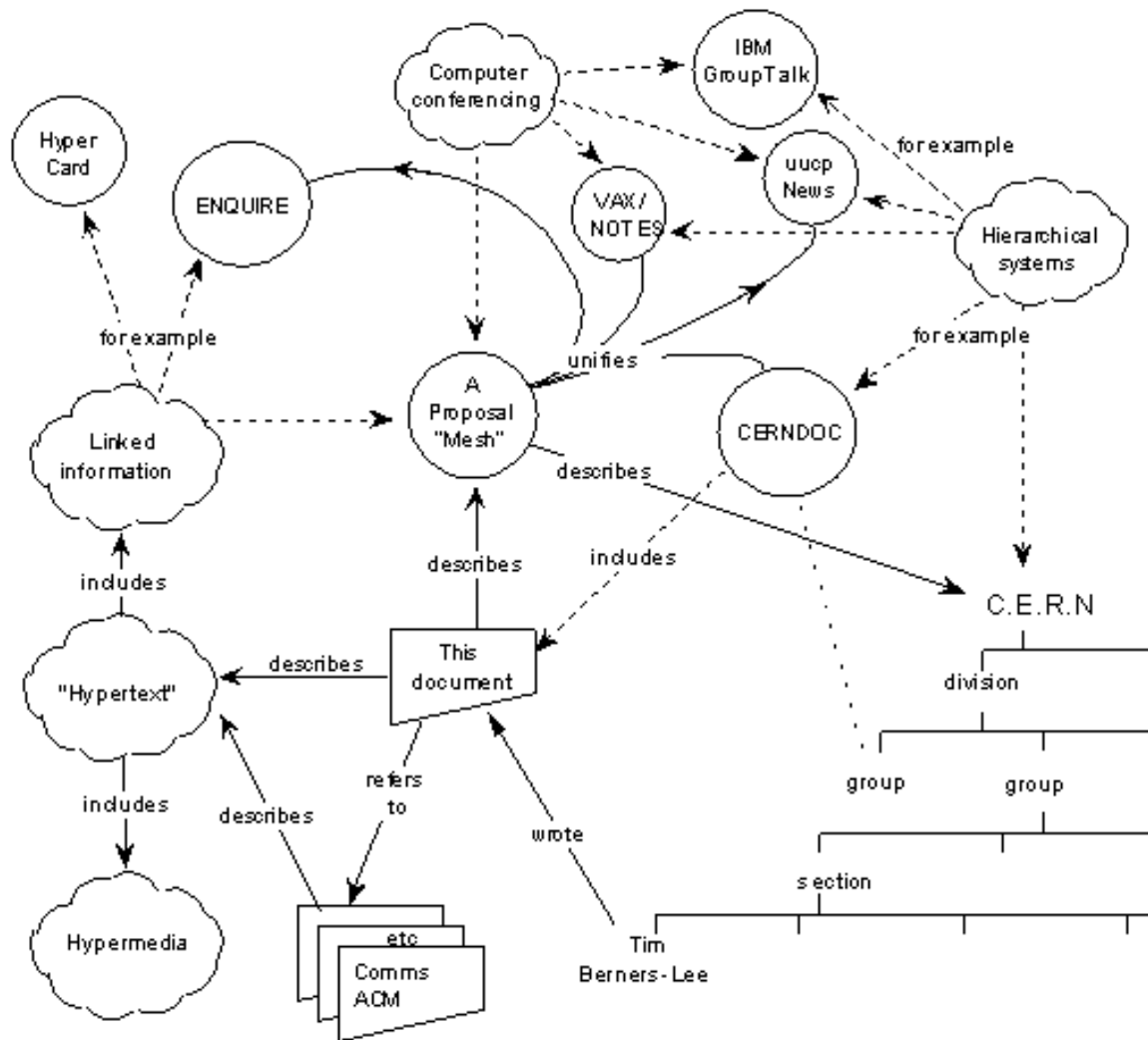
Albert Weichselbraun

23. April 2004

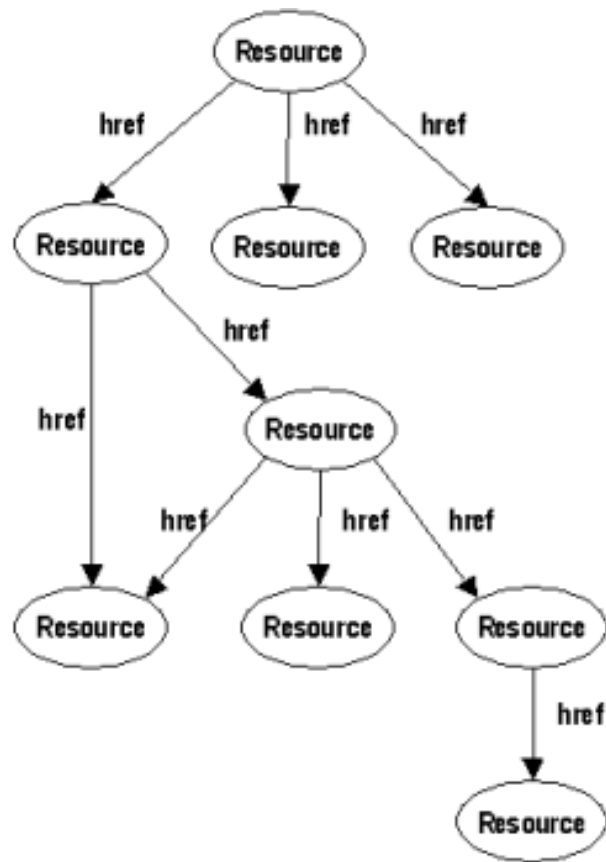
# Agenda

---

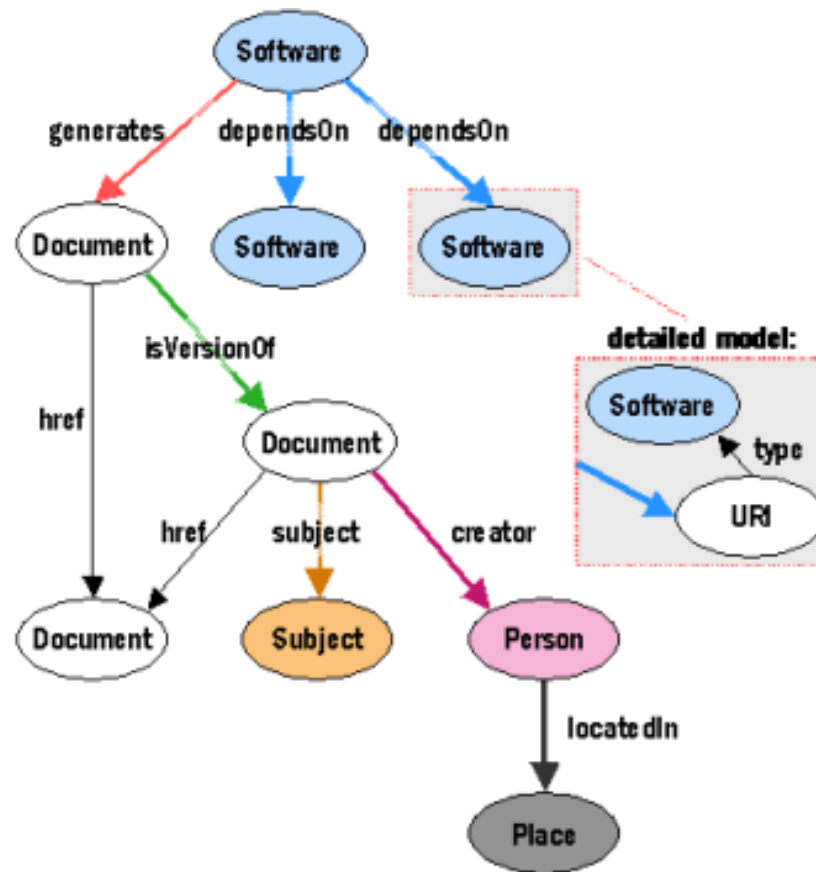
- Semantische Web
- Ontologien
  - Definitionen,  $\Delta$ : Syntax - Semantik
  - Motivation
  - Technologie
- Wissenschaftliche Problemstellung
  - Problematik
  - Dissertationsprojekt



## Ursprüngliches World Wide Web Proposal (Berners-Lee, 1989).



a) Current Web



b) Semantic Web

## World Wide Web vs. Semantisches Web (Koivunen und Miller, 2001).

# Semantisches Web: Erweiterungen

- strukturiert bestehende Webressourcen
- für Maschinen *und* Menschen verstehbar (Erfassen von Bedeutung!)
- Erweiterung des derzeitigen Webs
- dezentral
  - Recommendersysteme
  - Web of Trust
  - Communities können eigene (auch widersprüchliche!) Ontologien definieren

# Ontologien/Definition

---

- “Ontology”: Philosophie [Metaphysik IV, 1; Aristoteles]  
Wesen der Dinge; Welche Eigenschaften haben alle Dinge gemeinsam?
- “onotology”: System von Kategorien, die eine bestimmte Welt - unabhängig von der jeweiligen Sprache - beschreiben  
(Guarino, 1998)

**versprechen:** “shared and common understanding” einer Domaine  
- für Menschen und Applikationen kommunizierbar

# Ontologien/Syntax vs. Semantik

- **Syntax:** Struktur der Daten  
XML (Start/Endtag)
- **Semantik:** Bedeutung der Daten

Interoperabilität bedingt

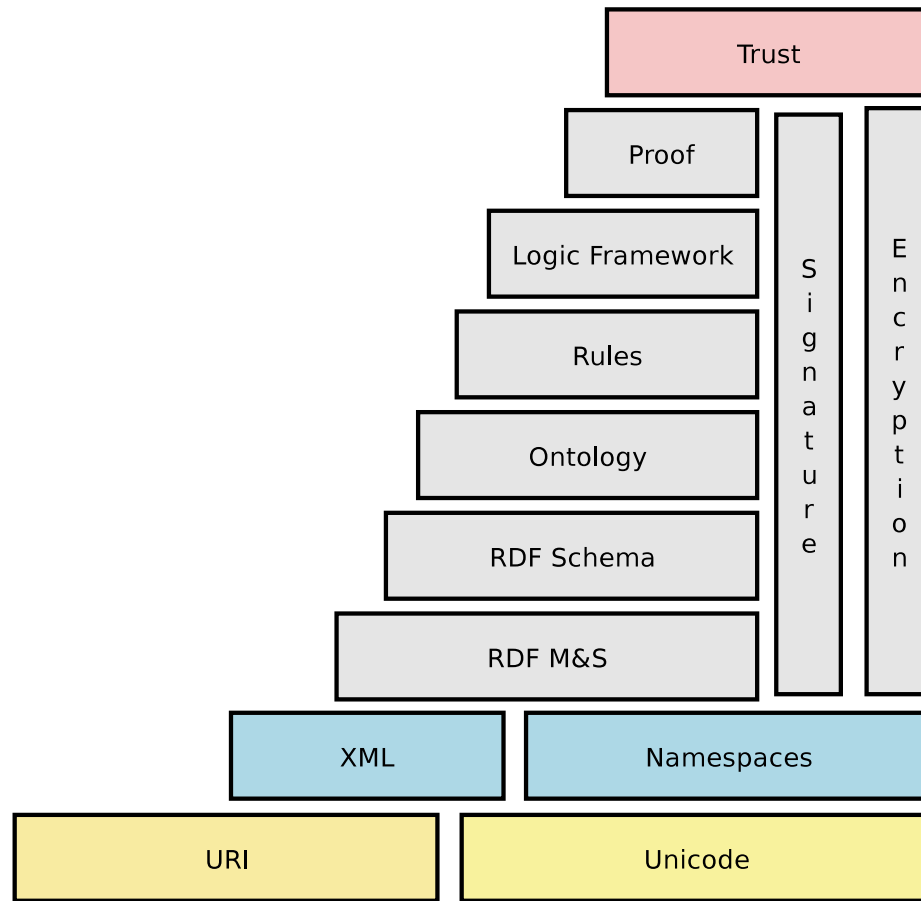
- gemeinsame Struktur zum Parsen der Daten  
(XML → DOM, SAX)
- die Möglichkeit die Daten zu verstehen (OWL)

Beispiel (Costello und Jacobs, 2003):

```
<SLR> . . . </SLR>
```

# Ontologien/Syntax vs. Semantik

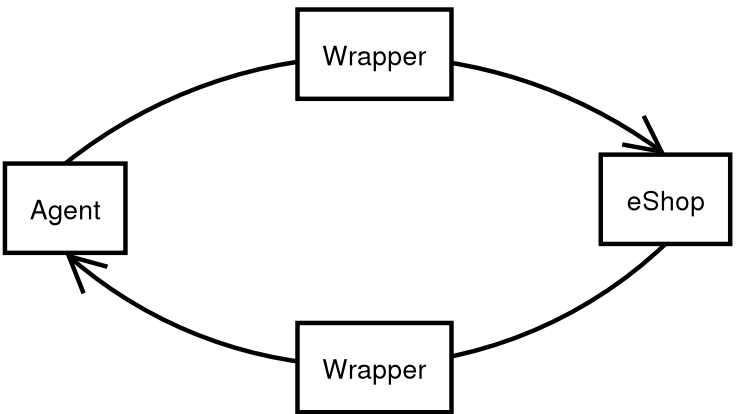
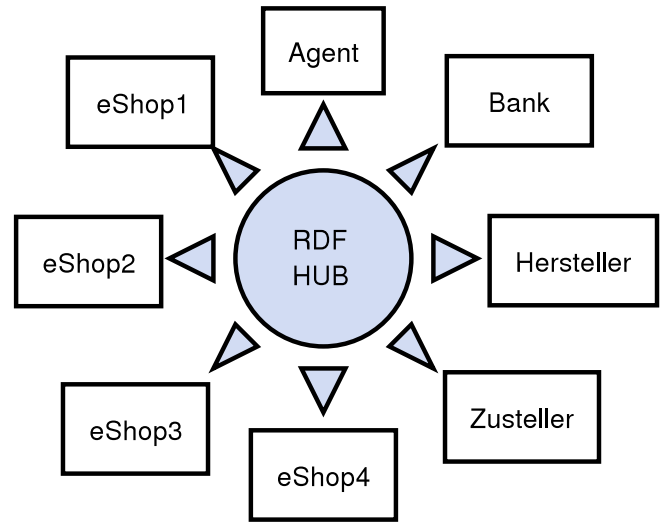
SLR Single Lens Reflex (camera)  
SLR Satellite Laser Ranging  
SLR Sending Loudness Rating (telecommunications)  
SLR Service Level Report  
SLR Side Looking Radar  
SLR Single Linear Recording  
SLR Slide Raft (aircraft door)  
...  
SLR Sri Lanka Rupee (national currency)  
SLR Statutory Liquidity Ratio  
SLR Stock Level Report  
SLR Stock Level Requirement  
SLR Straight Leg Raise  
SLR System Level Requirement(s)



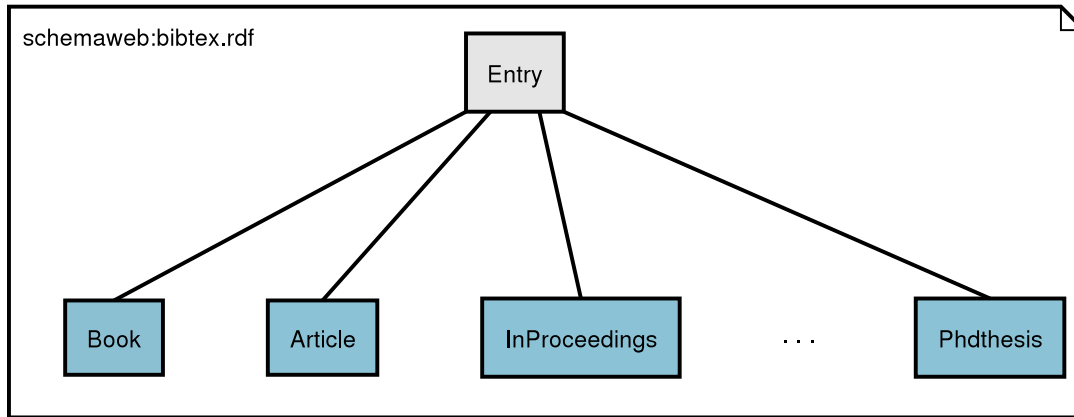
## Semantic Web enabling Technologies (Berners-Lee, 2002)



Diese Information ist nicht verfügbar  
 - verfügbar sind: Article\_Price, Article\_Tax, Article\_DeliveryTime



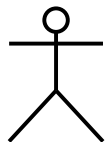
# Wrapper versus Ontologien



```

...
<owl:imports http://schemaweb:bibtex.rdf" />
...
<rdf:Property rdf:ID="hasJournal">
  <rdfs:domain rdf:resource="#Article">
  <rdfs:range rdf:resource="#Journal">
</rdf:Property>
...

```

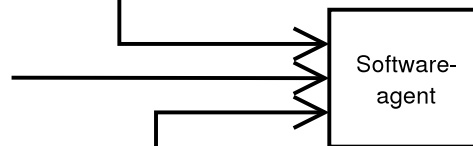


"In welchen Journalen haben Mitarbeiter des Institutes für Angewandte Informatik Artikel publiziert?"

```

...
<rdf:Description resource="wwwai:art02_2003.pdf">
  <hasJournal resource="scd:statPlanInference" />
</rdf:Description>
...

```

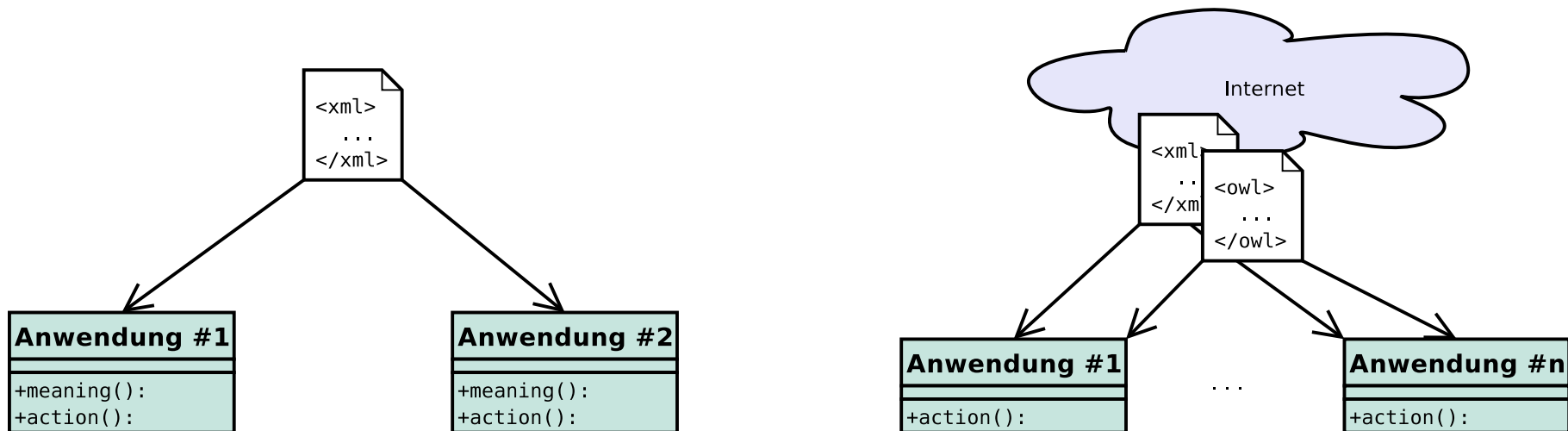


Journal of Statistical Planning and Inference

# Deduktion mittels Ontologien

# Ontologien/Software Engineering

- Code für die Interpretation von Daten
  - in die Applikation integriert
  - über OWL Beschreibungen realisiert



# Ontologie/Internet

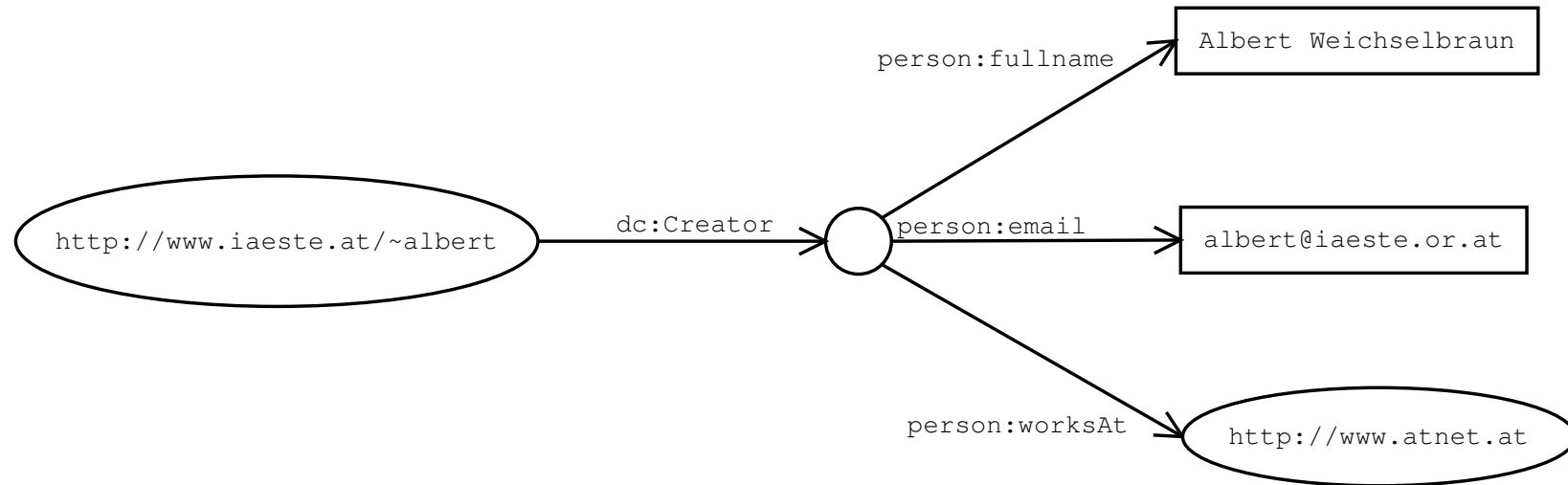
---

- Web als *eine* Datenbasis nutzbar
- *Autor* der Daten versieht diese mit entsprechenden Beschreibungen der Semantik  
→ ersetzt Information Retrieval
- weitere Schlüsse durch Deduktion erzielbar  
(RDF-Schema vs. XML-Schema) (einmal getroffene Aussage für alle Schlüsse weiter verwendbar)

```
<owl:sameAS rdf:resource="aw:price" />
```

# Ontologien/Technologie/RDF

---



RDF Aussagen als Graphen

# Ontologien/Technologie/RDF+XML

```
<rdf:Description
  rdf:resource="http://www.iaeste.at/~albert">
  <dc:Creator>
    <person:fullname>
      Albert Weichselbraun
    </person:fullname>
    <person:email>
      albert@iaeste.or.at
    </person:email>
    <person:worksAt resource="http://www.atnet.at" />
  </dc:Creator>
</rdf:Description>
```

Serialisierung von RDF als XML (Lassila und Swick, 1999)

# Ontologien/Technologie/Tripels

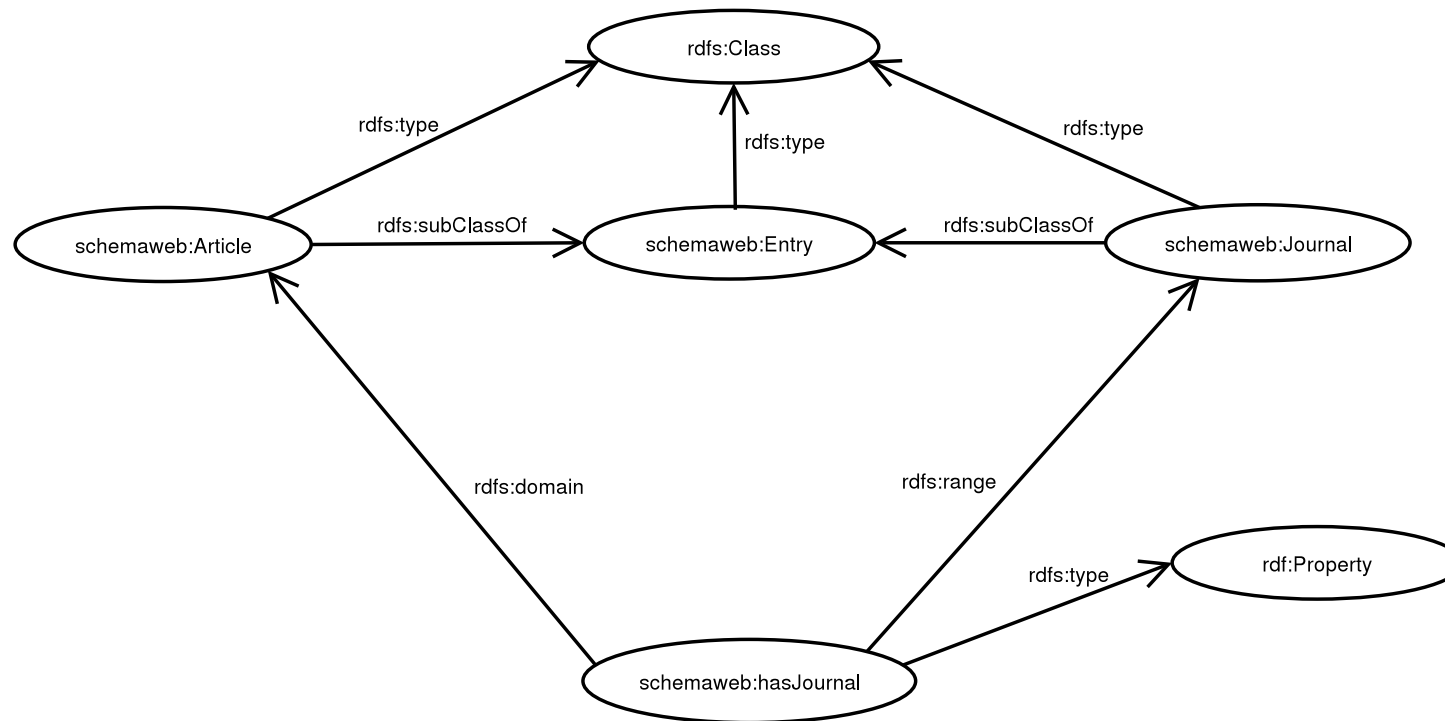
Vorteile:

- einfache Darstellung
- Abbildung in Datenbanken

```
<http://www.iaeste.at/~albert> dc:Creator  
  _:genid01.
```

```
  _:genid01 person:fullname "Albert Weichselbraun".  
  _:genid01 person:email    "albert@atnet.at".  
  _:genid01 person:worksAt  <http://www.atnet.at>.
```

Serialisierung von RDF in Form von N-Tripel (Grant und Beckett, 2003)



## RDF Schema: Ausschnitt des RDF Schemas der bibTeX-Ontologie

# XML Schema vs. RDF Schema

- XML Schema

- spezifiziert die Struktur von Dokumenten (Syntax)
- definiert *zulässige* Datentypen
- das Dokument *validiert* gegen das Schema

- RDF Schema

- beschreibt die Eigenschaften von Dokumenten (Semantik)
- Zielsetzung: Ermöglichen von zusätzlichen Aussagen durch Deduktion

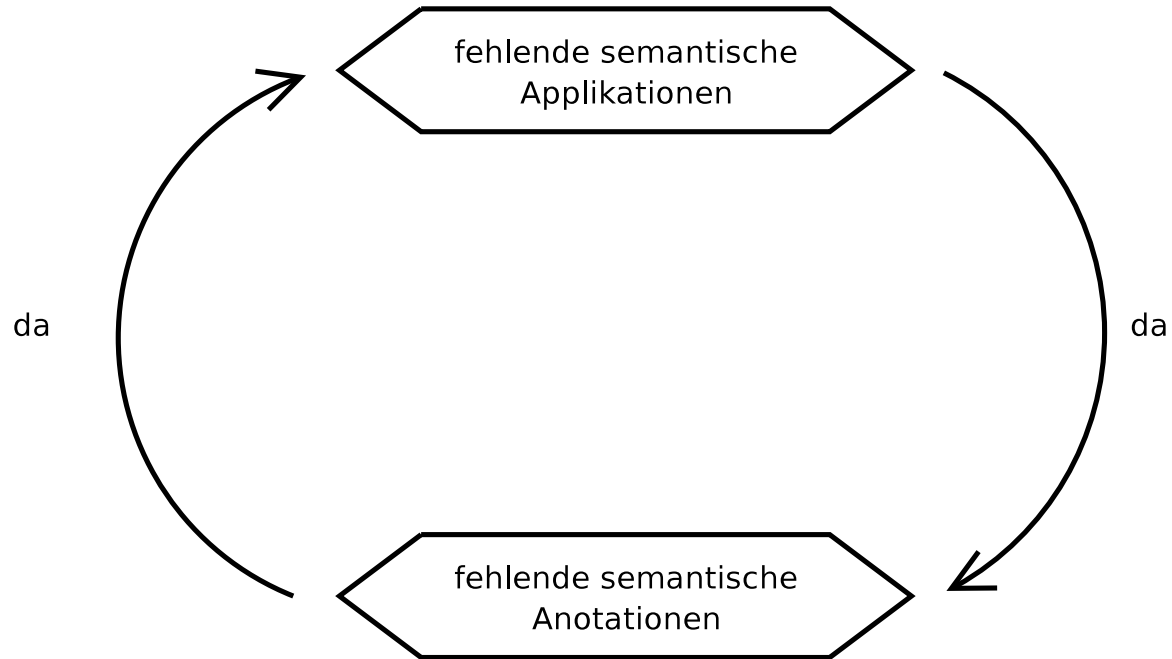
# Ontologien/Technologien/OWL

Erweiterungen von RDF Schema:

- Relationen zwischen Klassen
- Kardinalitäten
- Gleichheit
- Charakteristiken von Eigenschaften
- enumerierte Klassen

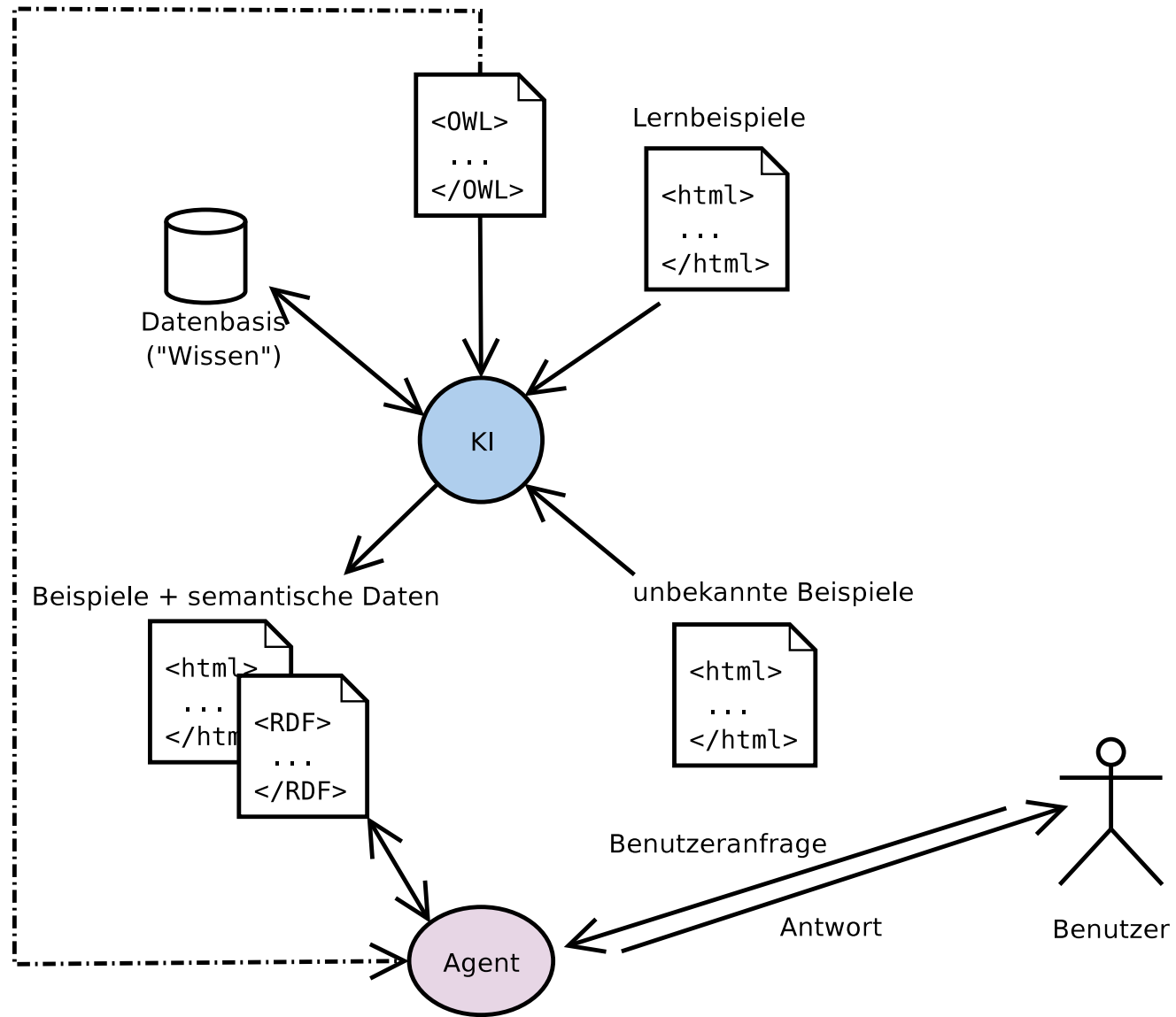
unterschiedlich komplex: OWL Lite - OWL DL - OWL Full

# Problemstellung/Semantic Web



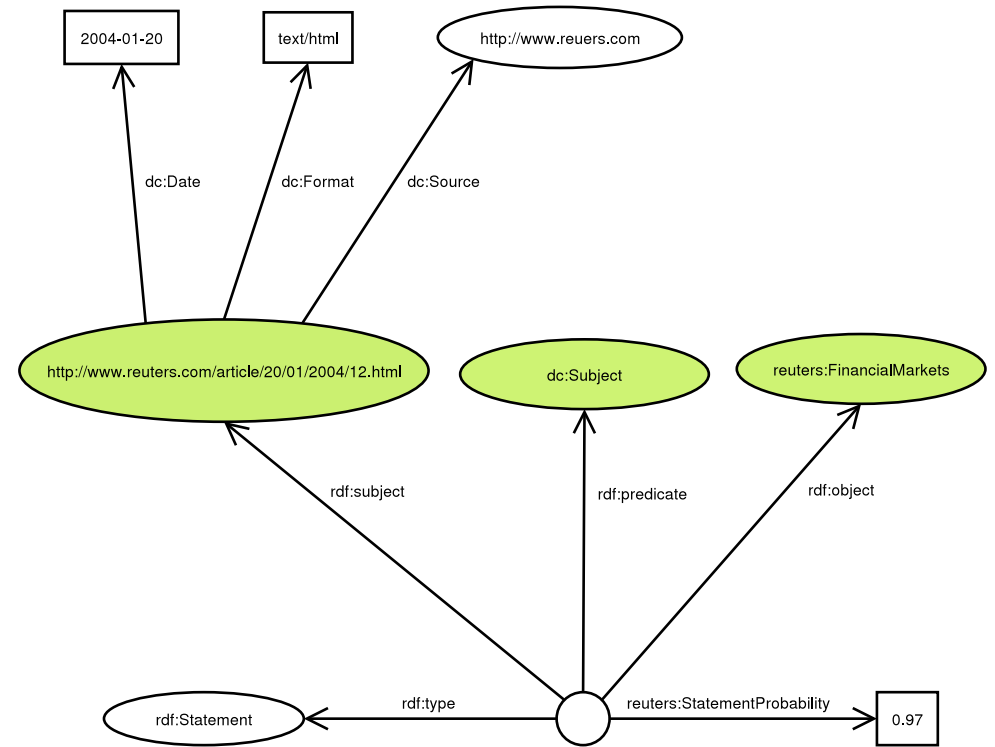
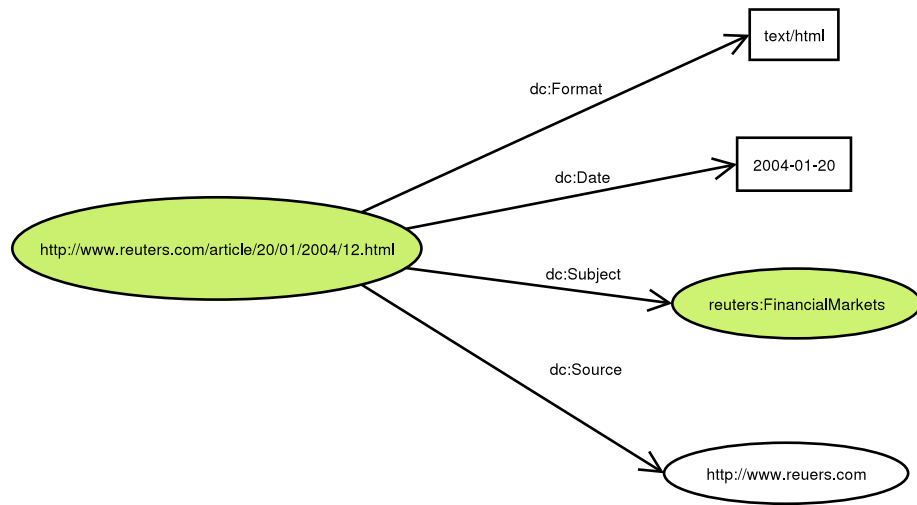
# Problemstellung/Dissertationsprojekt

- vollautomatische Klassifikation von Ressourcen
- Integration von Ontologien
- lernfähiges System
- Datenkennzeichnung
- strukturierte Ablage der gewonnenen Informationen im RDF-Format



# Problemstellung/Statistik

- Klassifikation erfolgt mit einer bestimmten Wahrscheinlichkeit
- muss in die Aussage über die Ressource integriert werden ⇒ Reification
- erlaubt Communities Aussagen über Ressourcen zu treffen und anhand dieser Aussagen zu handeln. (Trust, für Suchmaschinen, Einkäufe, etc...)
- notwendig, da verschiedene Communities Dinge anders bewerten!



# Probleme

---

- Kollision von Ontologien - Beheben von Widersprüchen
  - unterschiedliche Metadaten zu einer Ressource → wem glaubt man?
  - RDF/OWL-Poisoning
- Technologieebene:
  - URI/URL noch nicht Unicode fähig
  - Widersprüche im Klassenmechanismus von RDF-Schema (Russell's paradox)
  - endliche Bearbeitungszeit
  - skaliert diese Technologie?

\*

---

## Literatur

[Berners-Lee, 1989] Tim Berners-Lee. *Information Management: A Proposal*, March 1989. URL <http://www.w3.org/History/1989/proposal.html>.

[Berners-Lee, 2002] Tim Berners-Lee. *The Semantic Web*, 2002. URL <http://www.w3.org/2002/Talks/04-sweb/>.

[Costello und Jacobs, 2003] Roger L. Costello und David B. Jacobs. *OWL Web Ontology Language*, 2003. URL <http://www.xfront.com/owl/>.

[Grant und Beckett, 2003] Jan Grant und Dave Beckett. *RDF Test*

Cases, 15. Dezember 2003. URL <http://www.w3.org/TR/2003/PR-rdf-testcases-20031215/>.

[Guarino, 1998] Nicola Guarino. Formal ontology and information systems. In *Proceedings of FOIS'98*, Seiten 3–15, Italy, June 1998.

[Koivunen und Miller, 2001] Marja-Riitta Koivunen und Eric Miller. *W3C Semantic Web Activity*, November 2001. URL <http://www.w3.org/2001/Talks/04-sweb/>.

[Lassila und Swick, 1999] Ora Lassila und Ralph R. Swick. *Resource Description Framework (RDF) - Model and Syntax Specification*, 22 February 1999. URL <http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/>.