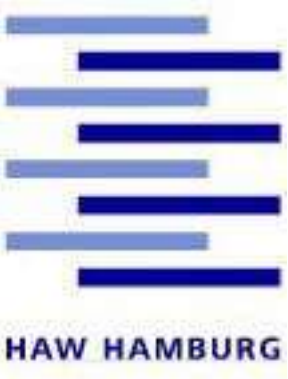
SS2005
Vortrag im Rahmen von Anwendungen 1:

Semantic Web: Enrichment und Search

Vortragender:
Gerrit Diederichs



Ziele

HAW HAMBURG

Was hatten wir schon:

- Überblick, Visionen, SWS (Piotr)
- Ontologien und Werkzeuge (Artem)

Mein Beitrag:

- Problem des Information Overkill
- Enrichment von Internetressourcen
- Suche basierend auf Semantic Web

Was kommt noch ?

- Transformationen (Thomas)

Gliederung

HAW HAMBURG

- **Motivation**
- Lösungsansätze
- Grundlagen (kurze Wiederholung)
- Enrichment
- Search
- Protégé 2000
- Projektszenario

Problem: Information Overkill

- Datenflut wächst täglich
- Google hat über 8 Milliarden indizierte Webseiten
- Maschinen „sehen“ darin nur eine Verlinkung von Ressourcen

→Suche nach bestimmten Ressourcen wird durch diesen „Data Smog“ immer ineffektiver

Heutige Suche im Web

HAW HAMBURG

- Schlagwort basierte Volltextsuche
- Verbesserung durch den Einsatz komplexer „Ranking“ Funktionen (Google PageRank)

Probleme:

- Nicht Einbeziehung von Synonymen
- Ignoranz von Mehrdeutigkeiten (Homonymen)
- Ignoranz von Wortformvariationen
- Nichterkennung sinnverwandter Begriffe

Aus [WLEKLI03]

Beispiel: Synonyme

HAW HAMBURG

- Google Suche
  - Begriff „Waldwirtschaft“
    → 85.700 Treffer
  - Synonym „Forstwirtschaft“
    → 2.060.000 Treffer

Unterschied Faktor 24 !

Beispiel: Homonyme

- Google Suche
  - Begriff „Java“
    - 210.000.000 Treffer
  - Begriff „Java + Urlaub“
    - 1.150.000 Treffer

Unterschied Faktor 182 !

Es gibt weitere Beispiele...

Gliederung

HAW HAMBURG

- Motivation
- **Lösungsansätze**
- Grundlagen (kurze Wiederholung)
- Enrichment
- Search
- Protégé 2000
- Projektszenario

Wie können wir finden was wir suchen ?

HAW HAMBURG

Idee:

Hinterlegung maschinenlesbarer, semantischer Information

Ansätze:

- Syntaktische Anreicherung der Suchanfrage (OntoSeek, Dipl.Arbeit A.Christensen)
- Semantische, maschinenlesbare Anreicherung von Webressourcen basierend auf Ontologien (Semantic Web)

Ansatz 1: OntoSeek

- Projekt des National Research Council, Landseb-CNR u.a.
- Inhaltsbasierte Suche in Produktkatalogen und Yellow Pages
- Anfragen werden durch in Ontologien spezifiziertem Wissen analysiert (Wortverwandschaften etc.)
- Anfrage wird mittels Ersetzungen präzisiert

Aus [CHRIST05]

Ansatz 1: Dipl.Arbeit A.Christensen

- Verbesserung der Websuche konventioneller Suchmaschinen
- Aufbau von Domänenwissen mittels Topic Maps
- Eingehende Anfragen werden hinsichtlich bekannter Topics überprüft
- Topic vorhanden → Anfrage verfeinern

Ansatz 1: Fazit

Vorteile:

- Nutzung bestehender Suchmaschinen möglich

Nachteile:

- Queries werden u.U. sehr komplex
- Relativ schwache Semantik

Es geht noch besser...

Ansatz 2: Semantic Web

- Modellierung von Wissen in Ontologien
- Population der Ontologien durch Annotation von Internetressourcen
  - Manuell
    - Webmasterprinzip
    - Community (Annotation Server, SHOE)
  - Automatisch

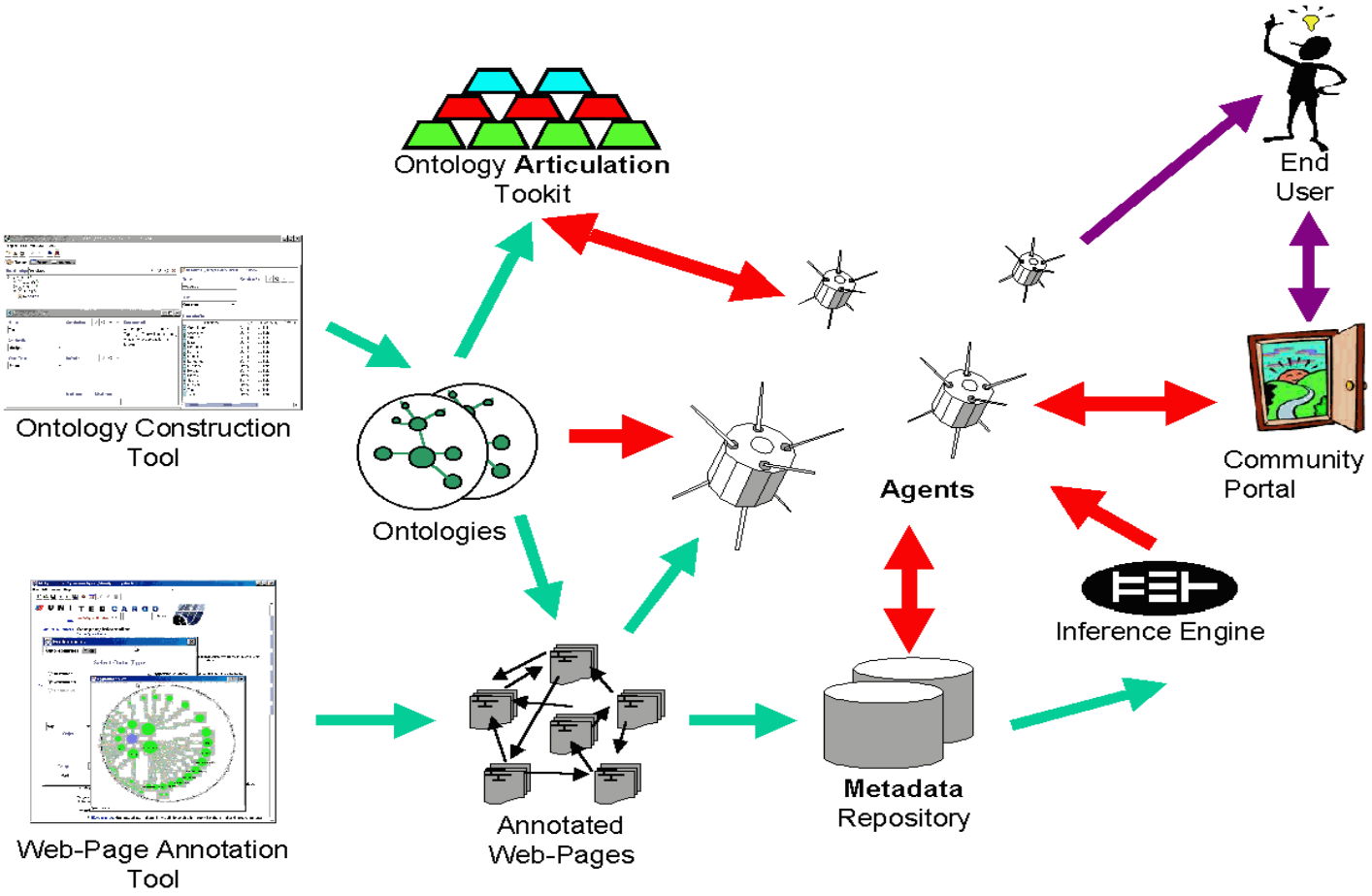
→Das Web als „globale DB“ (Berners-Lee)

Gliederung

HAW HAMBURG

- Motivation
- Lösungsansätze
- **Grundlagen (kurze Wiederholung)**
- Enrichment
- Search
- Protégé 2000
- Projektszenario

The Big Picture



RDF

- Metadatenmodell für Internetressourcen
- Basis sind Aussagen über Ressourcen (Subjekte)
- Aussagen sind aufgebaut als

  Subjekt-Prädikat-Objekt Triple
- Triples bestehen meist aus URIs

RDF: Ein Beispiel

HAW HAMBURG

Aussage:

„Der Autor von http://dietlweiss.de/ ist Tobias Dietl“

RDF Statement in N-Triples Notation:

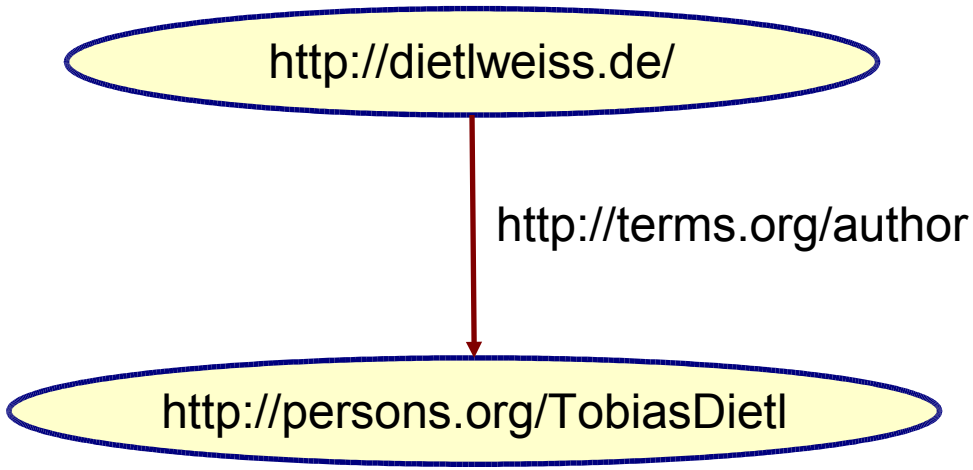
<http://dietlweiss.de/> ← subject
<http://terms.org/author> ← predicate
<http://persons.org/TobiasDietl> ← object

Bedeutung: http://dietlweiss.de/ hat den Autor Tobias Dietl

Aus [DIETL02]

RDF Notationen: Gerichteter Graph

RDF modelliert Statements mit Knoten und Pfeilen:



Aus [DIETL02]

RDF Notationen: RDF/XML

Offizielle RDF/XML Notation der
gleichen Aussage:

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xmlns:terms="http://terms.org/">
<rdf:Description rdf:about="http://dietlweiss.de/">
    <terms:author rdf:resource="http://persons.org/TobiasDietl" />
</rdf:Description>
</rdf:RDF>
```

Aus [DIETL02]

Ontologiesprachen

Aufgaben:

- Semantische Modellierung der durch RDF beschriebenen Aussagen
- Mapping von Ontologien
- Bestehen aus Klassen, deren Eigenschaften und Relationen
- Instanz wird über <rdf:type> erzeugt
- Quasi Standards sind **RDFS** und **OWL**
- Dabei gilt:

RDFS < OWL Lite < OWL DL < OWL Full

„<“ = syntaktisch und semantisch enthalten

Fazit

- RDF Triples

  →Instanzen eines Wissensmodells
- RDFS/OWL

  →Modellierung des Wissensmodells

→Technische Grundlage für (maschinenverwertbare) Semantik

→Formale Grundlage für logische Inferenz

Gliederung

- Motivation
- Lösungsansätze
- Grundlagen (kurze Wiederholung)
- **Enrichment**
- Search
- Protégé 2000
- Projektszenario

Enrichment in Knowledge Bases

- Möglichkeiten zur Annotierung von Ressourcen
  - Manuell einpflegen
  - Automatisiert einpflegen

Manuelle Klassifikation

- Experten erstellen Ontologie
- Experten erstellen Instanzen der Ontologie

→Bei größeren Datenmengen unbrauchbar

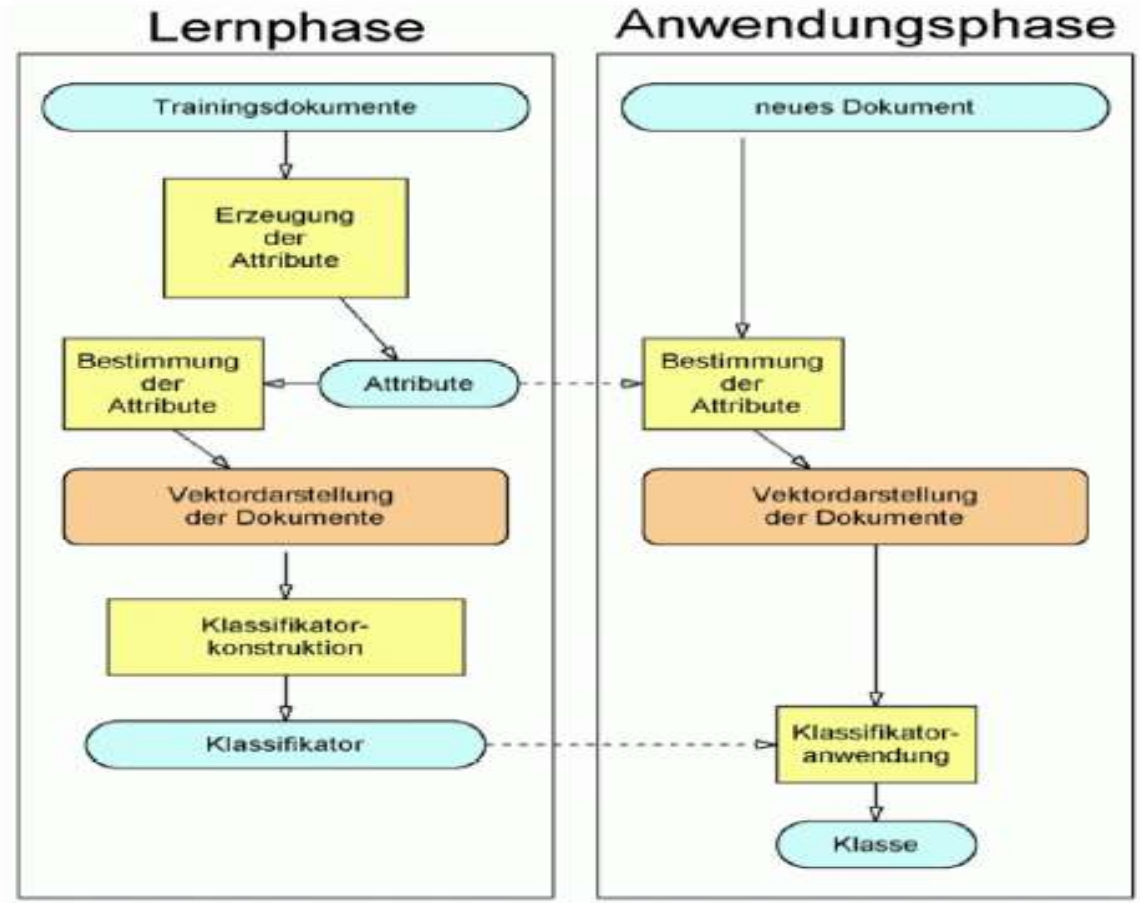
Automatische Klassifikation

HAW HAMBURG

- Lernphase
  - Erzeugung eines Sets von Trainingsdaten
  - Extraktion bestimmter Attribute
  - Erstellung eines Basismodells
- Anwendungsphase
  - Aufnahme neuer Dokumente
  - Extraktion der in der Lernphase identifizierten Attribute
  - Vergleich und Einordnung anhand des Klassifikationsmodells
  - Gegebenenfalls Erweiterung des Basismodells

Aus [HOFFMA02]

Automatische Klassifikation (2)



Aus [HOFFMA02]

Automatische Klassifikation (3)

HAW HAMBURG

- Identifizierung der Attribute durch Textanalyse
- Drei Verfahren werden unterschieden
  - Linguistische Analyse
  - Statistische Analyse
  - Begriffsorientierte Verfahren

Aus [HOFFMA02]

Automatische Klassifikation (4)

HAW HAMBURG

Linguistische Analyse

- Entfernung nicht sinntragender Wörter
  - Wörterbuchbasiert
  - regelbasiert
- Syntaktische Analyse auf Satzebene
- Semantische Analyse auf Dokumentebene

→Rein linguistische Verfahren bei natürlicher Sprache zu aufwendig

Aus [HOFFMA02]

Automatische Klassifikation (5)

HAW HAMBURG

Statistische Analyse

- Vorkommenshäufigkeit von Wörtern
- 5 Phasen in der Lernphase
  - Textnormalisierung
  - Termgenerierung
  - Attributauswahl
  - Attributgewichtung
  - Lernschritt

Aus [HOFFMA02]

Automatische Klassifikation (6)

Begriffsorientierte Verfahren

- Orientiert sich am menschlichen Klassifikationsverhalten
- Aufbau von Thesauren oder Wörterbüchern

Aus [HOFFMA02]

Fazit

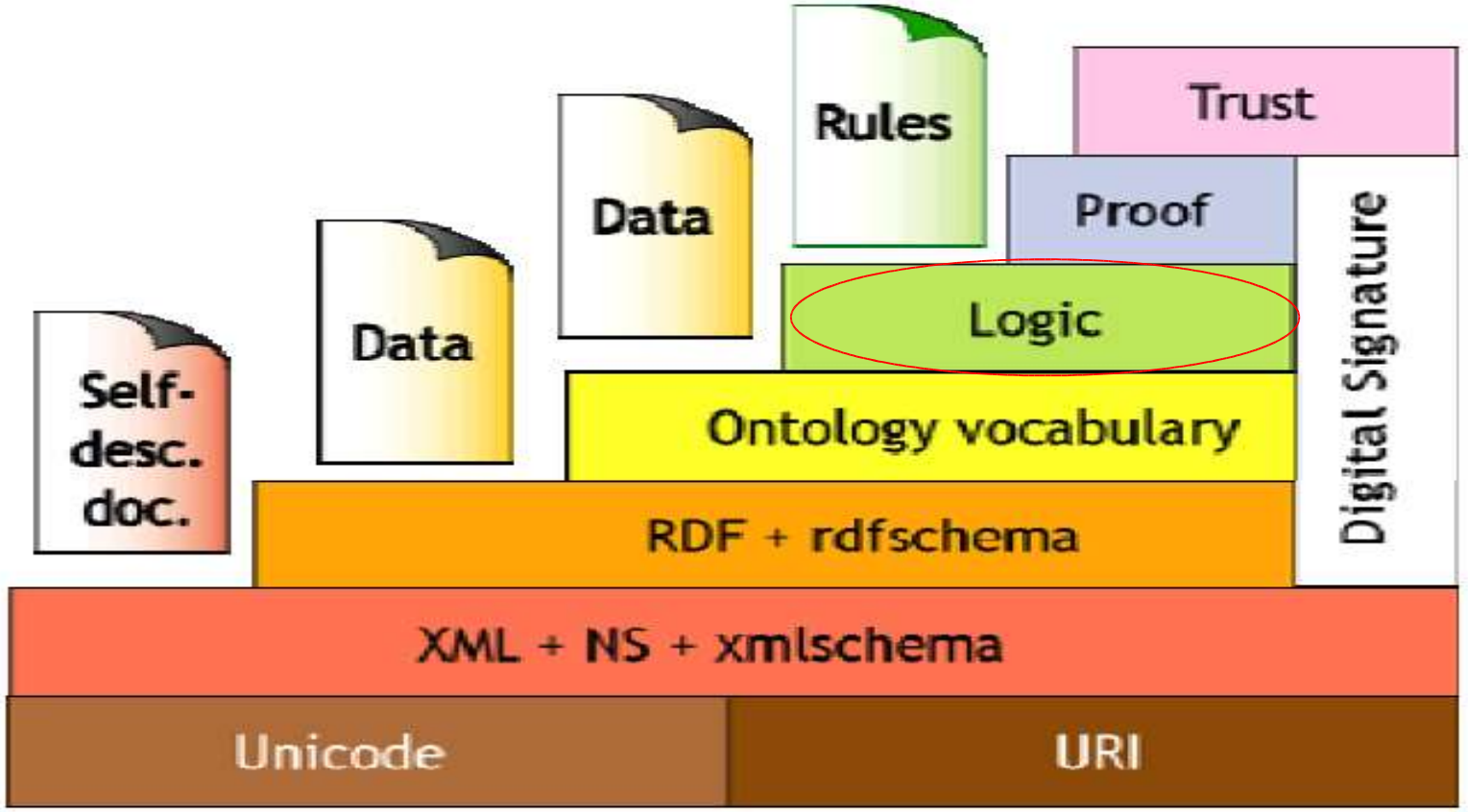
- Manuelle Klassifikation bei überschaubaren Datenmengen
- Automatische Klassifikation bei großen Datenmengen (z.B. Webmining)
  - Häufig Erstellung von Anfangstaxonomien durch Experten
  - Beispiel für Umsetzung einer automatischen Klassifikation in großem Stil → Web Fountain
  - i.d.R. sehr aufwendig bezüglich Ressourcen und Klassifikation

→Für Ferienclub Szenario reicht manuelle Klassifikation

Gliederung

- Motivation
- Lösungsansätze
- Grundlagen (kurze Wiederholung)
- Enrichment
- **Search**
- Protégé 2000
- Projektszenario

Der Semantic Web Stack



Quelle: Berners-Lee (1999)

Suche in OWL Modellen

- OWL Modelle bieten Inferenzmöglichkeiten
  → neues/nicht explizit modelliertes Wissen wird generiert
- Wissenserschließung durch Inferenzmaschine
- „Mächtigkeiten“ von Inferenzmaschinen
  - Higher Order Logic
  - Full First Order Logic (Prädikatenlogik)

  } Nicht entscheidbar
  - Description Logic
  - Logic Programming
- Generiertes Wissen als „virtuelle“ Triples
- Abfrage über RDF Queries

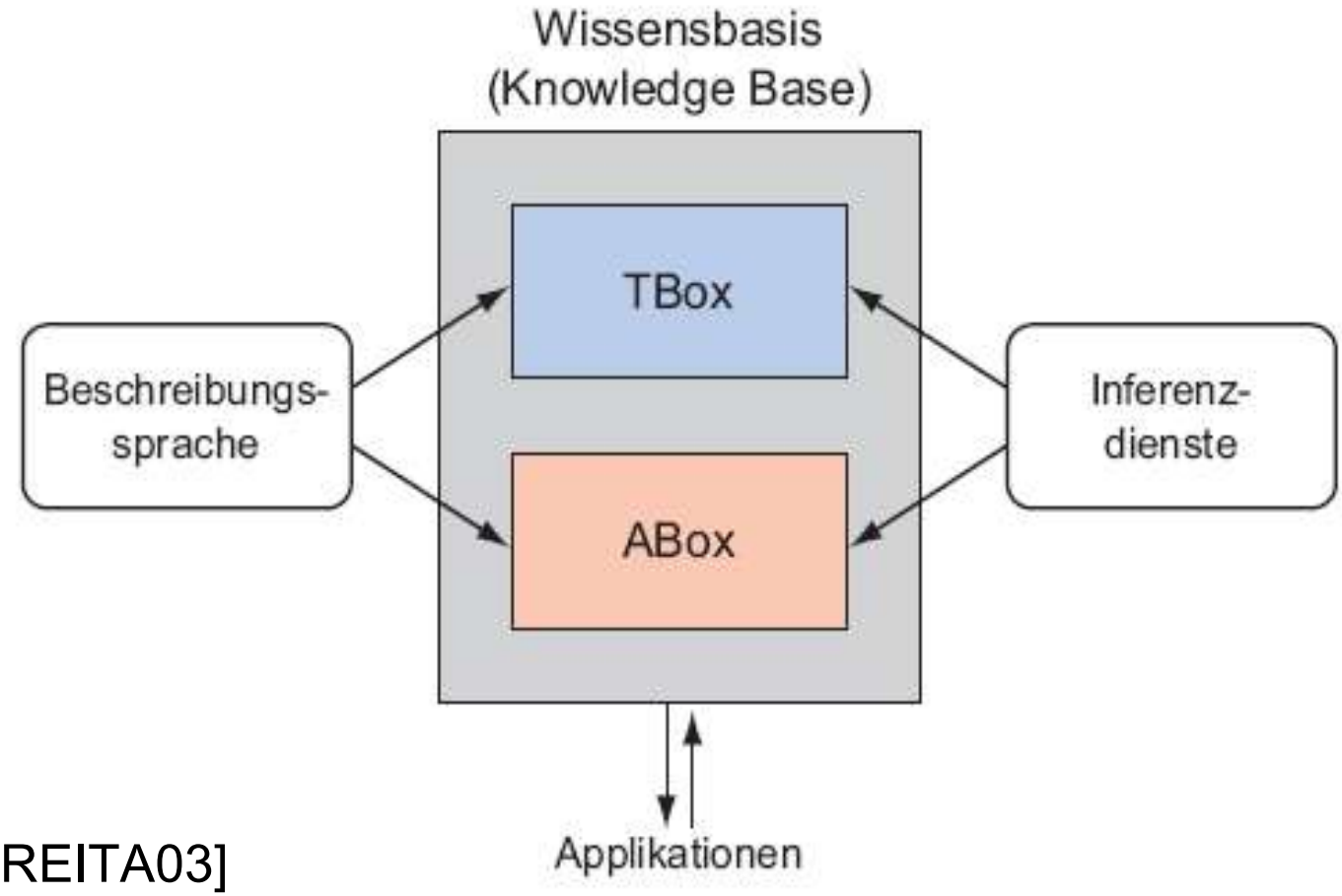
Inferenz (Reasoning)

Aufgaben:

- Konsistenz gewährleisten
- Klassifikation
- Äquivalenzen ermitteln
- Abgeleitete Bedingungen ermitteln → neues Wissen

Aus [FREITA03]

Description Logic



Aus [FREITA03]

Description Logic

HAW HAMBURG

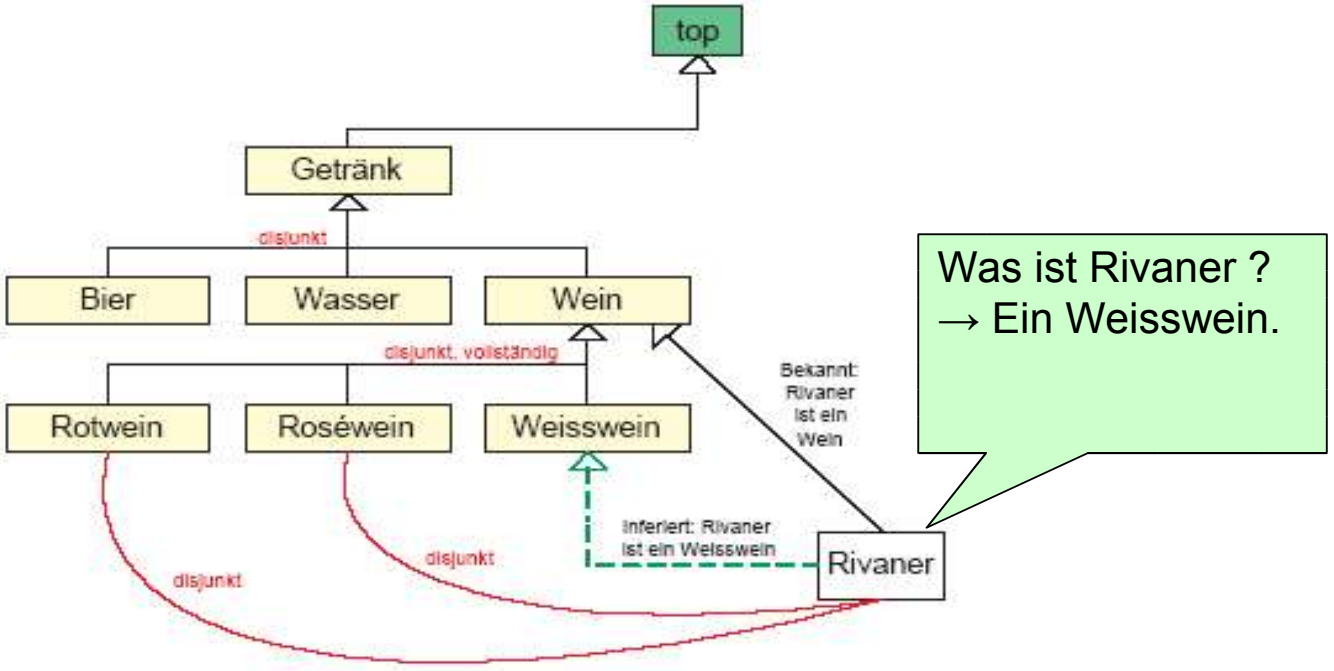
- Untermenge der Prädikatenlogik
- Prädikate: Konzept (Klasse), Rolle (Beziehung)
  →Beschreibt eine Klassenstruktur
- A-Box
  - Instanzen („Reale Welt“)
- T-Box
  - Modellierung der Ontologie (Konzepte, Rollen)
- Keine Variablen in Syntax

Description Logic: Operationen

1. $A \sqcup B$ *ist die Vereinigung der Mengen von* $A^{\mathcal{I}}$ *und* $B^{\mathcal{I}}$ *(entspricht „oder“)*
2. $A \sqcap B$ *ist der Schnitt der Mengen von* $A^{\mathcal{I}}$ *und* $B^{\mathcal{I}}$ *(entspricht „und“)*
3. $\neg A$ *ist die Grundmenge* $\Delta^{\mathcal{I}}$ *ohne* $A^{\mathcal{I}}$
4. $\exists R.A$ *fordert, dass das beschriebene Objekt zu mindestens einem Objekt aus der Menge* $A^{\mathcal{I}}$ *in der Beziehung R steht*
5. $\forall R.A$ *fordert, dass alle Objekte mit denen das beschriebene Objekt in der Beziehung R steht aus der Menge* $A^{\mathcal{I}}$ *stammen*

Inferenz: Beispiel

Wenn bekannt ist, dass Rivaner ein Wein ist, aber kein Rotwein und kein Roséwein, kann aufgrund der Vollständigkeit der Zerlegung von Wein in seine Unterkonzepte geschlossen werden, dass Rivaner ein Weisswein ist.



Aus [FREITA03]

Inferenzen in DL

In einer DL gibt es grundsätzlich zwei Arten von Inferenzen

1. *Ist Untermenge (Subsumption): $C \sqsubseteq D$ d.h. $C^{\mathcal{I}} \subseteq D^{\mathcal{I}}$ für alle $\mathcal{I}$?*
2. *Konsistenz (zu Tbox $\mathcal{T}$): Gibt es ein Modell $\mathcal{I}$ von $\mathcal{T}$ mit $C^{\mathcal{I}} \neq \emptyset$?*

Die Komplexität der Inferenzen steigt mit jeder Erweiterung der DLs.

Aus [GÖTTLI02]

Query Sprachen: RDQL

HAW HAMBURG

- weit verbreitet Abfragesprache (u.a. Jena Framework)
- basiert auf einer SQL ähnlichen Syntax
- Berücksichtigt Triple Notation von RDF
- Elemente:
  - *Select clause*
  - *Frome clause*
  - *Where clause*
  - *And clause*
  - *Using clause*

Aus [SCHMUD04]

Elemente RDQL Query

- Select
  - Projektionsmenge
- From
  - Durchsuchte Modelle
- Where
  - Selektion
- And
  - Verschärfung der Selektion
- Using
  - Abkürzung für URI‘s

Beispiel RDQL Query

```
SELECT ?resource, ?familyName
    FROM <http://example.org/someModel>
    WHERE (?resource info:age ?age)
                (?resource vCard:N ?y)
              (?y <vCard:Family> ?familyName)
    AND ?age >= 24
    USING info FOR <http://somewhere/peopleInfo#>
    vCard FOR <http://www.w3.org/2001/vcard-rdf/3.0#>
```

Ergebnis:

```
resource                                | familyName
=======================================================
<http://somewhere/JohnSmith/> | "Smith"
```

Fazit

- Die auf Ontologien basierende Infrastruktur des Semantic Web bietet ein formales Wissensmodell
- Inferenzmaschinen können darauf aufbauend das dargestellte Wissen um implizite Schlussfolgerungen erweitern
- Dabei stellen sie eine konsistente und korrekte Wissensbasis sicher
- RDF basierte Abfragesprachen können auf dieses Wissen zugreifen

→Die durch die Semantik ermöglichte Logik bietet eine weitaus mächtigere Alternative als die vorhin vorgestellte Anreicherung der Syntax

Gliederung

- Motivation
- Lösungsansätze
- Grundlagen (kurze Wiederholung)
- Enrichment
- Search
- **Protégé 2000**
- Projektszenario

Protégé 2000

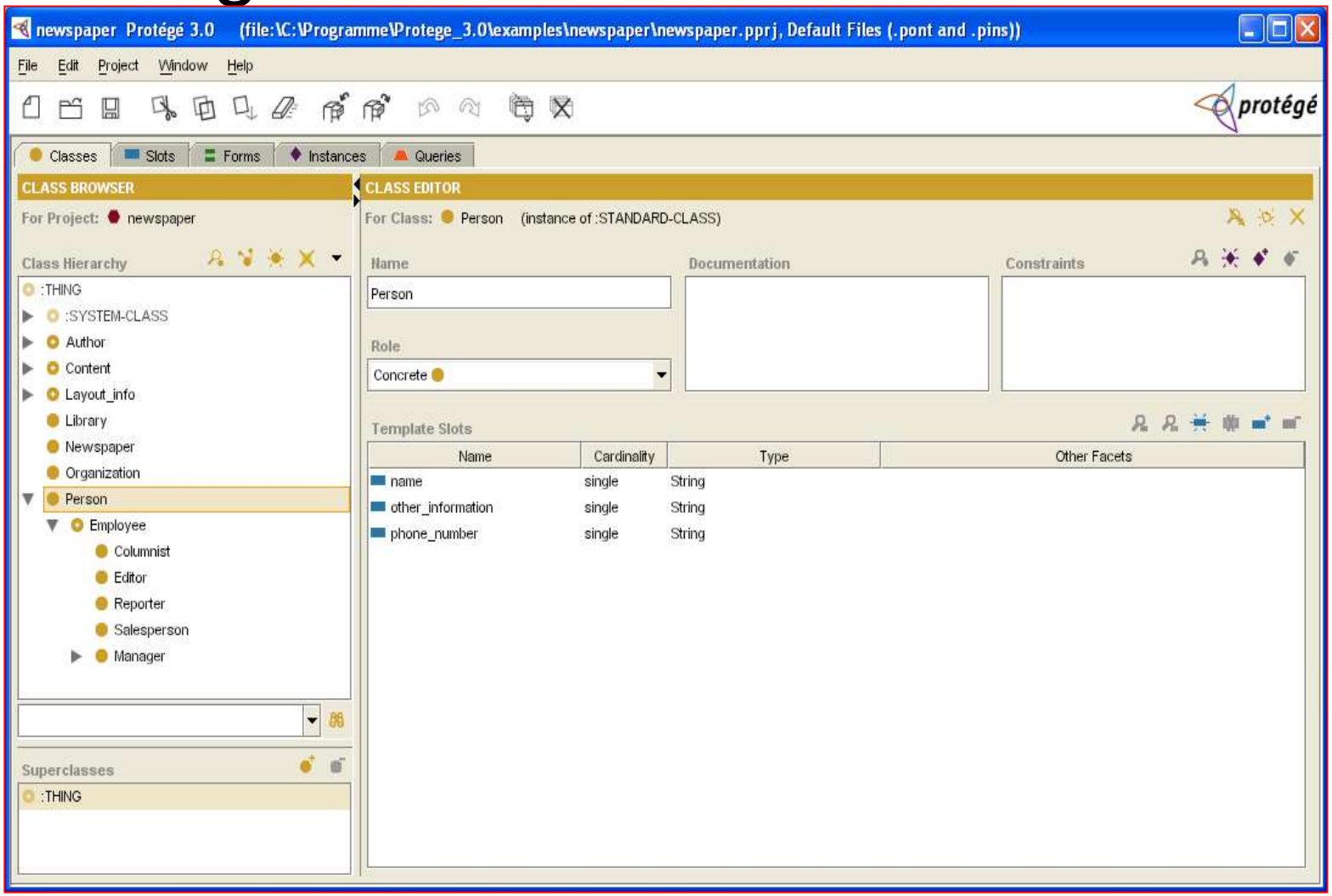
HAW HAMBURG

- Ontologieerstellungstool
  - Erstellung von Ontologien/Instanzen
  - Mapping von Ontologien
  - Erstellen von Queries
  - Plugin-Erweiterbarkeit
- Open Source (MPL)
- Java Anwendung
- Leicht benutzbare GUI
- Schnittstellen zu Inferenzmaschinen (RACER)

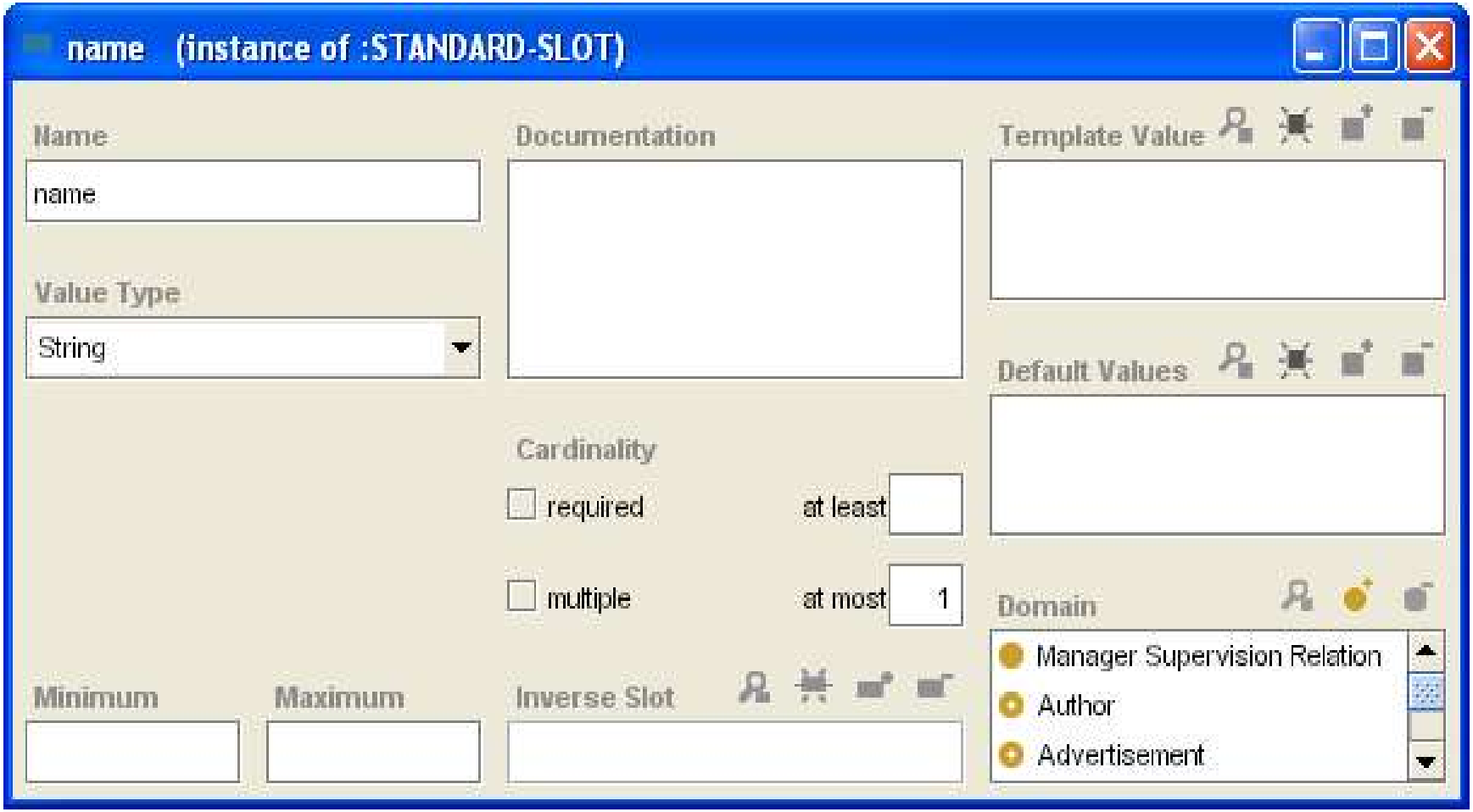
Aufbau

- Klassen
- Slots (Eigenschaften)
- Forms (auf Basis der Klassenbeschreibung)
- Instanzen
- Queries

Protégé: Klassenansicht

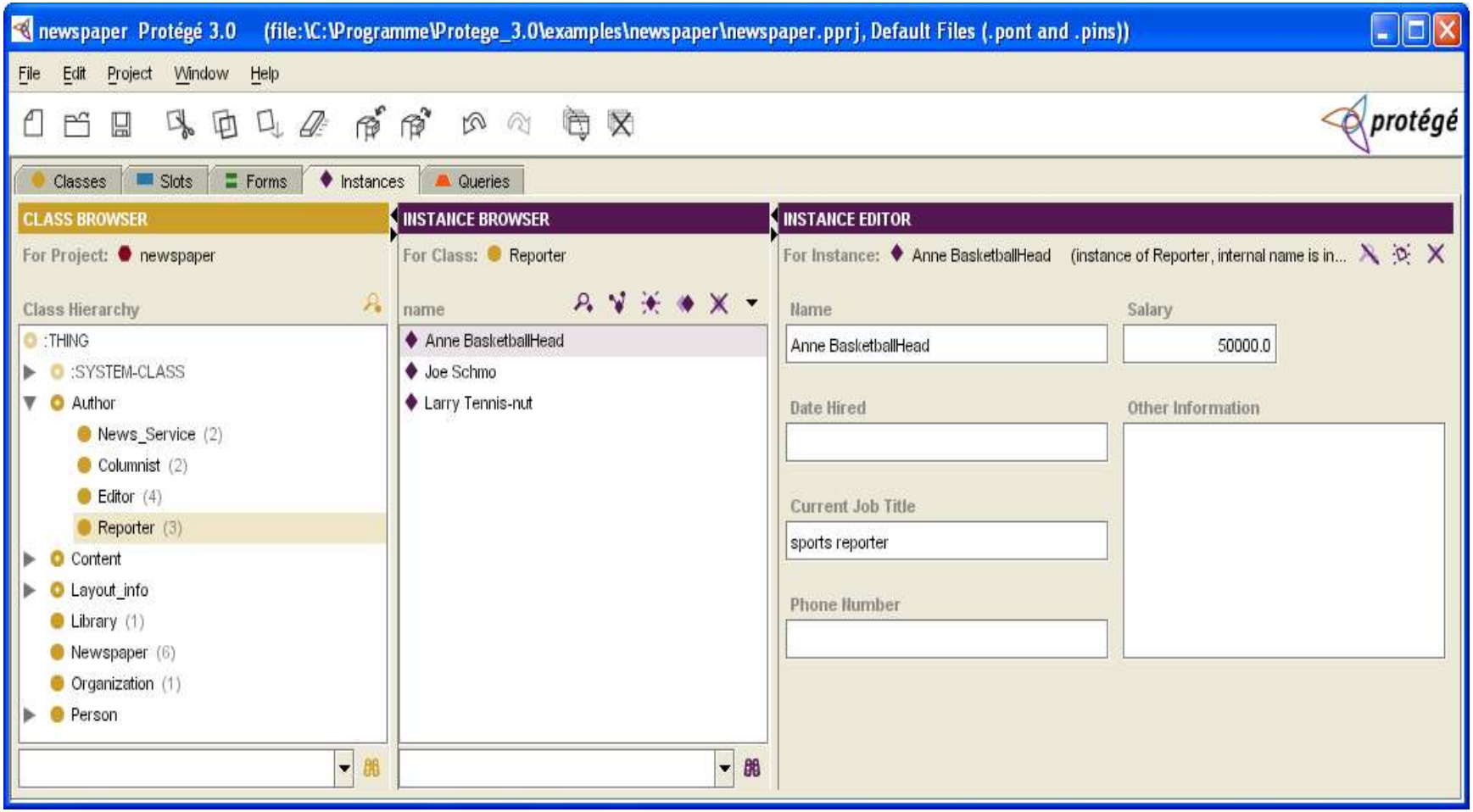


Protégé: Slot

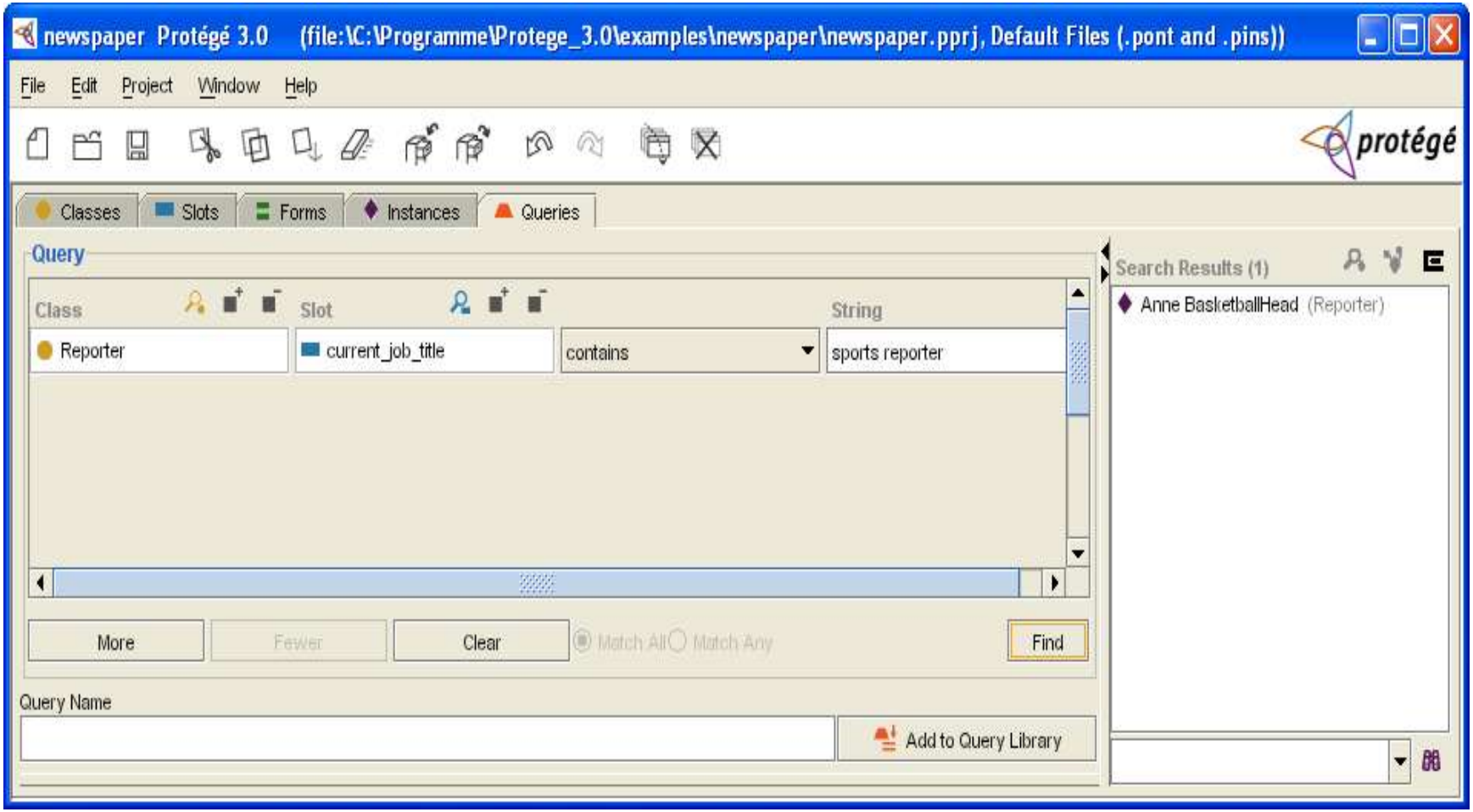


Protégé: Instanzen

HAW HAMBURG



Protégé: Query



Gliederung

- Motivation
- Lösungsansätze
- Grundlagen (kurze Wiederholung)
- Enrichment
- Search
- Protégé 2000
- **Projektszenario**

Informationsportal für den Ferienclub

HAW HAMBURG

Angebot:

- Aufbau eines Informationsportals für die Clubbesucher

ToDo:

- Auswahl von Semantic Web Tools für die Umsetzung
- Evaluierung bestehender Ontologien
- Ggf. Entwurf einer eigenen Ontologie
- Entwicklung einer benutzerfreundlichen Anfragesprache (easy RQL)
- Web Applikation „on-top“

Informationsportal für den Ferienclub

Zu klären:

- Welche Tools benutzen (Sem Web Gruppe)
- Was sind unsere „Top-Level“ Ontologien
  - Sind das schon bestehende
  - Eigenentwicklung
- Speicherung der Ontologien

Sinnvolle Ausbaustufe:

Personalisierte Agenten sammeln die relevanten Informationen für die Clubbesucher

Literatur

URLs:

[DIETL02]: http://www11.informatik.tu-muenchen.de /lehre /seminare /seminarSW-SS2002/extension/sprachen.ppt

[GÖTTLI02]: http://www11.informatik.tu-muenchen.de /lehre /seminare /seminarSW-SS2002/extension/logik1.ppt

[FREITA03]: http://www.im.uni-passau.de/lehre/ws0304/DLON/DLON.4in1.pdf

[HOFFMA02]: www.iicm.edu/thesis/rhoff/Hoffmann_DA.pdf

[SCHMUD04] : http://swt-www.informatik.uni-hamburg.de/publications/files/Dipl/Schmude_OntologiebasierteNavigation.pdf

www.semanticweb.org

http://www.w3.org/2001/sw/

Literatur

HAW HAMBURG

Sonstiges:

[CHRIST05]: Andreas Christensen

Diplomarbeit:

Eignung von Topic Maps zur Verbesserung von Suchanfragen am Beispiel der Studierenden an der HAW im Fachbereich Informatik

[WLEKLI03]: Fabian Wleklinski

Diplomarbeit:

Suche im Semantic Web

Bücher:

Stuckenschmidt, van Harmelen:

Information Sharing on the Semantic Web

ISBN: 3-540-20594-2

Fragen ?

Hat jemand die Zeit gestoppt ?