Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Schema Matching Seminar
WS 2007/08
Themen & Organisation

Prof. Felix Naumann,

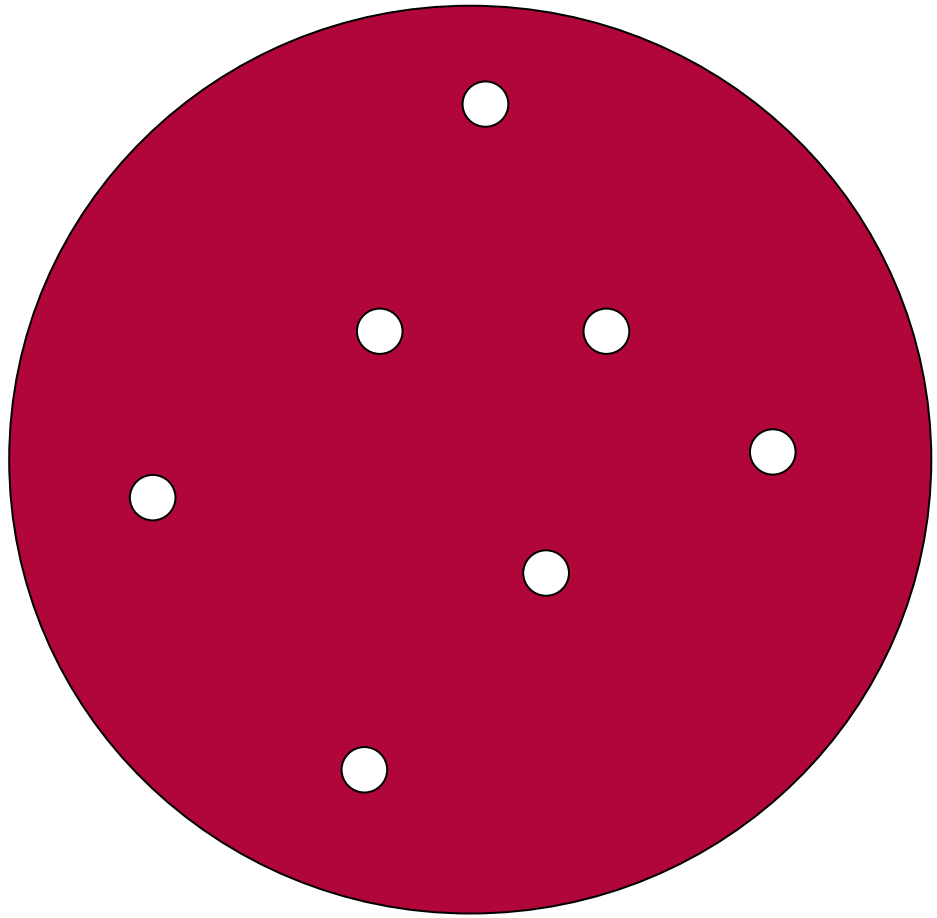
Alexander Albrecht, Frank Kaufer, Melanie Weis

Hasso-Plattner-Institut, Potsdam

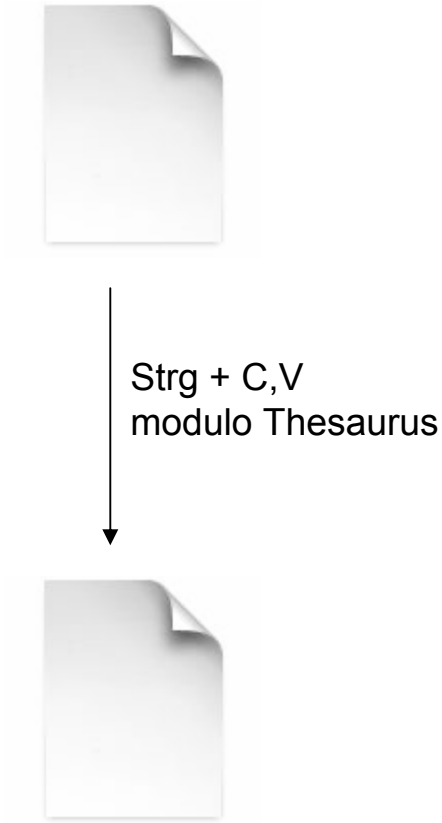
24. Oktober 2007

Seminarmodus - klassisch

- Großes Themengebiet
- 1-2 Artikel
- Seminarausarbeitung (20 Seiten in TeX)
- Vortrag



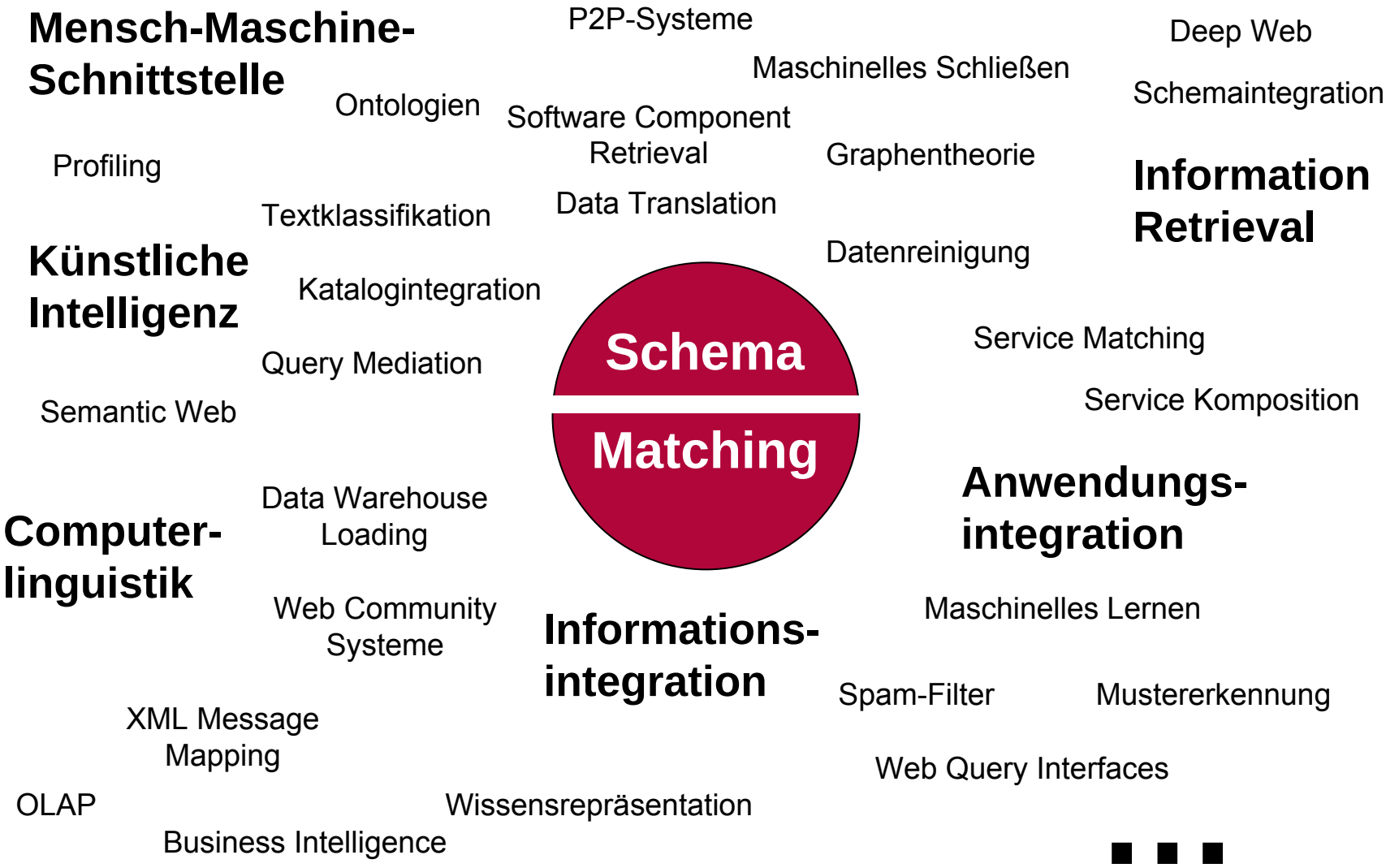
Seminarmodus - klassisch



Seminarmodus – FUBU Style

- Wissenschaftliche Kollaboration
  - Arbeitsgruppe, Forschungsprojekt
  - Forschungskolleg
- Gemeinsames Erarbeiten eines neuen, kohärenten Themengebiets
  - Theorie verstehen
  - Techniken herausarbeiten
  - Systeme evaluieren
- Ausarbeitung: Wiki (Wikipedia als Bonus)
- Talk im Tutorialstil
- Hands-on (6 CP)
  - Kleine Demoimplementationen
  - Testkollektionen erstellen (positive & negative Beispiele)
  - Evaluieren

Werde ich hier zum Fachidioten?



Themenkomplexe

A) Modellierung und Heterogenität

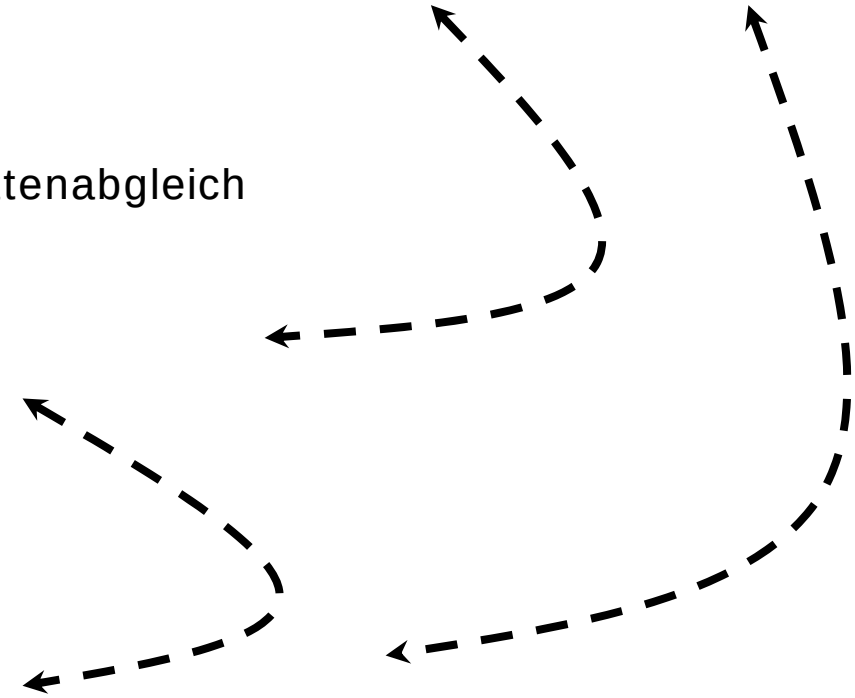
- Datenmodelle: Relational, XML-Schema, OWL, OO, ER, RDFS, F-Logic
- Heterogenitäten

B) Matchingtechniken

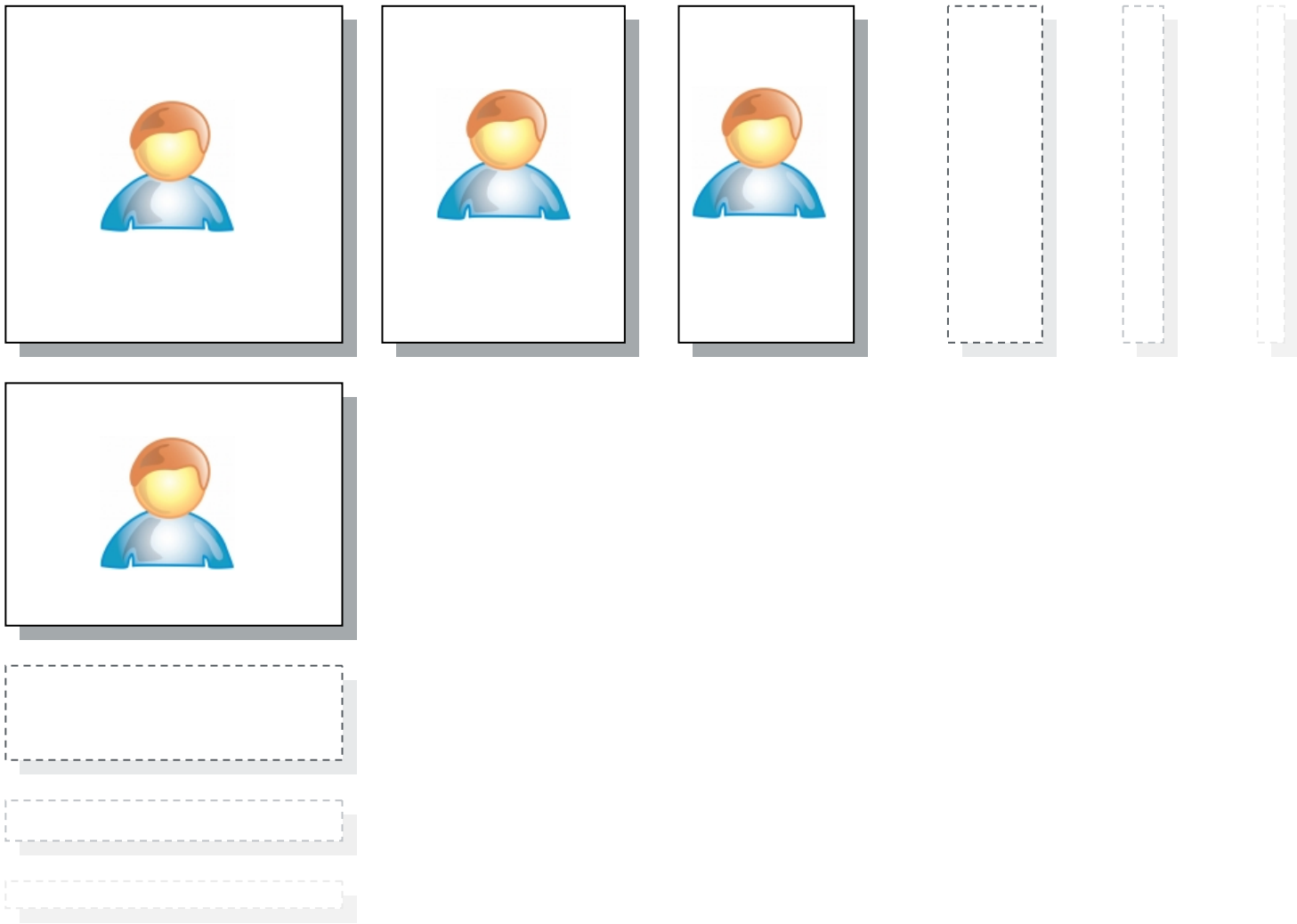
- Elementabgleich: Linguistisch, Metadatenabgleich
- Instanzabgleich
- Graphabgleich
- Lerntechniken
- Logisches Schlussfolgern

C) Matchingsysteme & Evaluation

- kommerzielle Systeme
- wissenschaftliche Prototypen
- Evaluation



Themenkomplexität



Die Themen

A. Datenmodelle

Datenmodelle

- Relational
- Bäume, XML-Schema
- OO (Objektrelationales Mapping, Java), Entity-Relationship
- Ontologien: OWL, RDF/S, F-Logic

Heterogenitäten …

- in allen Modellen
- in bestimmten Modellen (Modelleigenheiten)
- zwischen den Modellen

A. Datenmodelle
Datenmodellheterogenität: XML-RDB

```
<Products>
    <Product id="3551566666" productgroup="DVD">
        <title>Pulp Fiction</title>
        <details>http://www.amazon.de/dp/B00004TH46/</details>
        <actor>John Travolta</actor>
        <author>Quentin Tarantino</author>
    </Product>
    <Product id="3551566666"  productgroup="Book">  ... <Product>
    <Product id="B00004TH46" productgroup="Book"> ... <Product>
</Products>
```

Product

id	title	details
3551566666	Harry Potter ...	http://...
3899407288	Harry Potter ...	http://...
B00004TH46	Pulp Fiction	http://...

DVD

id	actor	author
B00004TH46	John Travolta	Quentin Tarantino

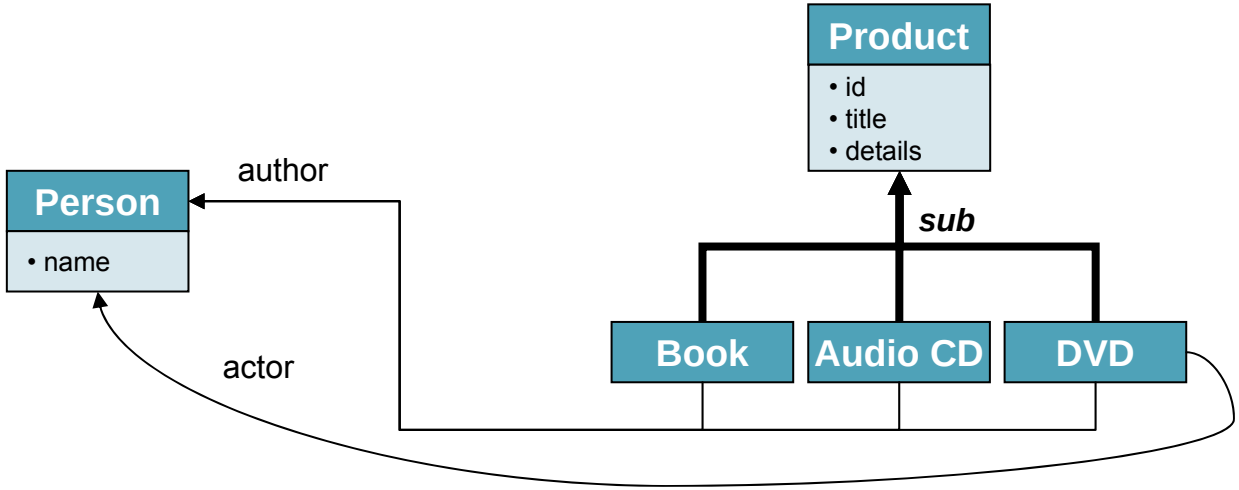
Book

id	author
3551566666	Joanne K. Rowling

Audio CD

id	author
3899407288	Rufus Beck

A. Datenmodelle
Datenmodellheterogenität: OO-RDB



Product

id	title	details
3551566666	Harry Potter ...	http://...
3899407288	Harry Potter ...	http://...
B00004TH46	Pulp Fiction	http://...

DVD

id	actor	author
B00004TH46	John Travolta	Quentin Tarantino

Book

id	author
3551566666	Joanne K. Rowling

Audio CD

id	author
3899407288	Rufus Beck

A. Datenmodelle
Schematische Heterogenität in XML

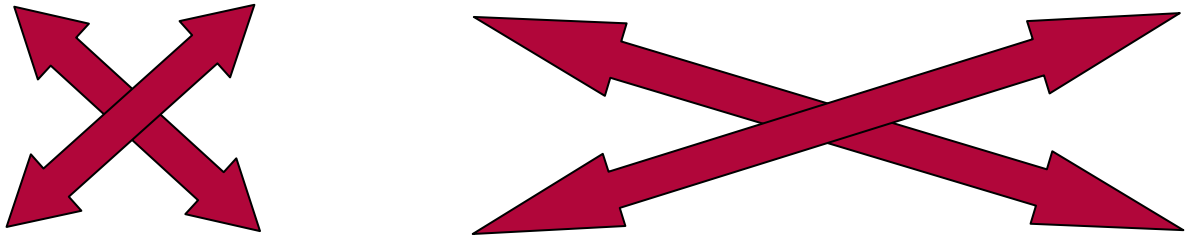
```
<Products>
    <Product id="3551566666" productgroup="DVD">
        <title>Pulp Fiction</title>
        <details>http://www.amazon.de/dp/B00004TH46/</details>
        <actor>John Travolta</actor>
        <author>Quentin Tarantino</author>
    </Product>
    <Product id="3551566666"  productgroup="Book">  ... <Product>
    <Product id="B00004TH46" productgroup="Book"> ... <Product>
</Products>
```

```
<Products>
    <Product title="Pulp Fiction">
        <id>3551566666</id>
        <attribute>
            <name>productGroup</name>
            <value>DVD</value>
        </attribute>
        <details>http://www.amazon.de/dp/B00004TH46/</details>
        <title>Pulp Fiction</title>
        <actor>John Travolta</actor>
        <author>Quentin Tarantino</author>
    </Product>
    <Product id="3551566666"  productgroup="Book">  ... <Product>
    <Product id="B00004TH46" productgroup="Book"> ... <Product>
</Products>
```

B.1. Intrinsisches Elementmatching

- Ähnlichkeit der Attributnamen wird gemessen
- Ähnliche Attributnamen sind potentiell bessere Matches als unähnliche Attributnamen.

ID	NName	VName	Tel	Adresse
1	Müller	Michael	123456	Hauptstr. 3, Berlin
2	Schmidt	Maria	987654	Dorfstr. 45, Golzow



PNR	Vorname	Nachname	Addresse	Telefon
1001	Michel	Müller	Hauptstrasse 3, Berlin	123456
1002	Peter	Schmidt	Forststr. 74, Hintertupfingen	918736

B.1. Intrinsisches Elementmatching String-Ähnlichkeit

Berechnung der Ähnlichkeit von Attributnamen

- Durch ein Ähnlichkeitsmaß
  - Je größer das Ergebnis des Ähnlichkeitsmaßes, desto ähnlicher sind die Strings.
    - Jaccard similarity
    - Cosine similarity
    - ...
- Durch ein Distanzmaß
  - Je kleiner das Ergebnis des Distanzmaßes, desto ähnlicher sind die Strings.
    - Edit distance
    - Q-Gram distance
    - ...
- Bei normalisierten Maßen: Ähnlichkeit = 1 - Distanz

B.1. Intrinsisches Elementmatching
Edit-Distance

Beispiel „HASE“ ⇒ „RASEN“

- triviale Umformung durch Einfügung und Löschung mittels Leerzeichen # am Anfang und Ende
- HASE##### ####RASEN

Transkript: DDDDIIIII

- Kosten 9
- Nicht minimal!

B.1. Intrinsisches Elementmatching
Edit-Distance contd.

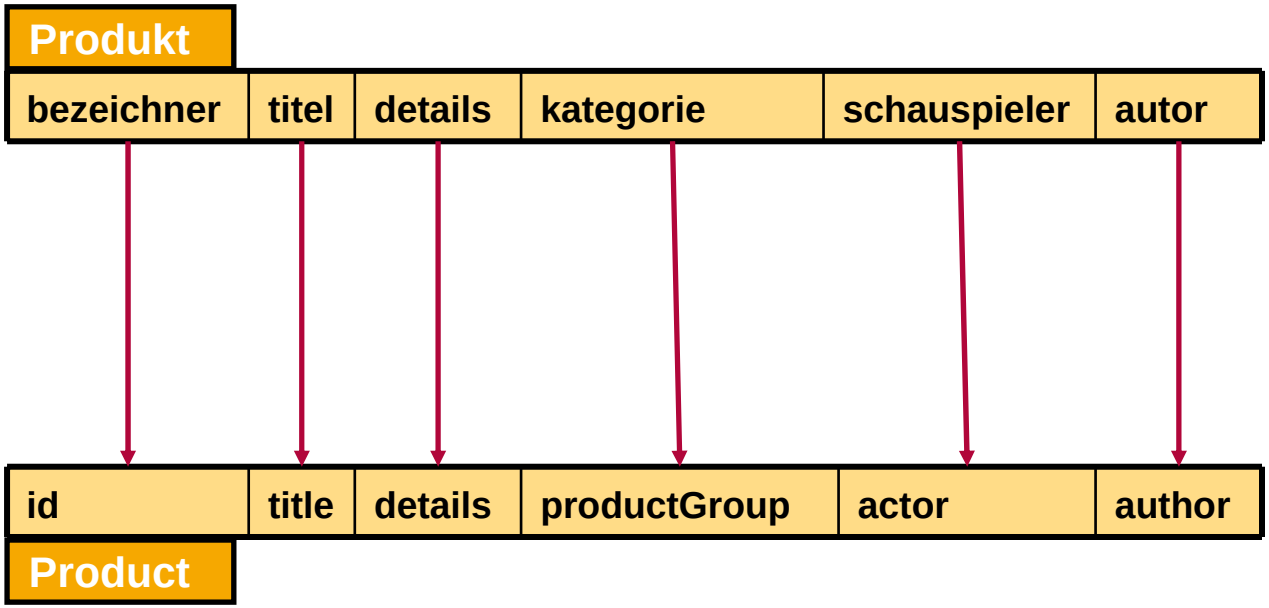
Beispiel „HASE“ ⇒ „RASEN“

- H → R durch Ersetzung (R)
- A → A durch Übereinstimmung (M)
- S → S durch Übereinstimmung (M)
- E → E durch Übereinstimmung (M)
- „ “→ N durch Einfügung (I)

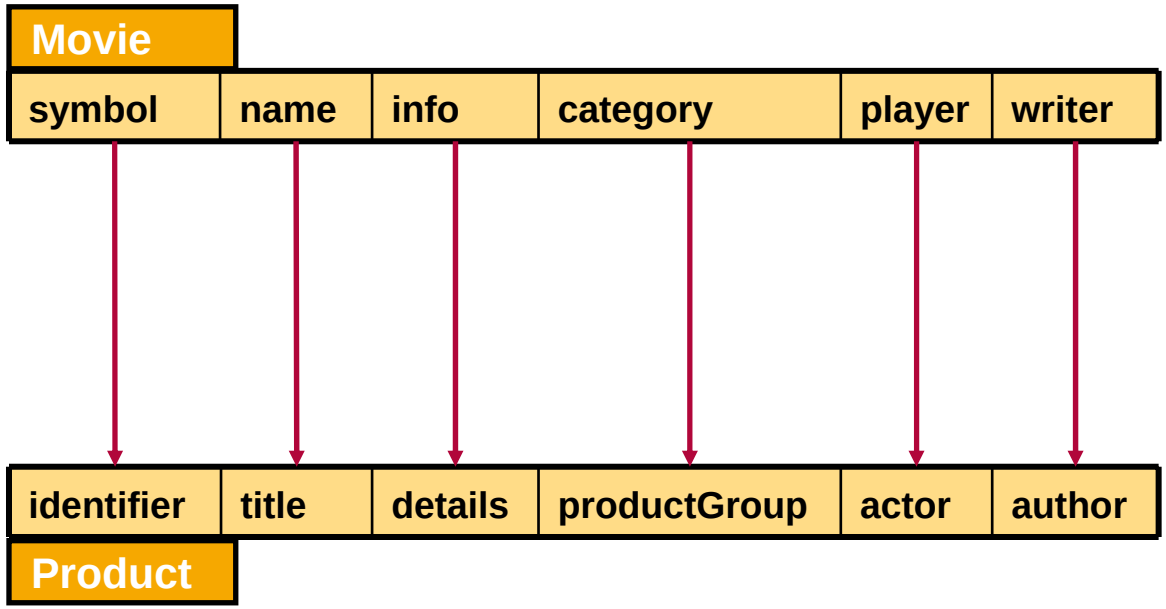
Transkript: RMMMI

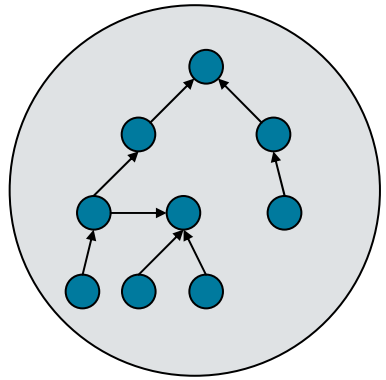
- EditDistance(HASE,RASEN) = 2
- Minimal! Woher weiß man das?

B.2. Extrinsisches Elementmatching Wörterbücher



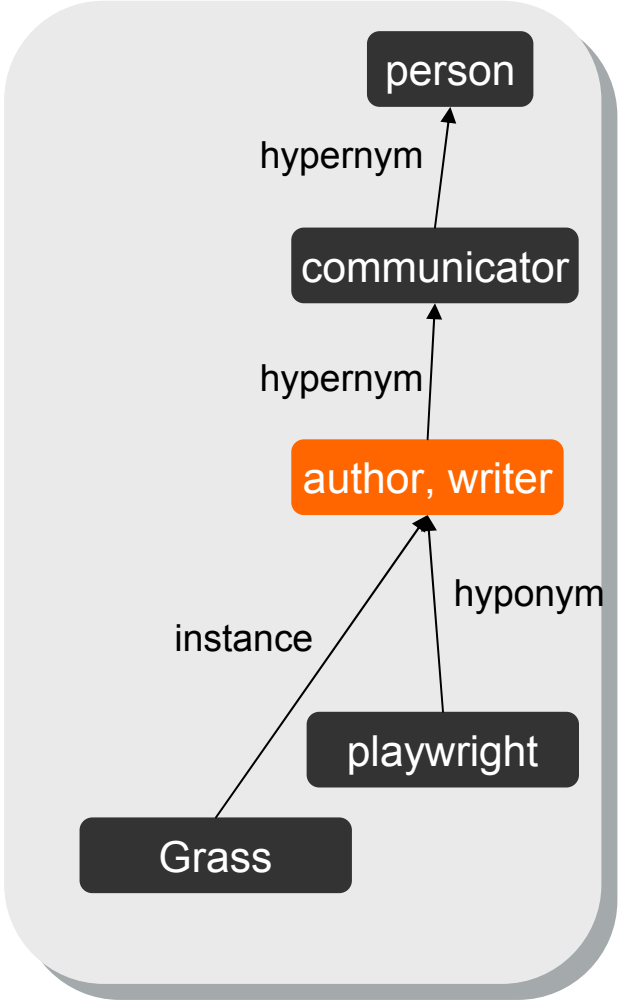
B.2. Extrinsisches Elementmatching Lexika





B.2. Extrinsisches Elementmatching WordNet

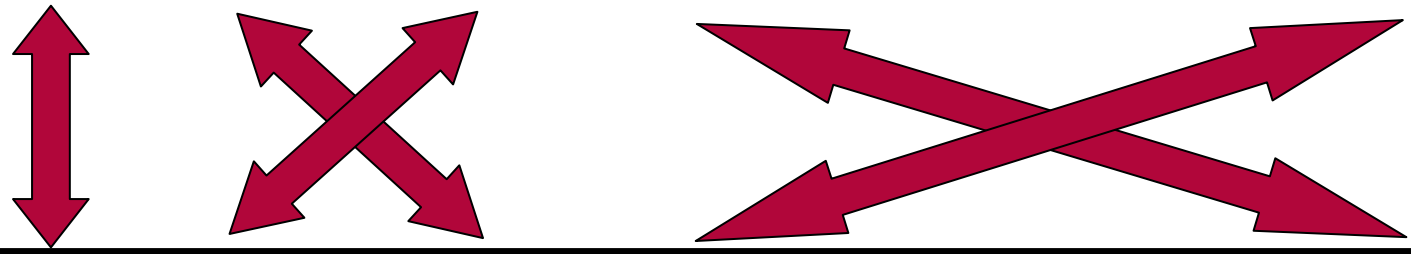
- Linguistische Ontologie
- Konzepte: Synonymmengen vs. Wörter
- **author**
  - **{author[1], writer}**
    Schriftsteller, Drehbuchautor (noun)
  - **{author[2], generator, source}**
    Verursacher, Initiator (noun)
  - **{author[3]}**
    verfassen (verb)
- Semantische Relationen
  - hypernym, hyponym, holonym, ...



B.3. Extensionales Matching

- Es werden ähnliche Daten oder Datencharakteristika ermittelt.
- Attribute / Attributkombinationen mit ähnlichen Daten sind potentiell bessere Matches als Attribute mit unähnlichen Daten.

ID	NName	VName	Tel	Adresse
1	Müller	Michael	123456	Hauptstr. 3, Berlin
2	Schmidt	Maria	987654	Dorfstr. 45, Golzow



PNR	Vorname	Nachname	Addresse	Telefon
1001	Michel	Müller	Hauptstrasse 3, Berlin	123456
1002	Peter	Schmidt	Forststr. 74, Hintertupfingen	918736

B.3. Extensionales Matching Ansätze

Betrachtung der Instanzdaten

- Betrachtung von Statistiken (Verteilung, Muster, …)
- Erkennung von Duplikaten zwischen den zu matchenden Tabellen
  - auf Attributebene
  - auf Tupelebene
- Erkennen von Transformationsregeln zwischen den Instanzen

B.3. Extensionales Matching
Ähnlichkeitsmaße

Ähnlichkeitsmaße für

- lange Strings
- Strings, die aus mehreren Worten (Token) bestehen
- komplexe Strukturen (Tupel statt Strings, XML Daten)

Beispiele

- Jaccard Similarity
- Cosine Similarity
- SoftTFIDF
- ...

B.3. Extensionales Matching
Jaccard Similarity

ID	NName	VName	Tel	Adresse
1	Müller	Michael	123456	Hauptstr. 3, Berlin

PNR	Name	Anschrift	Telefon
1001	Michel Müller	Hauptstrasse 3, Berlin	123456

1. Tokenmenge (TM) erzeugen

$$TM(T1) = \{1, \text{Müller}, \text{Michael}, 123456, \text{Hauptstr.}, 3, \text{Berlin}\}$$

$$TM(T2) = \{1001, \text{Michael}, \text{Müller}, \text{Hauptstrasse}, 3, \text{Berlin}, 123456\}$$

2. Ähnliche Token identifizieren

Michael = Michel

Hauptstr. = Hauptstrasse

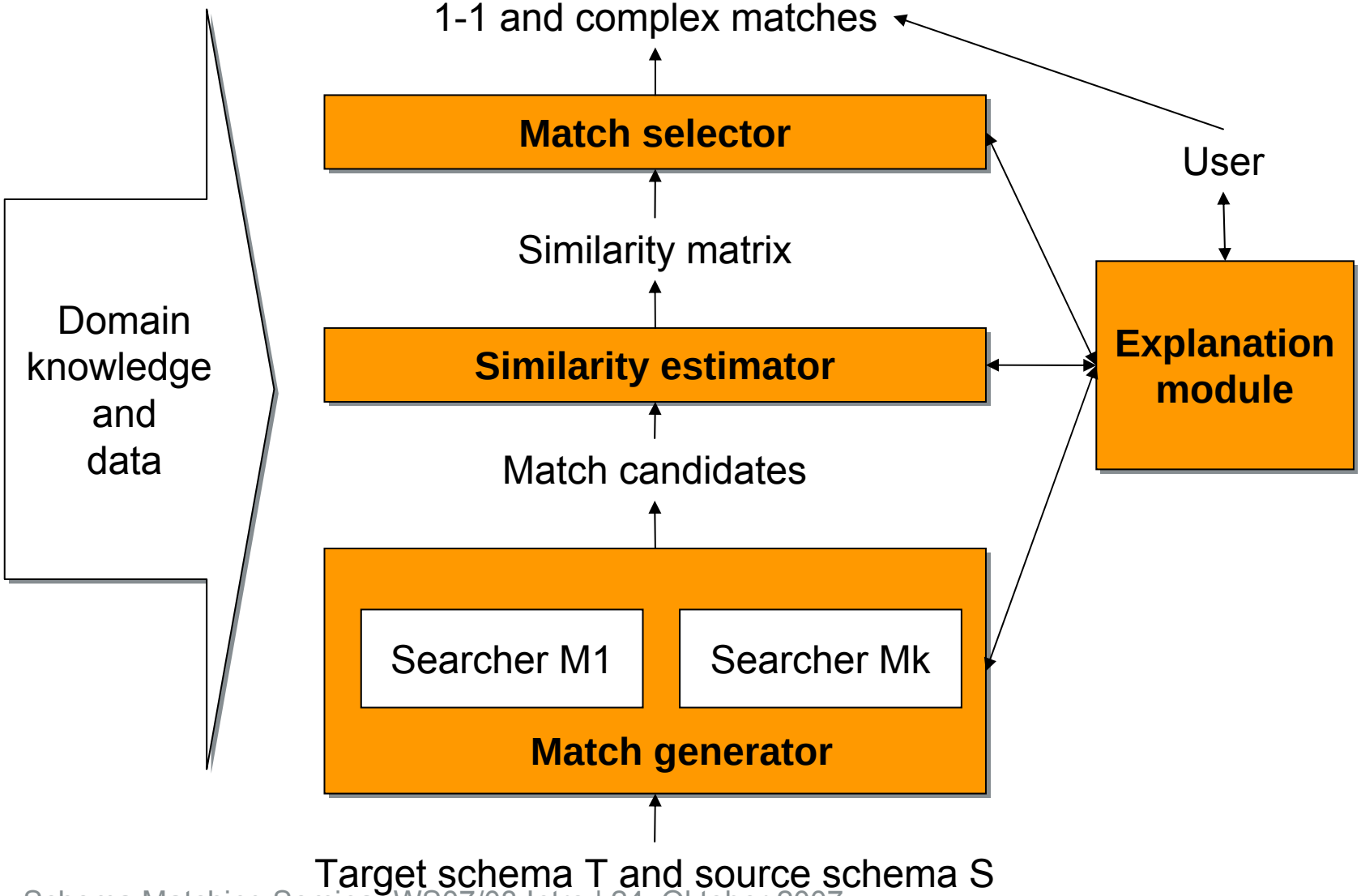
3. Jaccard Ähnlichkeit berechnen

$$Jaccard(TM(T1), TM(T2)) = \frac{|TM(T1) \cap TM(T2)|}{|TM(T1) \cup TM(T2)|} = \frac{6}{7}$$

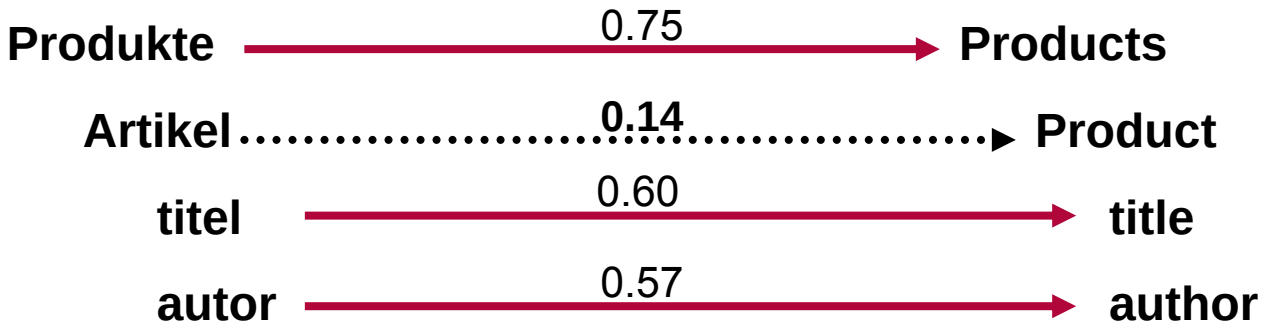
B.3. Extensionales Matching System iMap

- Finden von 1:1 und komplexen Matchings
- Durchsucht den Raum aller möglichen Matches anhand von spezialisierten Searcher Modulen
  - text searcher: concat(Name, Vorname)
  - numeric searcher: preis = preis + Mwst
  - date searcher: datum = concat(Monat, „.“, Jahr)
- Setzt frühzeitig Domänenwissen ein
  - Integritätsbedingungen
  - Betrachtet überlappende Tupel (Duplikate)
  - Erlaubt, Daten außerhalb der Datenbank zu verwenden

B.3. Extensionales Matching System iMap

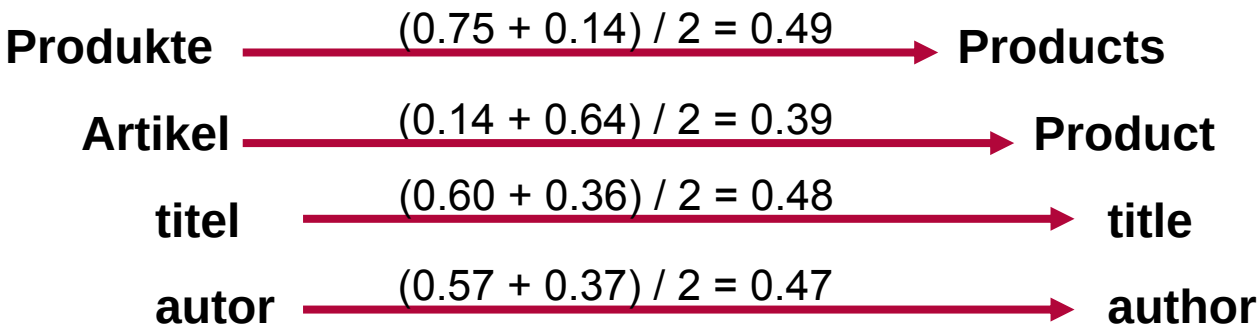


B.4. Strukturbasiertes Matching



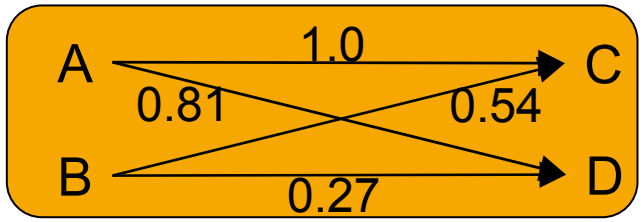
Min: 0.30

Idee: Ähnlichkeit = (Elementähnlichkeit
+ durchschnittliche Elementähnlichkeit der direkten Nachbarn) / 2



B.5. Globales Matching

- Ergebnis der vorherigen Verfahren

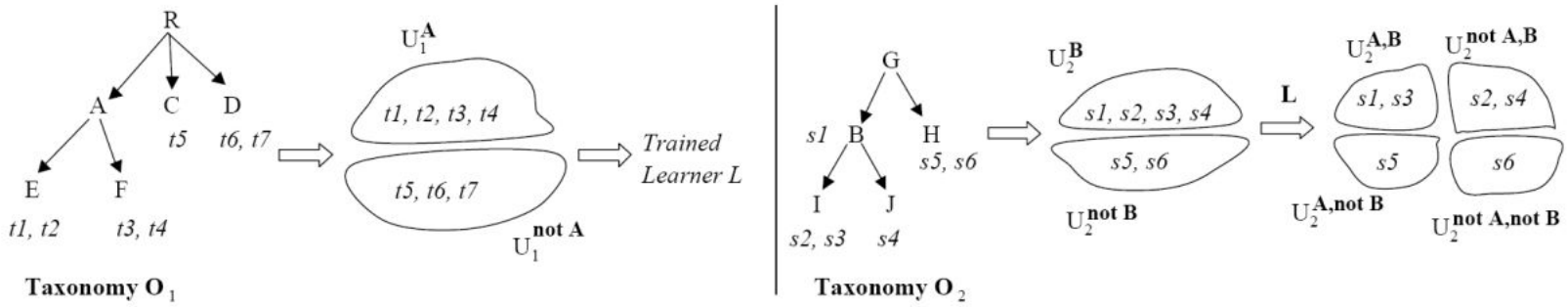


- Algorithmen zur Extraktion des eigentlichen (besten) Mappings
  - Stable Marriage
  - Maximum Weighted Bipartite Graph Matching
- Vergleich der beiden Verfahren
- Praktische Aufgabe:
  - Enummerierung mehrerer Stable Marriages
  - Finden der besten Stable Marriage

B.6. Maschinelles Lernen

- Stochastik
- Künstliche Intelligenz
- Unüberwachtes vs. Überwachtes Lernen

- Ansätze
  - Naive Bayes - **GLUE (LSD)**
  - Neuronale Netzwerke (Semint)
  - Entscheidungsbäume
  - ...
- Survey



B.7. Semantisches Matching
Ontologien vs. Schemas

Ontologien: formale Spezifikation einer gemeinsamen Konzeptualisierung

- Schemas (Konzepte, Relationen)
- Instanzen
- Logische Axiome

Unterschiede

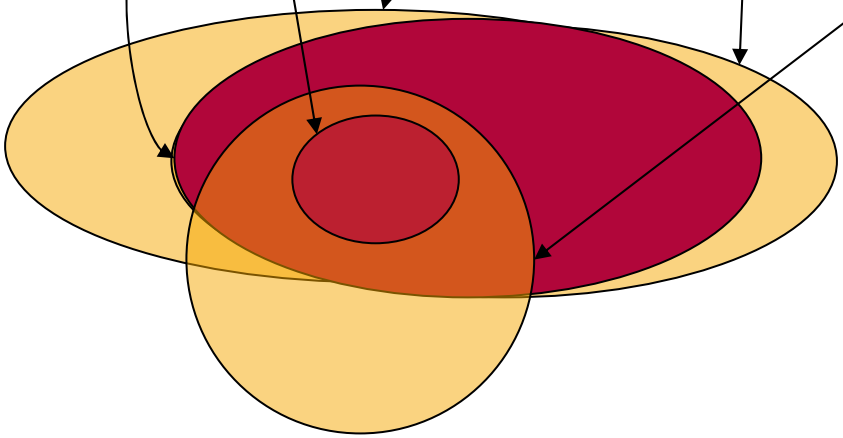
- Objekt-Orientierung
- Betonung auf konzeptionelle Hierarchien (subclass, hypernym)
- Metamodellierung (Instanzen <-> Konzepte)
- Anwendung: Wissensmodell vs. Datenmodell (expliziter)
- **Formale Logik und Semantik**

B.7. Semantisches Matching
Logik, Semantik, Inferenz

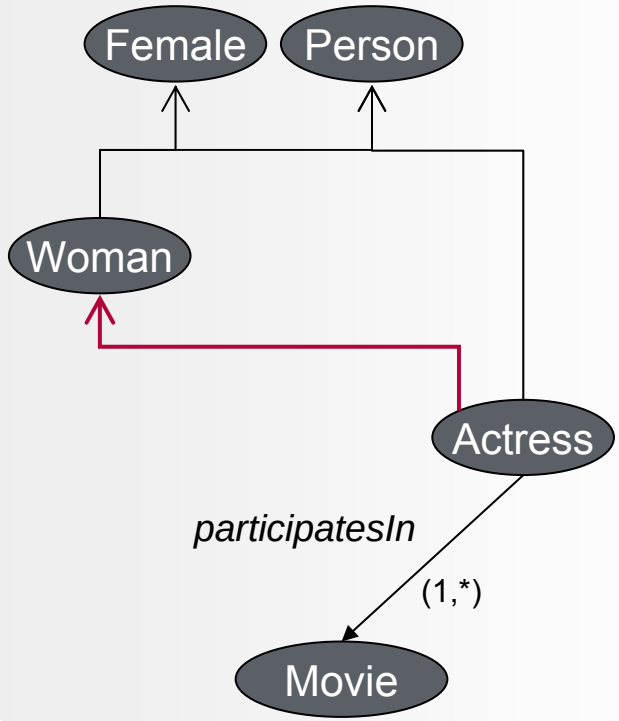
$\mathcal{O}$

Woman $\equiv$ Female $\sqcap$ Person.

Actress $\sqsubseteq$ Female $\sqcap$ Person $\sqcap \exists$ participatesIn.Movie.



$\mathcal{O} \models$ Actress $\sqsubseteq$ Woman

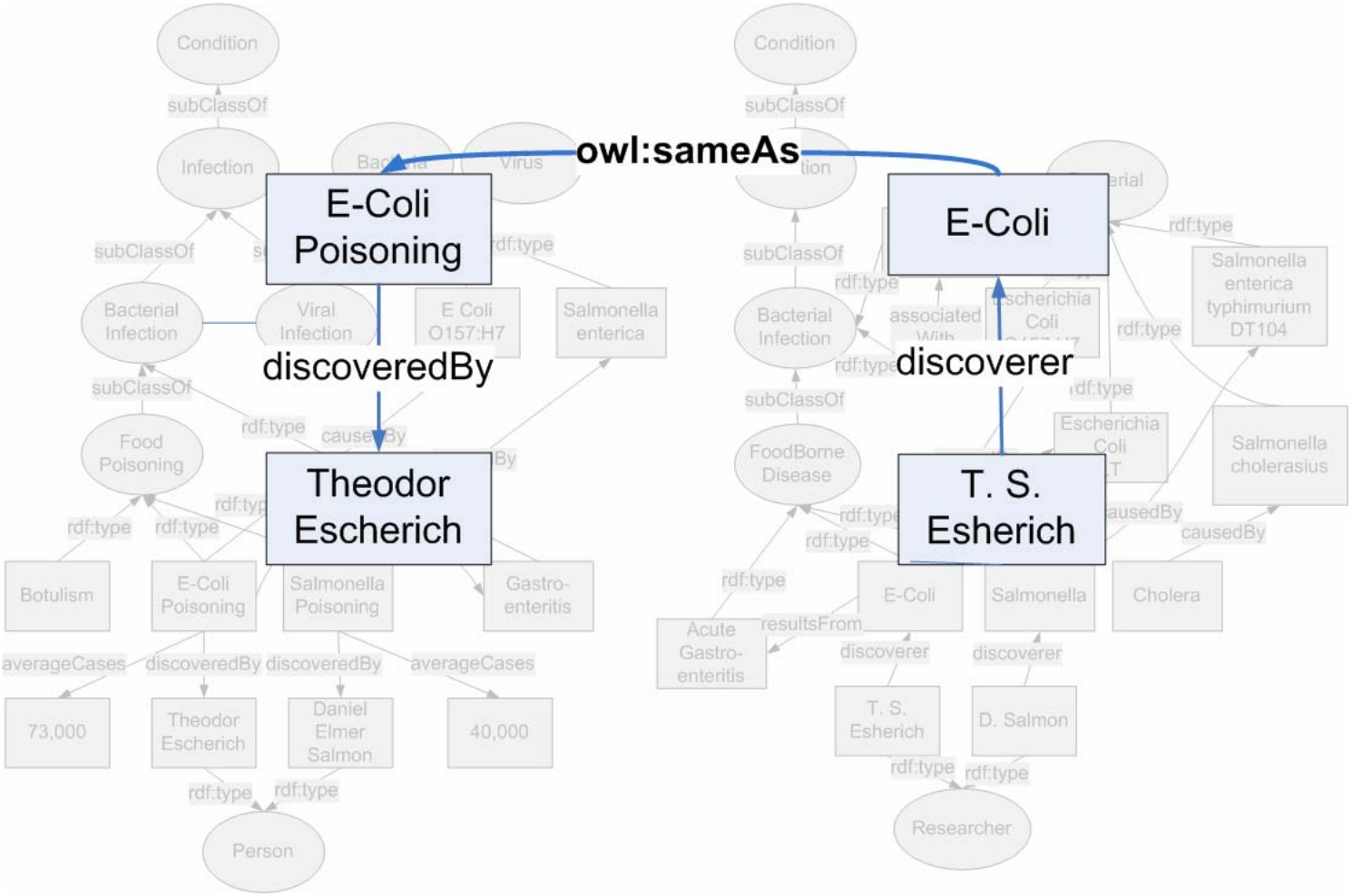


B.7. Semantisches Matching
Ontology Matching

Schema Matching mit zusätzlichen Methoden & Herausforderungen

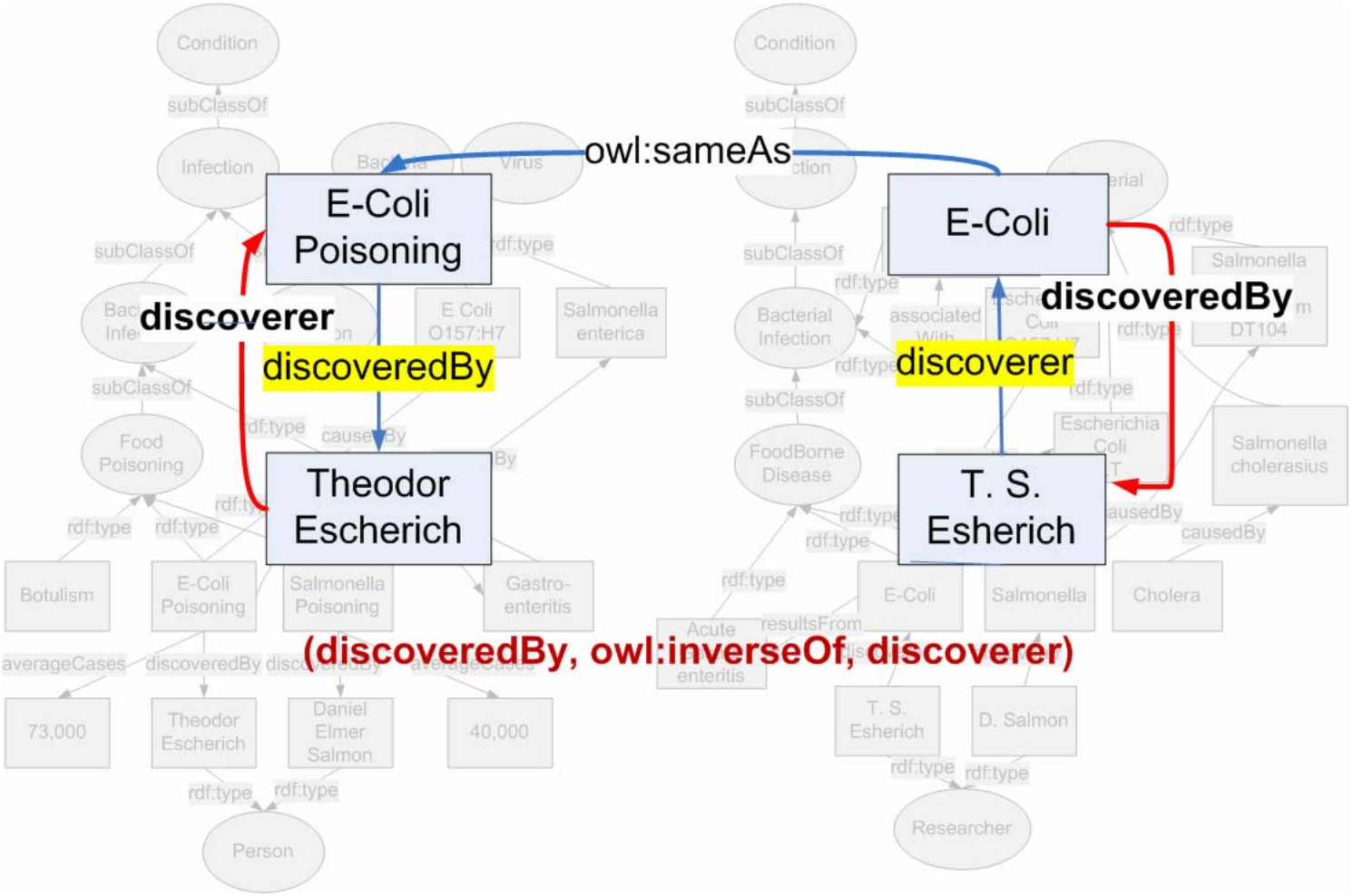
- Semantisches Matching ~ Logische Inferenz (Schlussfolgern)
- Semantische Relationen statt [0,1]-Werte
  - Äquivalenz, Implikation, Disjunktion, Konjunktion
  - Gleichheit, Untermenge, Vereinigung, Schnittmenge, Disjunktheit
- Integration von syntaktischem und semantischem Matching
- Systeme
  - S-Match (Aussagenlogik)
  - ILIADS (OWL/Beschreibungslogik + syntaktisches Matching)

B.7. Semantisches Matching ILIADS



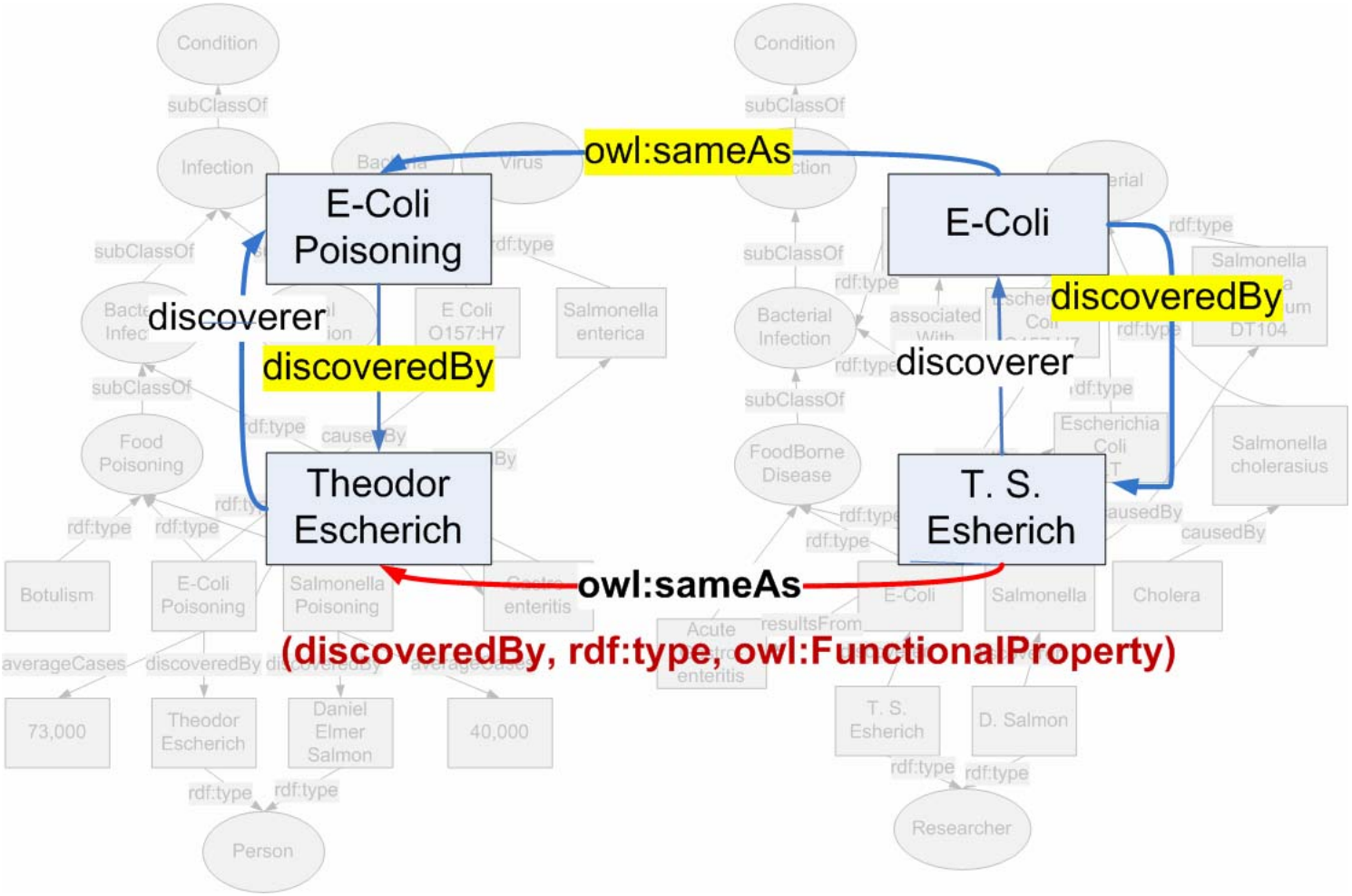
Source: http://www.cs.umd.edu/projects/linqs/iliads

B.7. Semantisches Matching
ILIADS



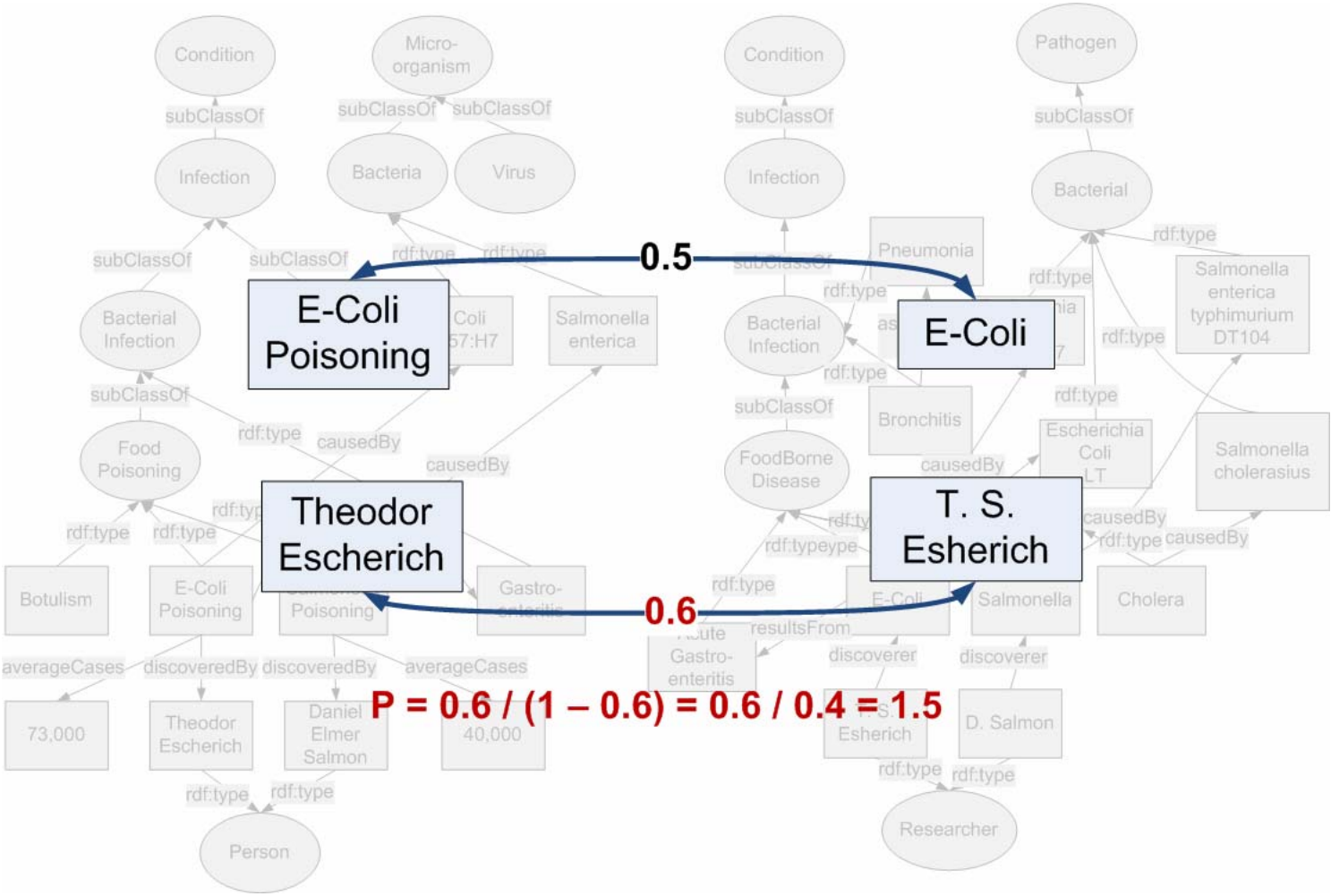
Source: http://www.cs.umd.edu/projects/linqs/iliads

B.7. Semantisches Matching
ILIADS



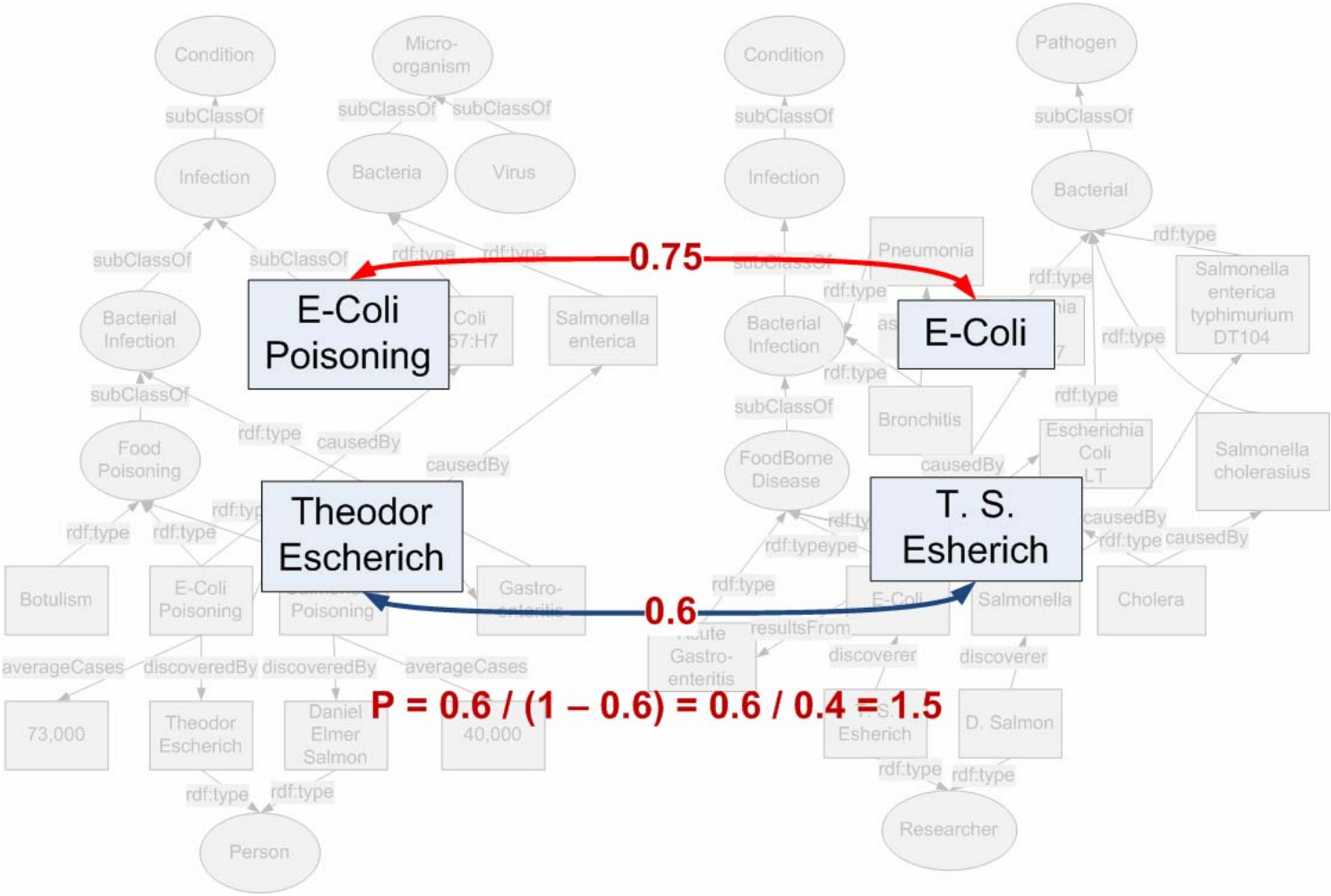
Source: http://www.cs.umd.edu/projects/linqs/iliads

B.7. Semantisches Matching
ILIADS



Source: http://www.cs.umd.edu/projects/linqs/iliads

B.7. Semantisches Matching
ILIADS

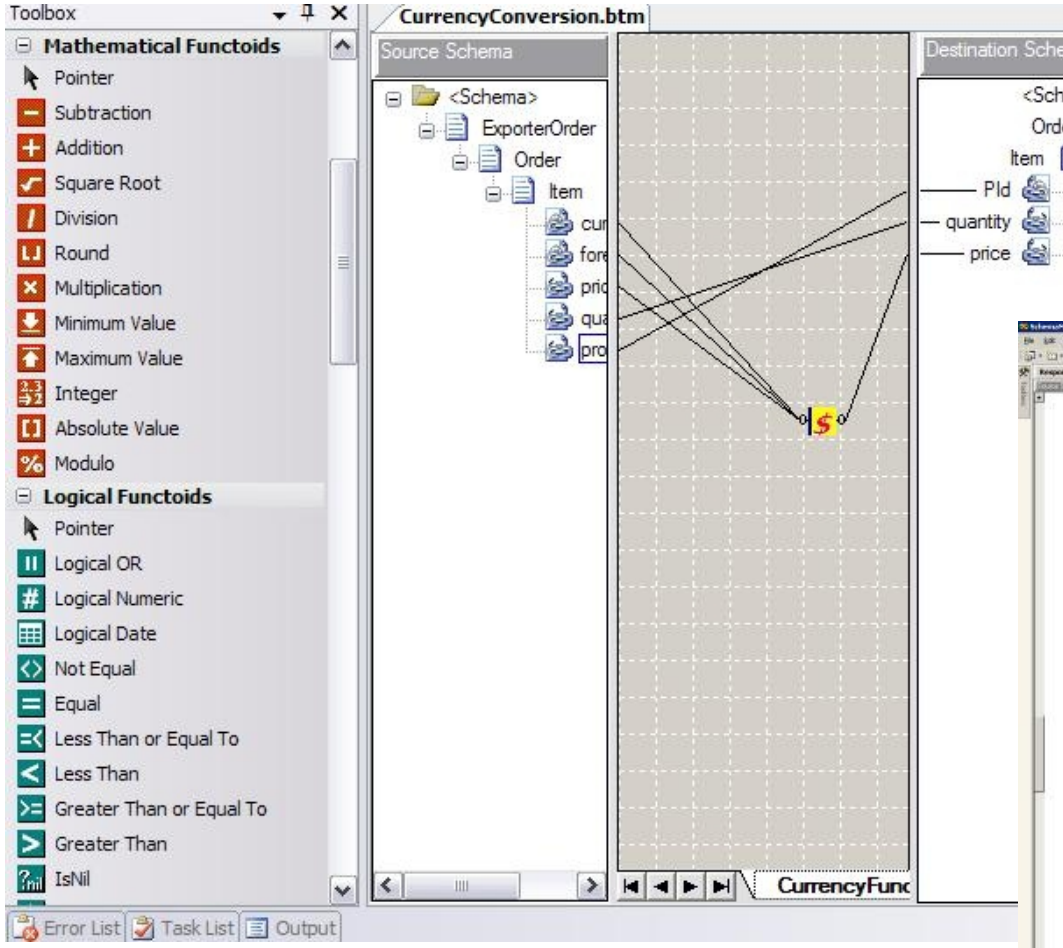


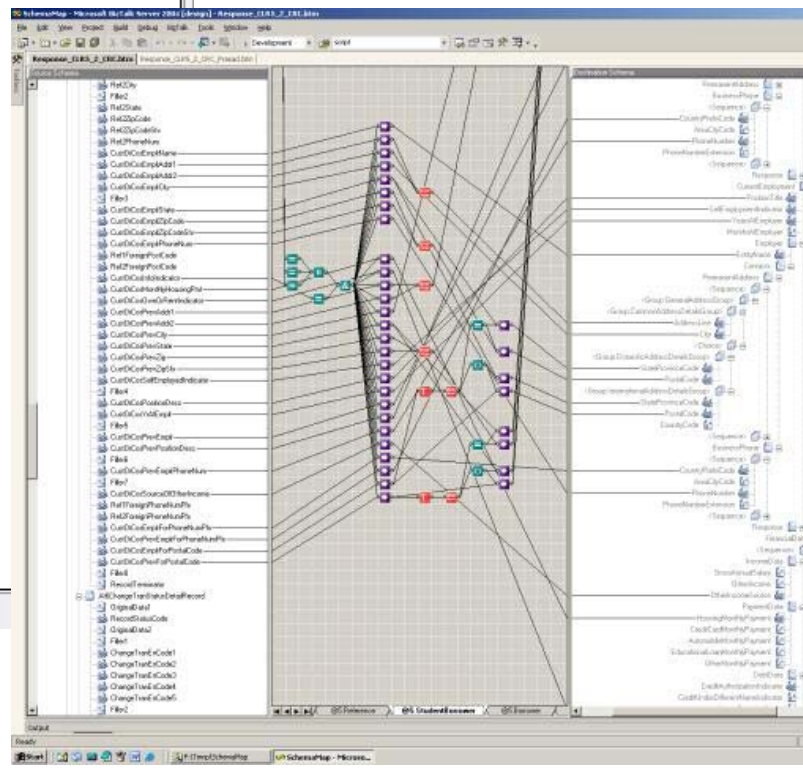
Source: http://www.cs.umd.edu/projects/linqs/iliads

C.1. Evaluationsverfahren

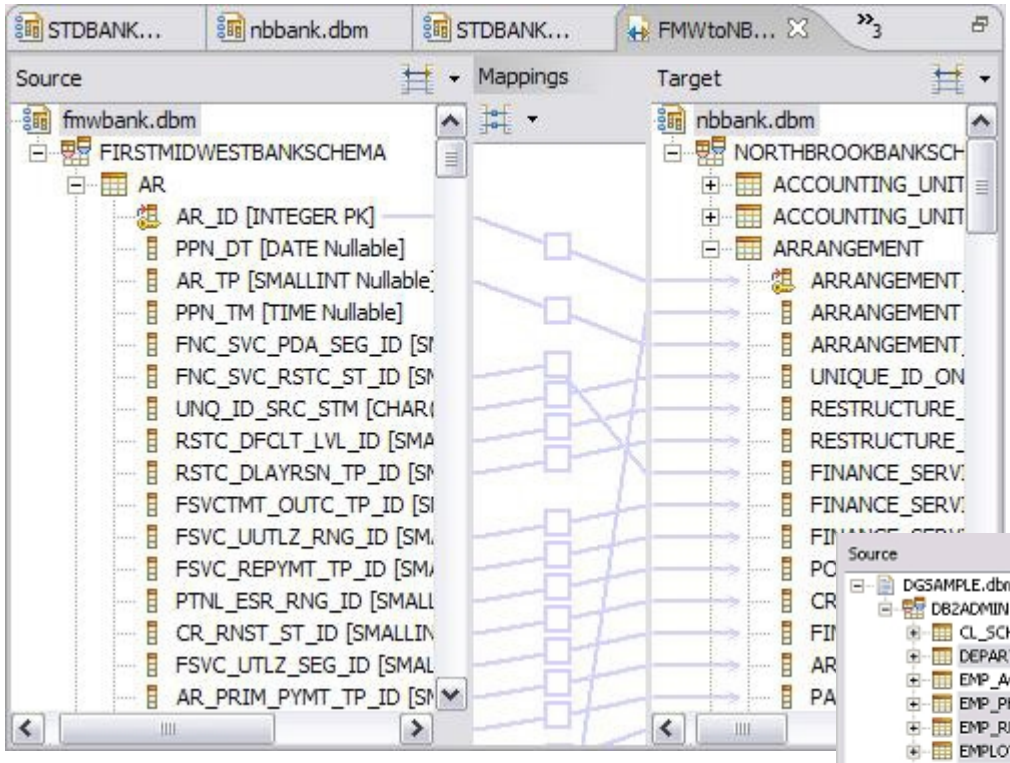
- Methoden
- Benchmarks
  - OAEI (Ontology Alignment Evaluation Initiative)
  - XBenchMatch
  - Thalia Benchmark
  - ...
- Diverse Datensätze
- Maße
  - Recall / Precision / F-Measure
  - OverAll Maß (Cupid)

C.2. Microsoft BizTalk-Mapper

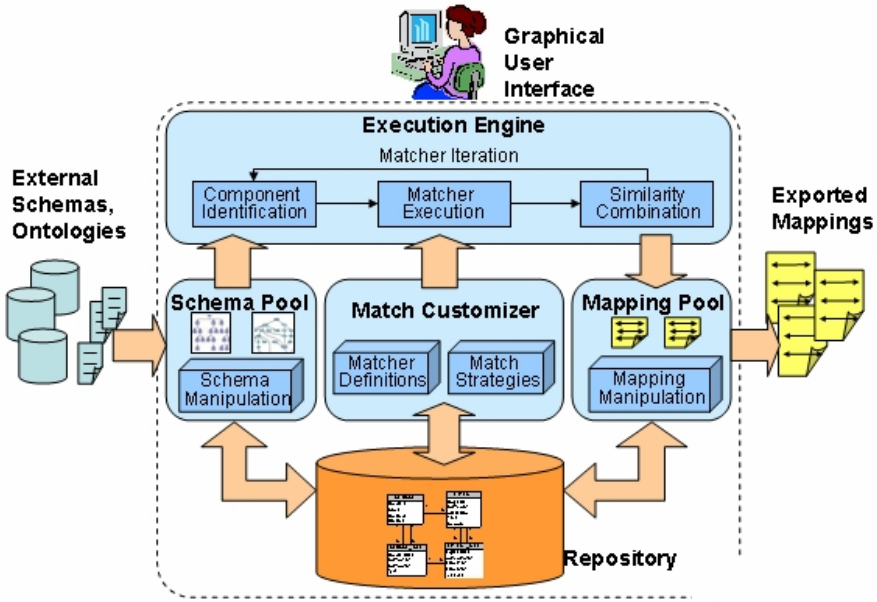


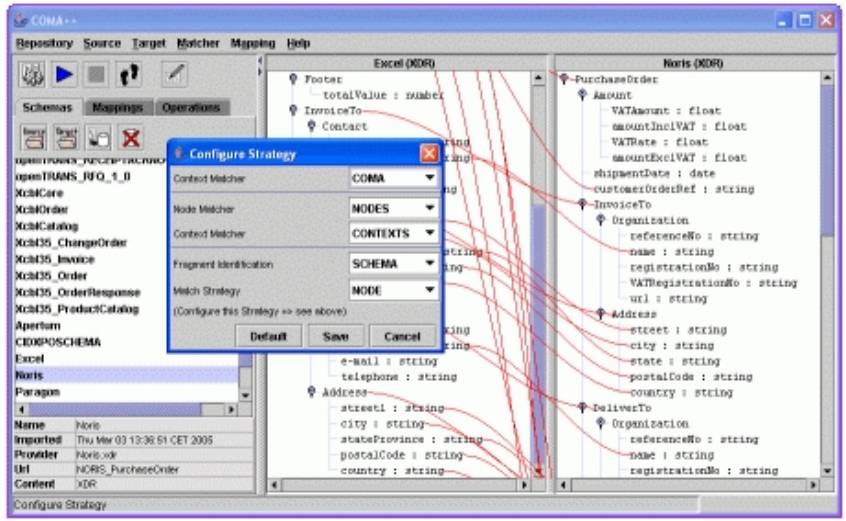


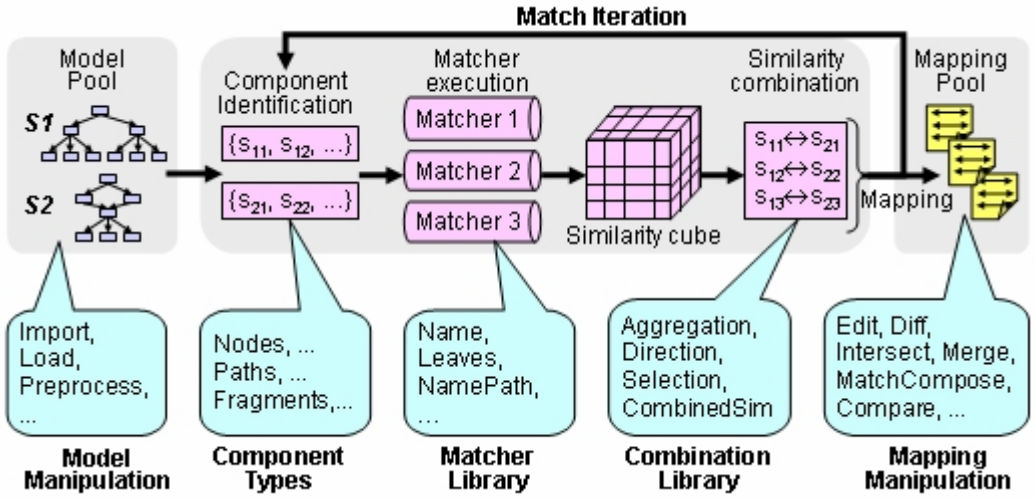
C.3. IBM Rational Data Architect



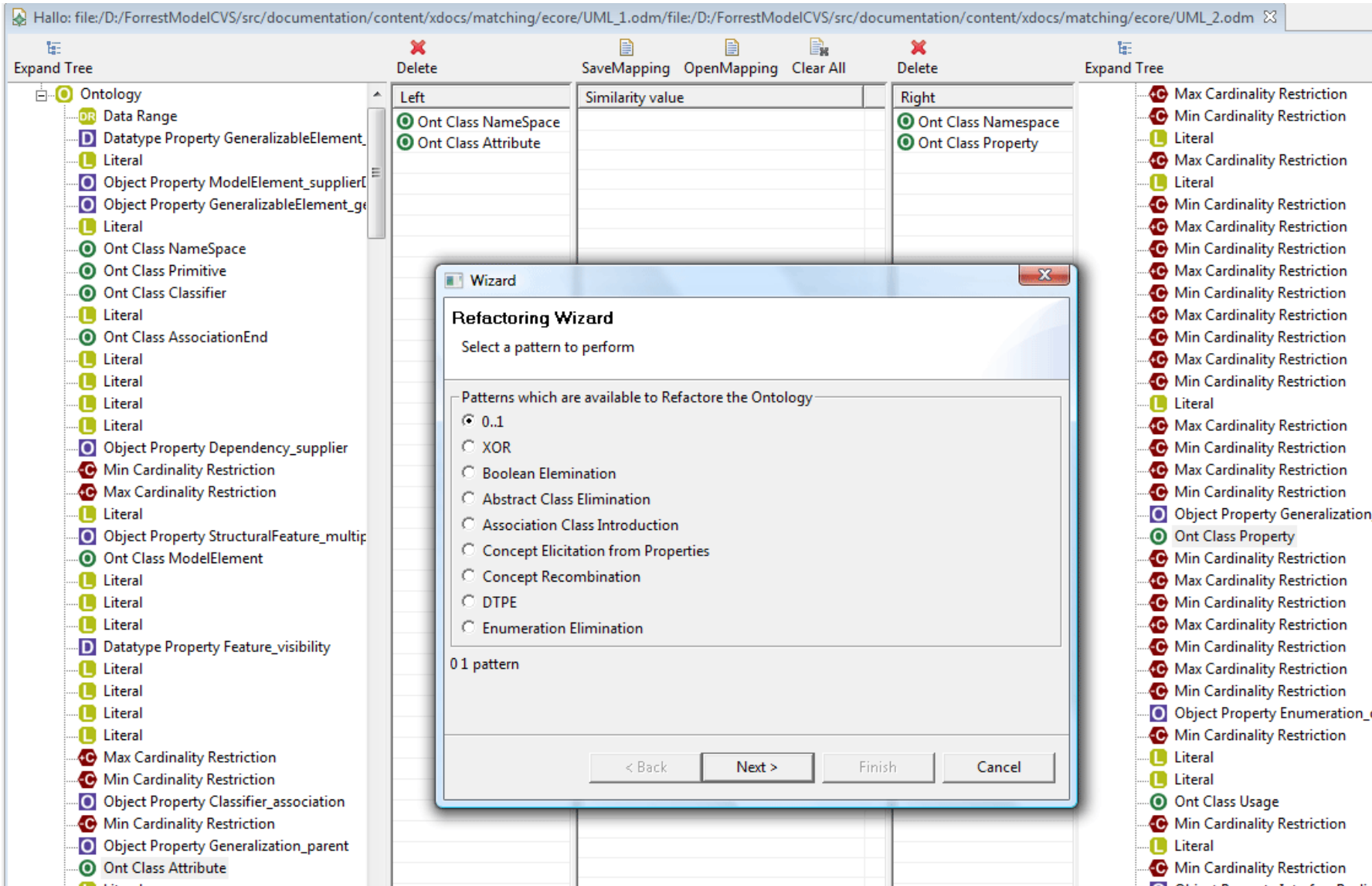
C.4. Coma++/QuickMig (Uni Leipzig/SAP)







C.4. Coma++/QuickMig Eclipse-Plugin Ontohammer



C.5. Model Management

Ziele

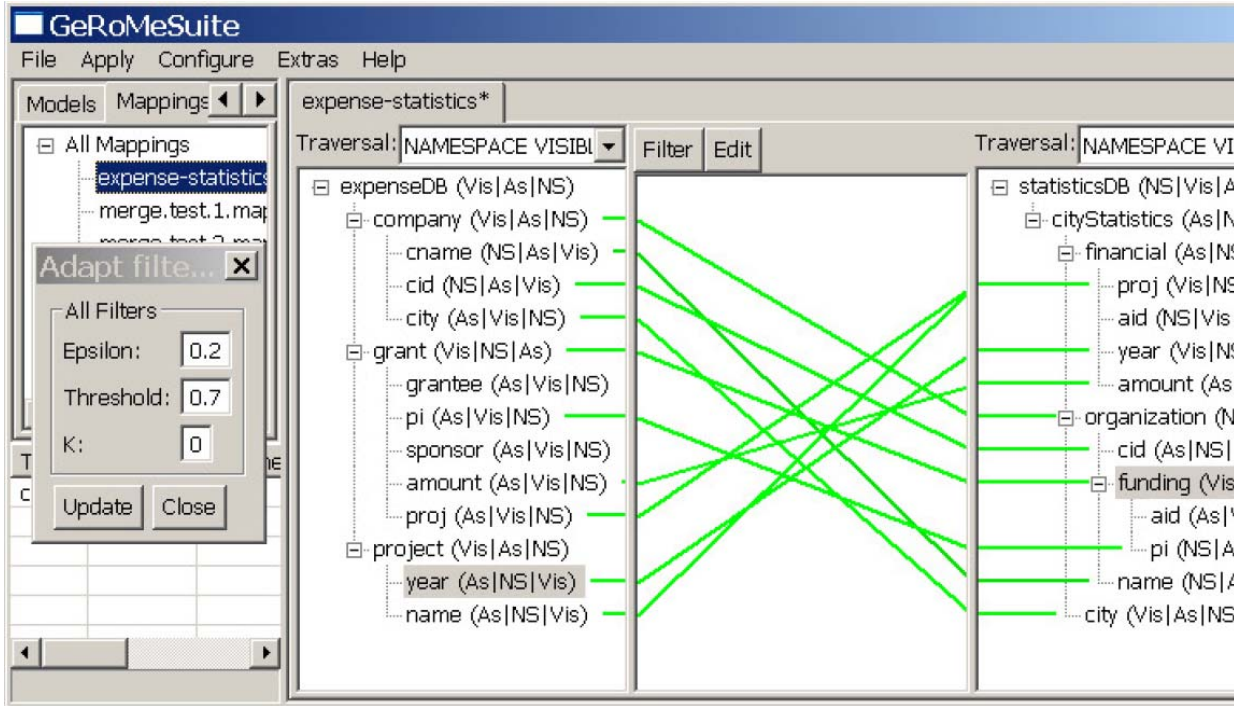
- Data Progammability
- Modellevolution
- Generische Operatoren für Datenmodelle

Operatoren

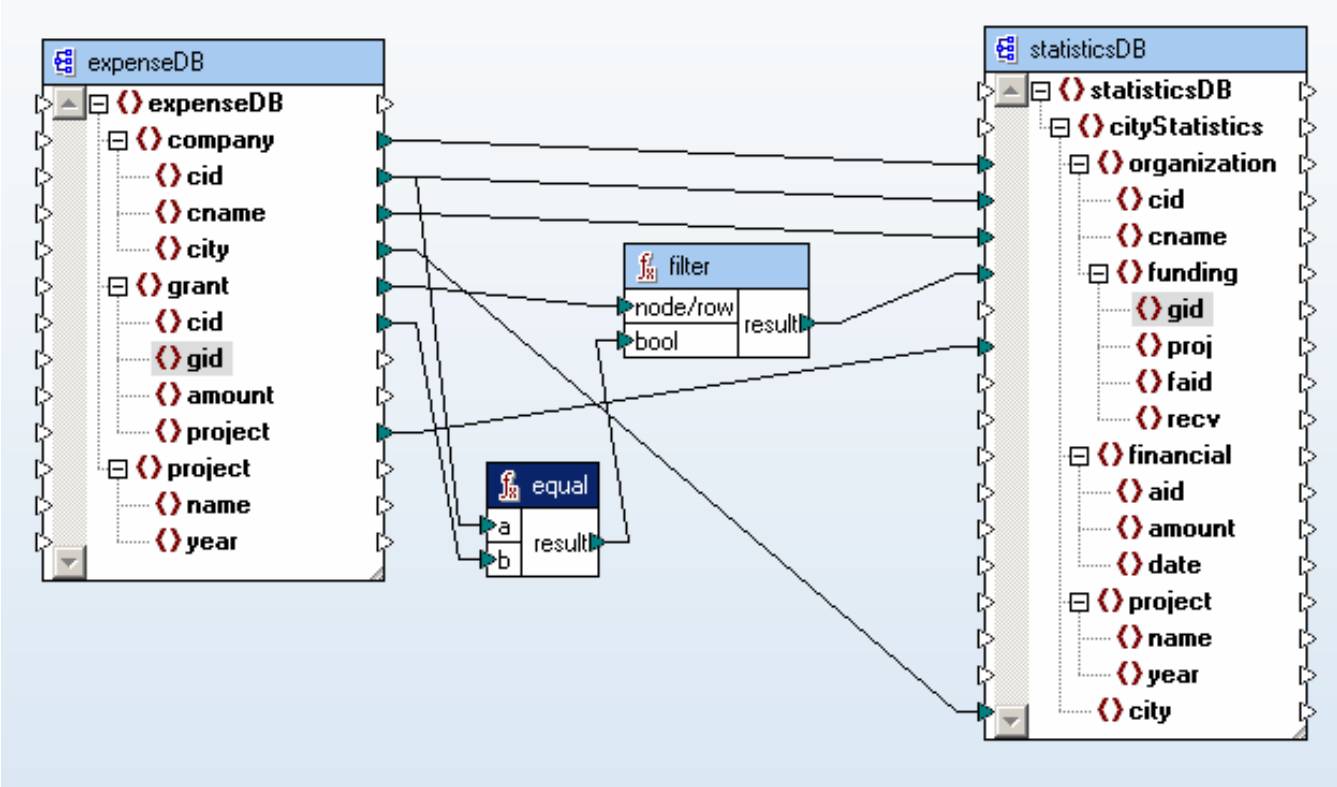
- Match
- Merge
- Compose
- Inverse
- Diff
- ModelGen
- ...

Systeme

- Rondo (Microsoft Research)
- GeRoMeSuite (RWTH Aachen)



C.6. Altova MapForce



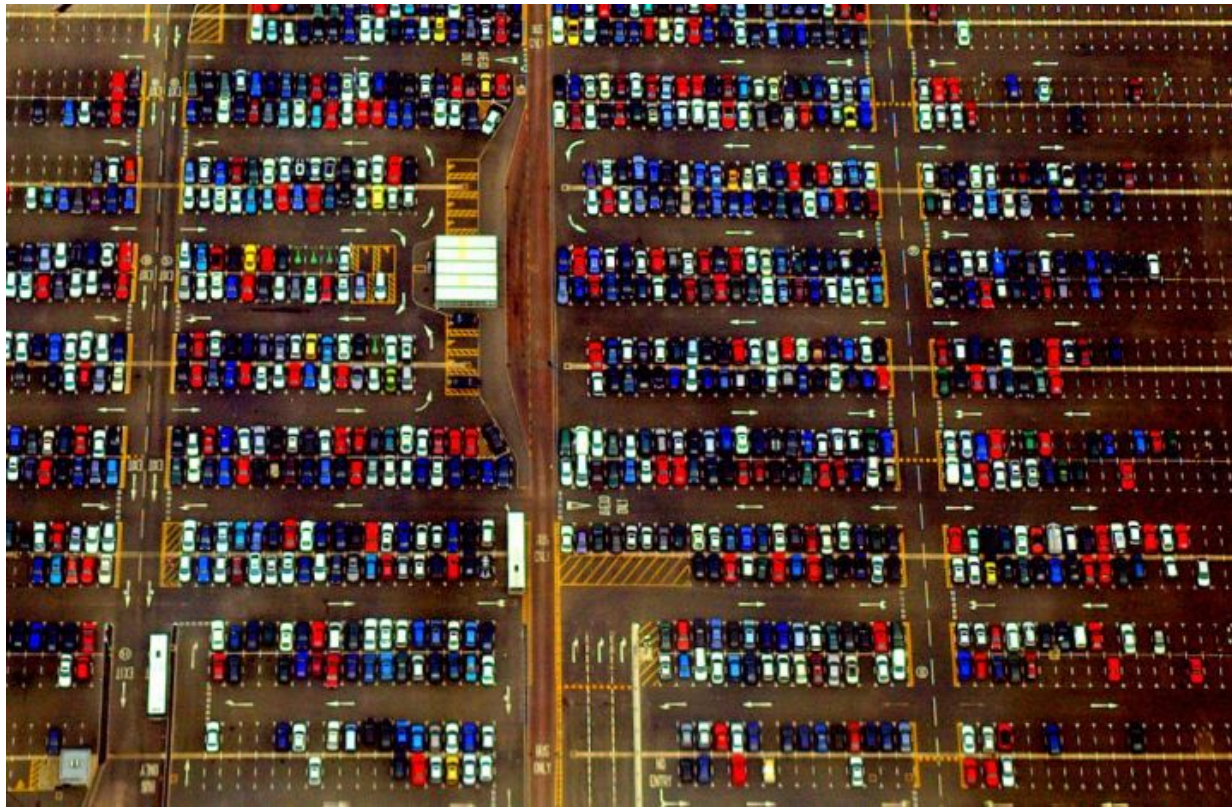
Themenübersicht (Betreuer)

A. Modellierung (Frank Kaufer)
B. Matchingtechniken
   1. Intrinsisches Elementmatching (Melanie Weis)
   2. Extrinsisches Elementmatching (Frank Kaufer)
   3. Extensionales Matching (Melanie Weis)
   4. Strukturbasiertes Matching (Frank Kaufer)
   5. Globales Matching (Frank Kaufer)
   6. Maschinelles Lernen (Frank Kaufer)
   7. Semantisches Matching (Frank Kaufer)
C. Matchingsysteme & Evaluation
   1. Evaluationsverfahren (Felix Naumann)
   2. Microsoft BizTalk-Mapper (Felix Naumann)
   3. IBM Rational Data Architect (Felix Naumann)
   4. Coma++/QuickMig (SAP/Uni-Leipzig) (Felix Naumann)
   5. Model Management (Alexander Albrecht)
   6. Altova MapForce (Alexander Albrecht)




Weitere Themen bei Bedarf

Organisation



Ablauf

- Mittwoch, 9:15 – 10:45: interne Treffen
- Vorträge en bloc im Dezember und/oder Februar
- Betreungstreffen: individuell per Email anfragen
- Mailingliste
  - seminar-sm0708@hpi.uni-potsdam.de
  - http://lists.hpi.uni-potsdam.de/listinfo/seminar-sm0708
  - Fachliche & organisatorische Diskussion
  - Archiv
- Wikizugang nach Anmeldung
  - http://www.hpi.uni-potsdam.de/naumann/a/matching

Themenvergabe

- 3 Wunschthemen
- Email an Frank Kaufer
- Vergabe nach
  - Priorität
  - Kohärenz
- Bewerbungsfrist: Donnerstag, 25.10.2007, 23:59 Uhr
- Anmeldefrist: 06.11.2007

Leistungserfassungsprozess

- Gesamtbewertung von
  - Wikiarbeit (Deadline 01.03.2008)
  - Vortrag 30 min + Diskussion
  - Praktische Arbeit
- Bonus
  - Wiki -> Wikipedia
  - Mailingliste
  - Gemeinsame, kombinierte Implementationen

Fragen!