Informationsintegration

Semantic Web

Ulf Leser

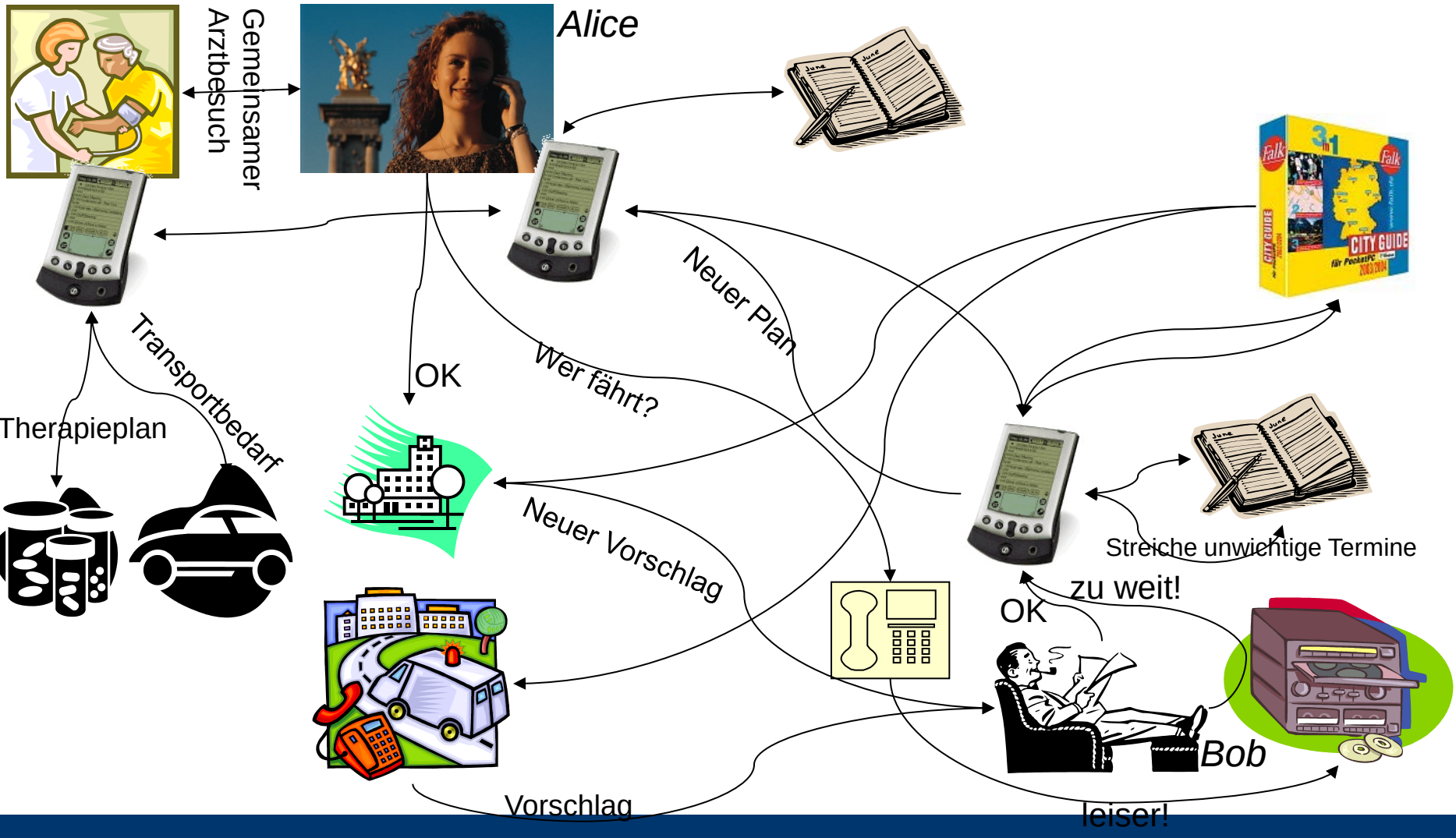
Inhalt dieser Vorlesung

- Grundidee des Semantic Web
- Layer Cake
- RDF und RDFS
- SparQL
- Die OWL Sprachfamilie

Tim Berners-Lee et al.

- „Das Semantic Web ist eine Erweiterung des gegenwärtigen Web, in der Informationen eine wohldefinierte Bedeutung erhalten, so dass Computer und Menschen besser zusammenarbeiten können“ [BHL01]
  - Erweiterung, kein Neudesign; Abwärtskompatibilität ist essentiell
  - Zusammenarbeit zwischen Menschen und Computern und Computern und Computern verbessern
    - Integration von Daten und Anwendungen
    - Intelligentere, persönliche, kontextbezogene, ... Dienste
  - Mittel: Explizite Definition der Bedeutung von Informationen
- Sollte als Vision verstanden werden
  - An deren Erfüllung man arbeitet (z.B. W3C)

Szenario [BHL01]



Ist-Zustand

- Web Seiten werden in HTML verfasst
- Geprägt von Layout-Informationen
  - Gut für Menschen
  - Kaum zu interpretieren für Rechner
- Informationen leben in zwei Welten
  - Für Menschen als Konsumenten
    - Gedichte, Filme, Text,...
  - Für Computer als Konsumenten
    - Daten, Programme,...
- Das Web betont den Menschen
- Das Semantic Web soll dies ausgleichen

Beispiel-Anwendungen

- Wissensmanagement
  - Intranet mit Millionen Dokumenten
  - Informationsbeschaffung, -wartung und -suche
  - Anfragen statt Suche
    - Liste alle Telefonnummern aller Mitarbeiter der HU Informatik
    - Wann wurde Rembrandt geboren?
    - Wann wurden die großen niederländischen Maler geboren?
- Web Commerce
  - Shopping-Agenten suchen bestes und billigstes Angebot
  - Online Shops präsentieren Waren zielgerichtet (Präferenzen)
  - Broker vermitteln gezielt zwischen Anbietern und Käufern (e-marketplace)
- E-Business
  - Virtuelle Unternehmen
  - Katalog-Integration

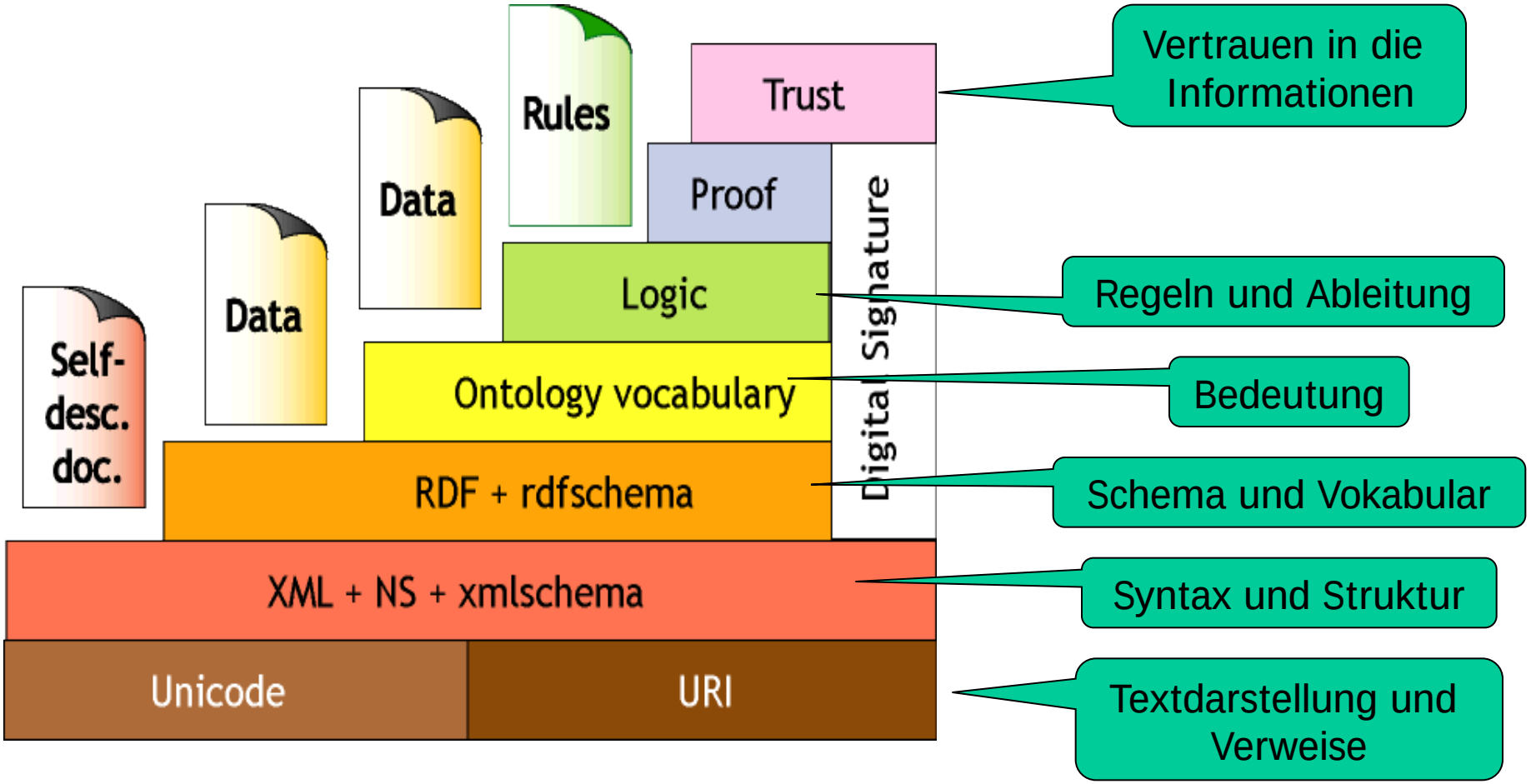
Grundprinzipien „Semantic Web“

- Semantik wird durch Annotation und Verweise spezifiziert
- Uniform Resource Identifier (URI)
  - Sage nicht „farbe“, sage "http://www.pantomime.com/std6#farbe"
  - Definitionen können im Laufe der Zeit ergänzt werden
  - Definitionen können jederzeit verändert oder ersetzt werden
- Keine Erzwingung von Konsistenz oder Kontinuität
  - Ist „http://www.pantomime.com/std6#farbe“ noch dasselbe?
  - Ist „http://www.pantomime.com/std6#farbe“ dasselbe wie „http://www.colors.com/colors“?
  - „Jeder kann Beliebiges über Beliebiges sagen"
- Sehr lose Koppelung
  - Kein einzelnes System weiß alles
  - Hochgradig dezentrales Design
  - Preis: Keine Sicherstellung von Konsistenz

Inhalt dieser Vorlesung

- Grundidee des Semantic Web
- Layer Cake
- RDF und RDFS
- SparQL
- Die OWL Sprachfamilie

Semantic Web „Layer Cake“



Quelle: [Hen02]

1+2. Unicode, URI und XML

- Unicode / URI
  - Semantikfrei
  - Unicode: Standard zur binären Repräsentation von Zeichen
  - URI: Identifikation von virtuellen oder physischen Ressourcen
  - URI ist ein Schlüssel
- XML / Namespaces / XML Schema
  - Standard zur Darstellung strukturierter Daten
  - Mit geeigneter Kodierung auch für nicht-hierarchische Daten
    - Serialisierung von Daten in XML
    - Z.B. für RDF, OWL, Relationen, ...
  - Definition von Schemata (inkl. Datentypen) für Daten
  - Austauschschema, nicht notwendigerweise Daten-/ Speicherschema

3. Resource Description Framework

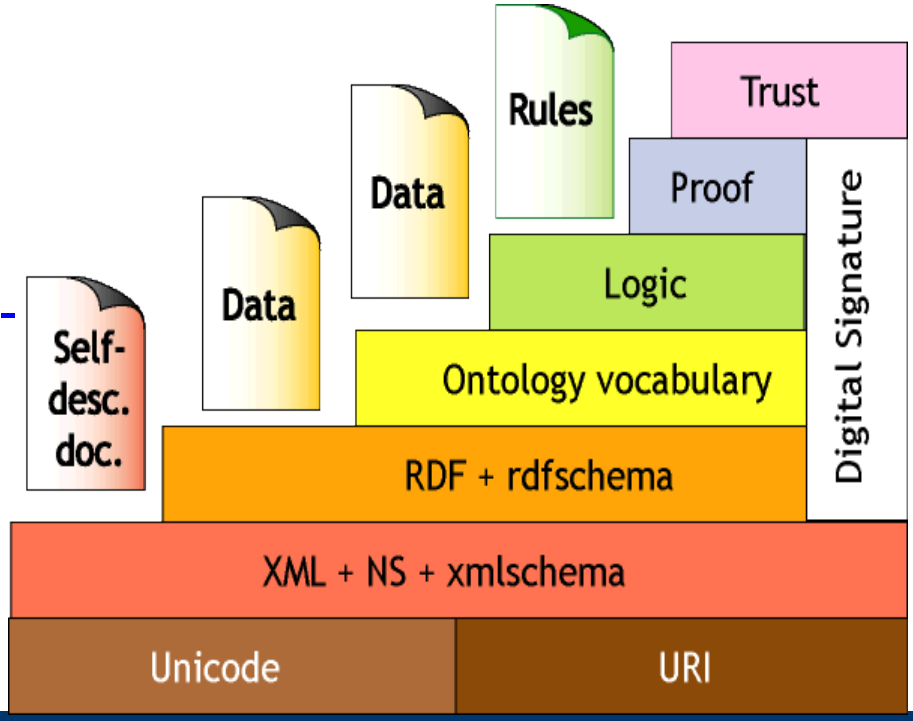
- Graphbasiertes Datenmodell
- Informationen werden als Tripel modelliert
  - (Subjekt Prädikat Objekt)
  - Beschreibung der Werte von Eigenschaften von Ressourcen
  - Erlaubt Aussagen über alles mögliche
    - Insbesondere auch über andere Aussagen -> später
- RDF Datenbasis = Menge von Tripeln
- RDFS („RDF Schema“)
  - Festlegung eines Vokabulars für RDF Datenbasen
  - Typisierung von RDF Daten
    - Datentypen, Spezialisierung, getypte Beziehungen, ...
- Gleich mehr dazu

4. Ontology Vocabulary

- Konzeptualisierung von Domänen
- Ressourcen erhalten Bedeutung durch Einbettung in eine formale Ontologie
- Interoperabilität von RDF Datenbasen durch Verwendung derselben Ontologie
  - Beziehungen von Konzepten über Datenbasis-Grenzen hinweg
  - Bei verschiedenen Ontologien: Ontologie-Alignment
- Alles freiwillig, lose Kopplung
  - RDF Dokumente müssen nicht zu RDFS-Definitionen konform sein
  - Konzepte müssen nicht durch Ontologien untermauert werden
  - Alles kann sich ständig ändern

5-7. Logic, Proof, Trust

- Klingt gut, aber ...
  - Keine klare Aufteilung
  - Logic: Inferenz einer Wissensrepräsentationssprache wie OWL
  - Proof: ?
- Trust
  - Maßnahmen zur Beurteilung des Vertrauens in Daten und Schlüsse
  - Schutz vor Spam-Seiten, Spam-RDF-Datenbanken, Spam-Ontologien
  - Sehr schwierige Umsetzung



Inhalt dieser Vorlesung

- Grundidee des Semantic Web
- Layer Cake
- RDF und RDFS
- SparQL
- Die OWL Sprachfamilie

RDF Grundlagen

- Grundlegendes Element sind Aussagen bestehend aus
  - Ressource (Subjekt)
  - Eigenschaft (Prädikat)
  - Ressource (Objekt)
- Beispiel: „Hitchcock ist der Regisseur von Marnie“
  - RDF-Tripel: **(Hitchcock, www.duden.de/regisseur, Marnie)**
  - Serialisiert in XML

```
<?xml version="1.0">
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#">
   <rdf:Description rdf:about="Alfred Hitchcock">
       <ist_regisseur_von> Marnie </ist_regisseur_von>
   </rdf:Description>
</rdf:RDF>
```

  - Prädikatenlogik: **istRegisseur(Hitchcock, Marnie)**

Mehr-arige Prädikate

- RDF Tripel können nur binäre Prädikate ausdrücken
- Für ternäre, quartäre, ... Prädikate müssen „künstliche“ Ressourcen erschaffen werden
  - Ein „Schlüssel“ pro Tupel
  - Können als Blank Nodes realisiert werden (später)
  - Im Grunde ist das Skolemisierung
- „Hitchcock hat 1964 den Film Marnie gedreht“
  - Neue künstliche Ressource **`MarnieFilm`**

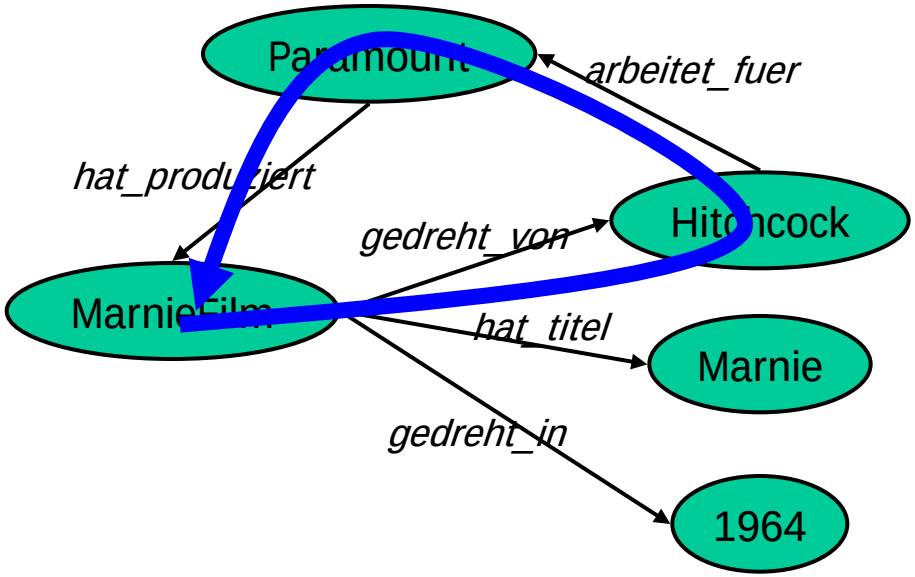
```
(MarnieFilm, gedreht_von, Hitchcock)
(MarnieFilm, hat_titel, Marnie)
(MarnieFilm, gedreht_in, 1964)
```

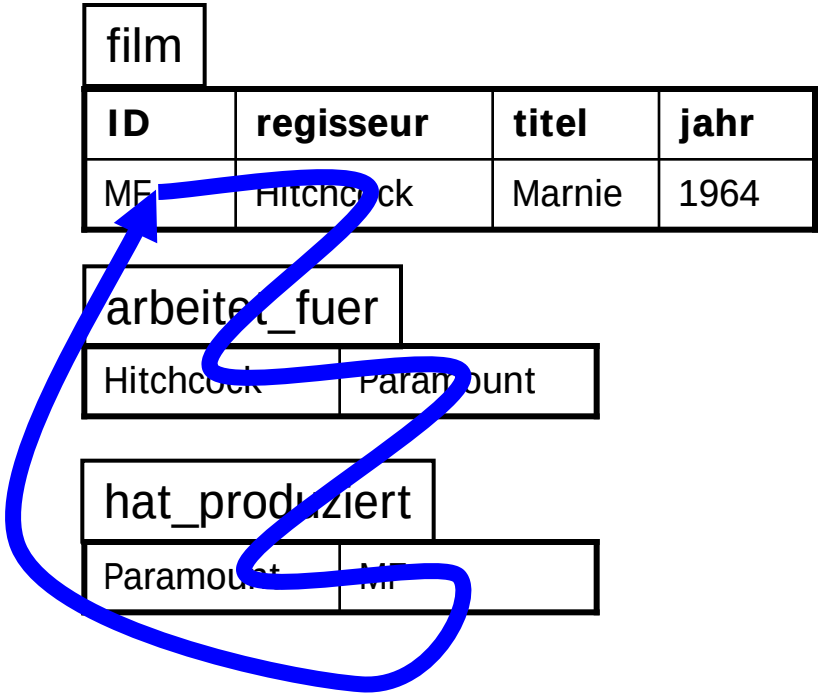
RDF als Graph

- Graphen als natürliche Repräsentation von RDF
- Naiver Ansatz (so stand es in den W3C Dokumenten)
  - Subjekte und Objekte werden Knoten
  - Prädikate sind Kanten
- Eigenschaften des Graphen
  - (Alle) Knoten haben eindeutige Label (URI oder Literal)
  - Kanten sind gerichtet und haben keine eindeutigen Label
  - Knoten können durch mehr als einer Kante verbunden sein (Multigraph)

RDF als Graph

```
(MarnieFilm gedreht_von Hitchcock)
(MarnieFilm hat_titel Marnie)
(MarnieFilm gedreht_in 1964)
(Paramount hat_produziert MarnieFilm)
(Hitchcock arbeitet_für Paramount)
```

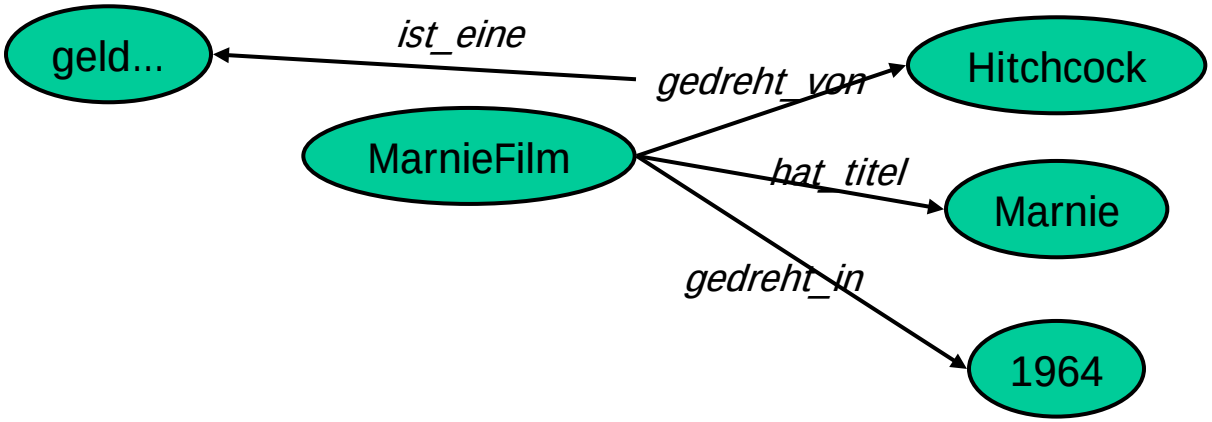




Problem

- Auch Prädikate sind Ressourcen
- Über die kann man Aussagen machen
- Erfordert Kanten von/auf Kantenlabel

```
(MarnieFilm gedreht_von Hitchcock)
(MarnieFilm hat_titel Marnie)
(MarnieFilm gedreht_in 1964)
(gedreht_von ist_eine gelderwerbstätigkeit)
```

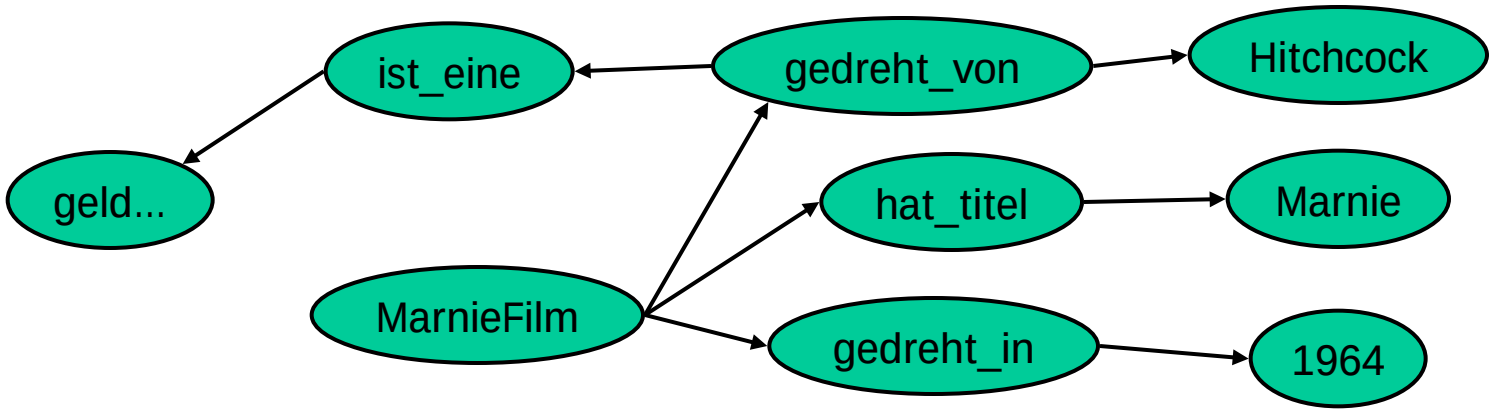


Warum ist **`gedreht_von`** kein Knoten?

Problem

- Auch Prädikate sind Ressourcen
- Über die kann man Aussagen machen
- Erfordert Kanten von/auf Kantenlabel

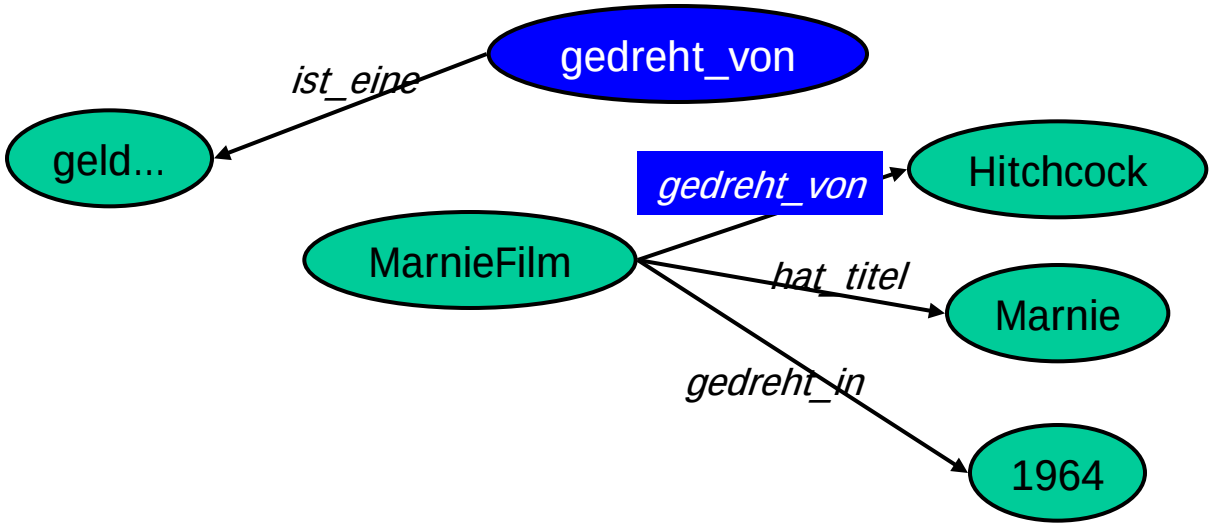
```
(MarnieFilm gedreht_von Hitchcock)
(MarnieFilm hat_titel Marnie)
(MarnieFilm gedreht_in 1964)
(gedreht_von ist_eine gelderwerbstätigkeit)
```



Wo sind unsere Aussagen?

Offizielle Version?

- Unklar [W3C Spezifikationen]
  - „The nodes of an RDF graph are its subjects and objects.”
  - „A URI reference or literal used as a node identifies what that node represents. A URI reference used as a predicate identifies a relationship between the things represented by the nodes it connects. A predicate URI reference may also be a node in the graph.”

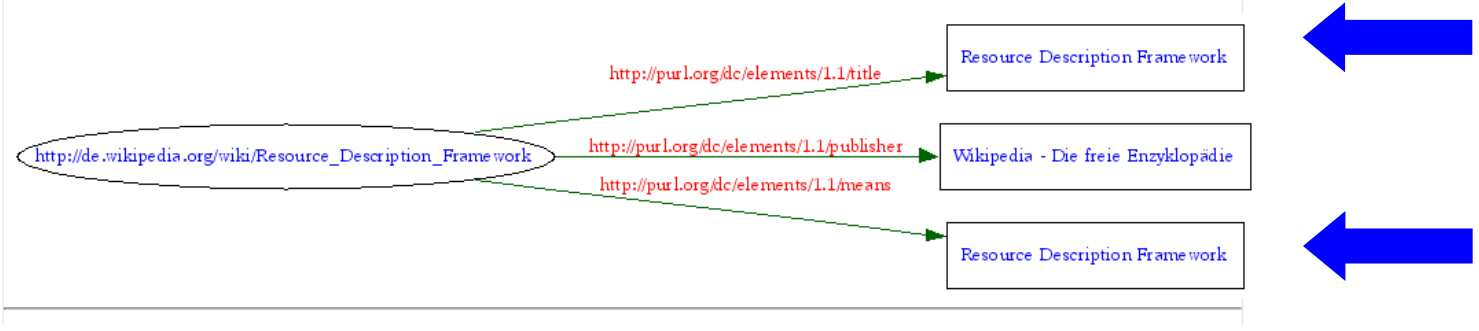


Eindeutigkeit von Labels

- Unklar
  - „An *RDF graph* is a set of RDF triples.
  - The set of *nodes* of an RDF graph is the set of subjects and objects of triples in the graph.“
  - Aber: W3C‘s RDF Validator

```
1: <?xml version="1.0" encoding="UTF-8" ?>
2: <rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
3:          xmlns:dc="http://purl.org/dc/elements/1.1/">
4:   <rdf:Description rdf:about="http://de.wikipedia.org/wiki/Resource_Description_Framework">
5:     <dc:title>Resource Description Framework</dc:title>
6:     <dc:publisher>Wikipedia - Die freie Enzyklopädie</dc:publisher>
7:   </rdf:Description>
8:   <rdf:Description rdf:about="http://de.wikipedia.org/wiki/Resource_Description_Framework">
9:     <dc:means>Resource Description Framework</dc:means>
10:  </rdf:Description>
11: </rdf:RDF>
```

Graph of the data model

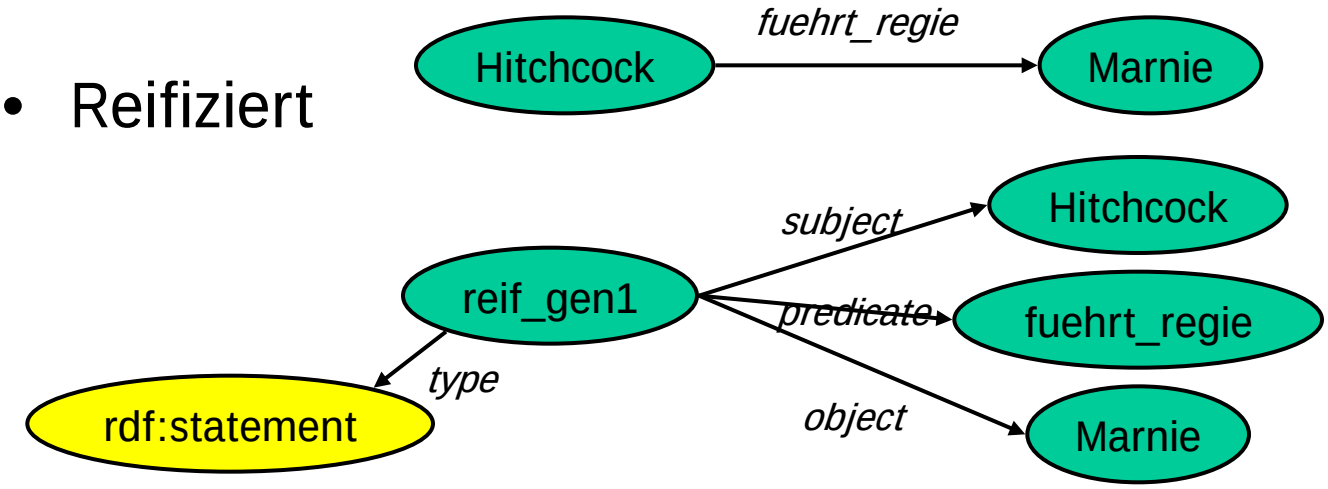


Aussagen über Aussagen

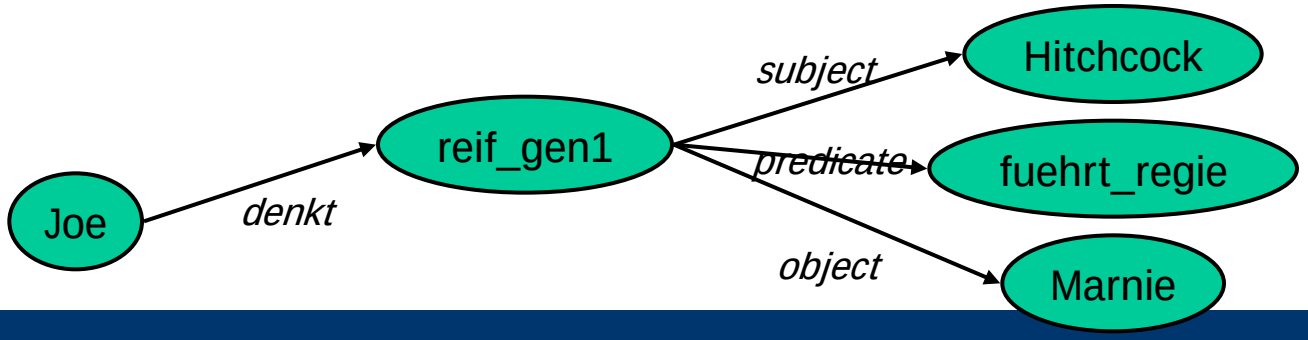
- In RDF kann man Aussagen über Aussagen machen
  - Hitchcock ist der Regisseur von Marnie
  - Joe denkt, dass Hitchcock der Regisseur von Marnie ist
- Um eine Aussage über eine Aussage X machen zu können, muss man X reifizieren
  - Reification = „Verdinglichung“
  - Eine Aussage wird als Ressource behandelt
- Vorgehen für Aussage (S P O)
  - Man schafft einen neue Ressource R vom Typ **`RDF:Statement`**
  - Drei Aussagen: **`(R subject S) (R predicate P) (R object O)`**
  - R kann normal verwendet werden: **`(Joe denkt R)`**

Grafisch

- „Hitchcock ist der Regisseur von Marnie“
- Reifiziert

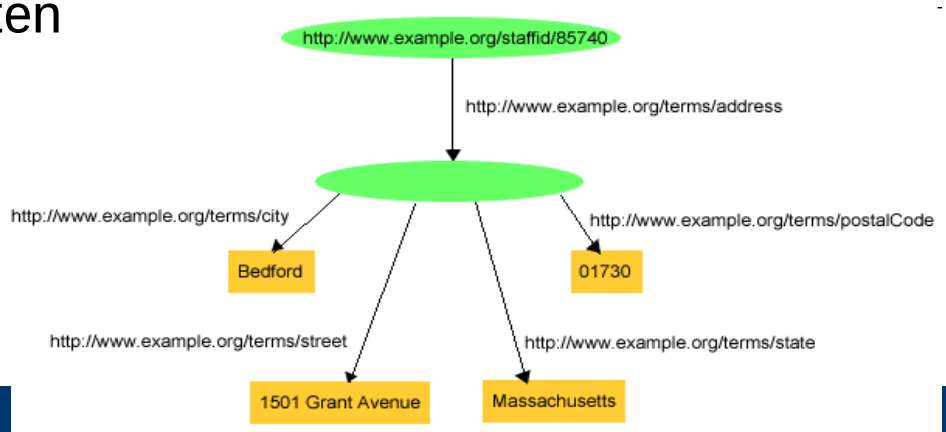


- „Joe denkt, dass Hitchcock der Regisseur von Marnie ist“



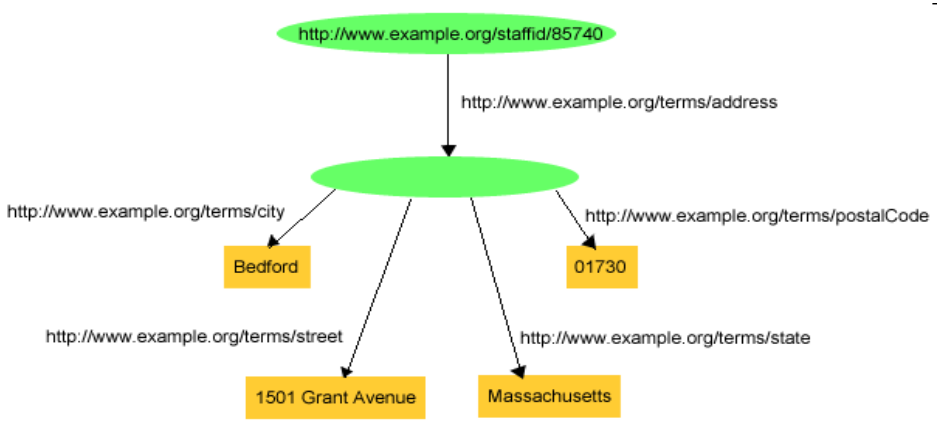
Weitere RDF Konzepte

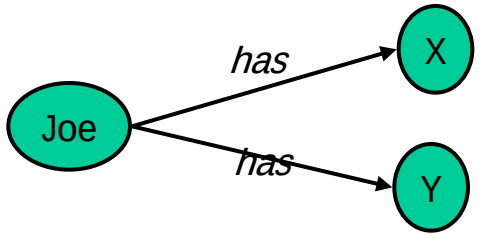
- Aussagen über Mengen von Ressourcen
  - Gruppierung von Aussagen in Bags und (geordneten) Sequenzen
- Ressourcen können einen Typ haben
  - Wird im Allgemeinen nicht weiter interpretiert
  - Spezielle interpretierte Typen wie rdf:statement, rdf:bag, ...
- „Blank Nodes“ – oft ignoriert
  - Statt URIs und Literalen können auch „Blank Nodes“ als Subjekt und Objekt verwendet werden
  - Bringt viele Schwierigkeiten
  - Gutartige Verwendung – n-arige Prädikate
  - Beispiel: Adresse mit 5 Elementen

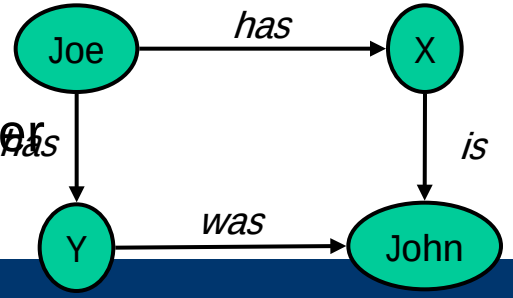


Probleme mit Blank Nodes

- Blank Nodes identifizieren sich nur über ihre Umgebung
- Sie nehmen praktisch die Rolle von Variablen ein
  - **(Peter address ?)**
  - **(? City Bedford)**
  - **(? Street „Grant Avenue“)**
  - **(…)**
  - **(Paul address ?)**
- Besser
  - **(Peter address X)**
  - **(X City Bedford)**
  - **(X Street „Grant Avenue“)**
  - **(…)**
  - **(Paul address Y)**
- RDF-Graphen nicht eindeutig
  - Verschiedene Graphen – Ergebnisse aller möglichen Anfragen sind gleich







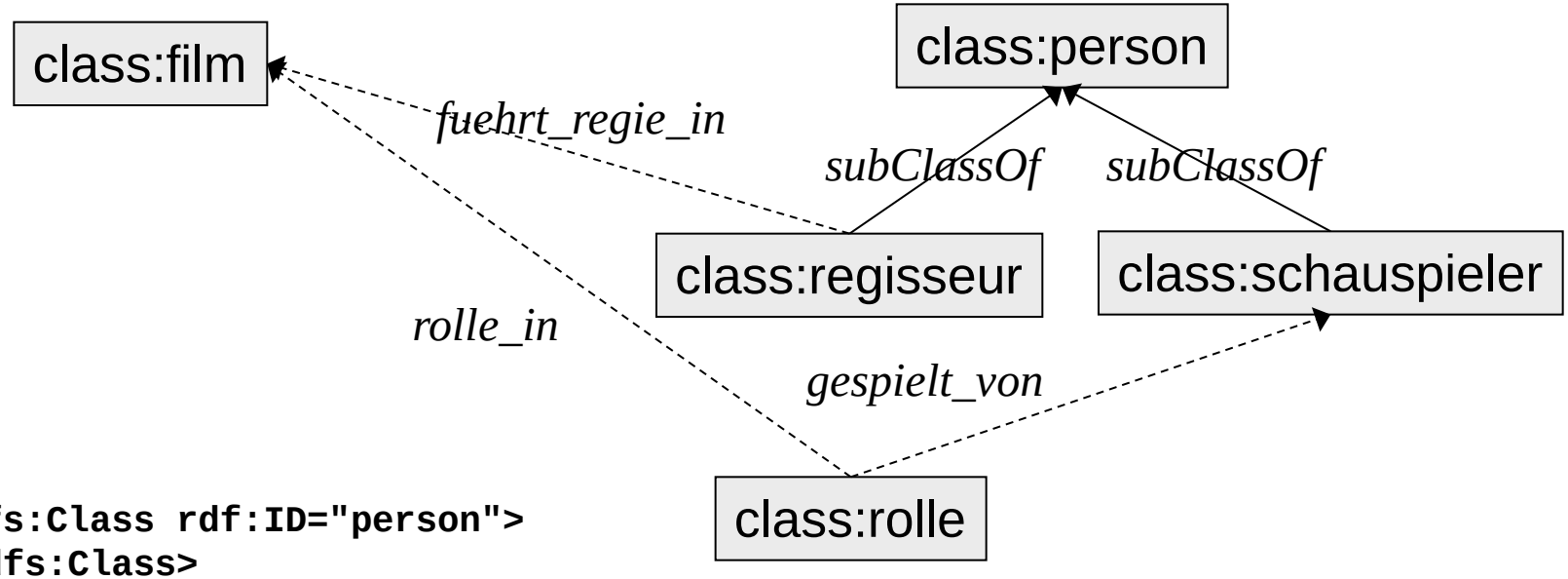
Bewertung von RDF

- Sehr flexibles Datenmodell
  - Keine Trennung von Dingen und Beziehungen
  - Blank nodes für n-arige Relationen
    - Sagen nur: Es existiert ein Knoten, der ... (aber geben keine Identifikation)
    - RDF Graphen sind eine Mischung aus Extension und Intension
  - Reifikation
    - Kann man nicht auf Aussagen in PL-I abbilden
- Für Massendaten nicht geeignet (und nicht gedacht)
- Geeignet für heterogene, schwach strukturierte und wissensintensive Anwendungen
  - Format zur Definition Semantischer Netze

RDFS – Schemata für RDF

- Eine RDF Datenbasis ist vollkommen frei in den verwendeten Begriffen
  - Kein Schema – das erschwert die Analyse
- RDFS
  - RDFS: RDF Vocabulary Description Language
  - Spezifikation von
    - Typen von Ressourcen (rdfs:class)
    - Subtypbeziehungen zwischen Typen (rdfs:subClassOf)
    - Eigenschaften eines Typen (rdfs:property)
    - Erlaubte Typen in Subjekt und Objekt von Prädikaten (rdfs:domain, rdfs:range)
    - ...
- Damit: Inferenz in Typhierarchien

Filmontologie in RDFS




```
<rdfs:Class rdf:ID="person">
</rdfs:Class>
<rdfs:Class rdf:ID="regisseur">
  <rdfs:subClassOf rdf:resource="#person"/>
</rdfs:Class>
…
<rdfs:Property rdf:ID="fuehrtRegieIn">
  <rdfs:domain rdf:resource="#regisser"/>
  <rdfs:range rdf:resource="#film"/>
</rdfs:Property>
…
```

Einschätzung RDFS

- RDFS ist relativ nahe an klassischen objektorientierten Modellen
  - Und eher weiter weg von Description Logics
- Man kann z.B. nicht
  - Klassen auf Basis anderer Klassen definieren
    - Union, Schnitt, Komplement, ...
  - Eigenschaften von Eigenschaften definieren
    - Transitivität, Symmetrie, ...
- Dafür gibt es eine keine Trennung zwischen Modell und Metamodell
  - Eingebaute Sprachelemente können redefiniert bzw. erweitert werden
    - Z.B.: Range oder domain von rdfs:subClassOf einschränken

RDFS versus XML

- Konformität mit einem Schema
  - RDFS definiert Klassen und deren Beziehungen
  - Die können in RDF benutzt werden
  - Man kann aber auch beliebige andere (undefinierte) benutzen
- Subklassenbeziehung versus Schachtelung
  - Schachtelung ist ein syntaktisches Konstrukt
  - Bezeichnet eine 1:n Beziehung, über die man nichts weiter weiß
  - Subklassen erben Eigenschaften der Superklasse und sind extensional eingeschlossen

Inhalt dieser Vorlesung

- Grundidee des Semantic Web
- Layer Cake
- RDF und RDFS
- SparQL
- Die OWL Sprachfamilie

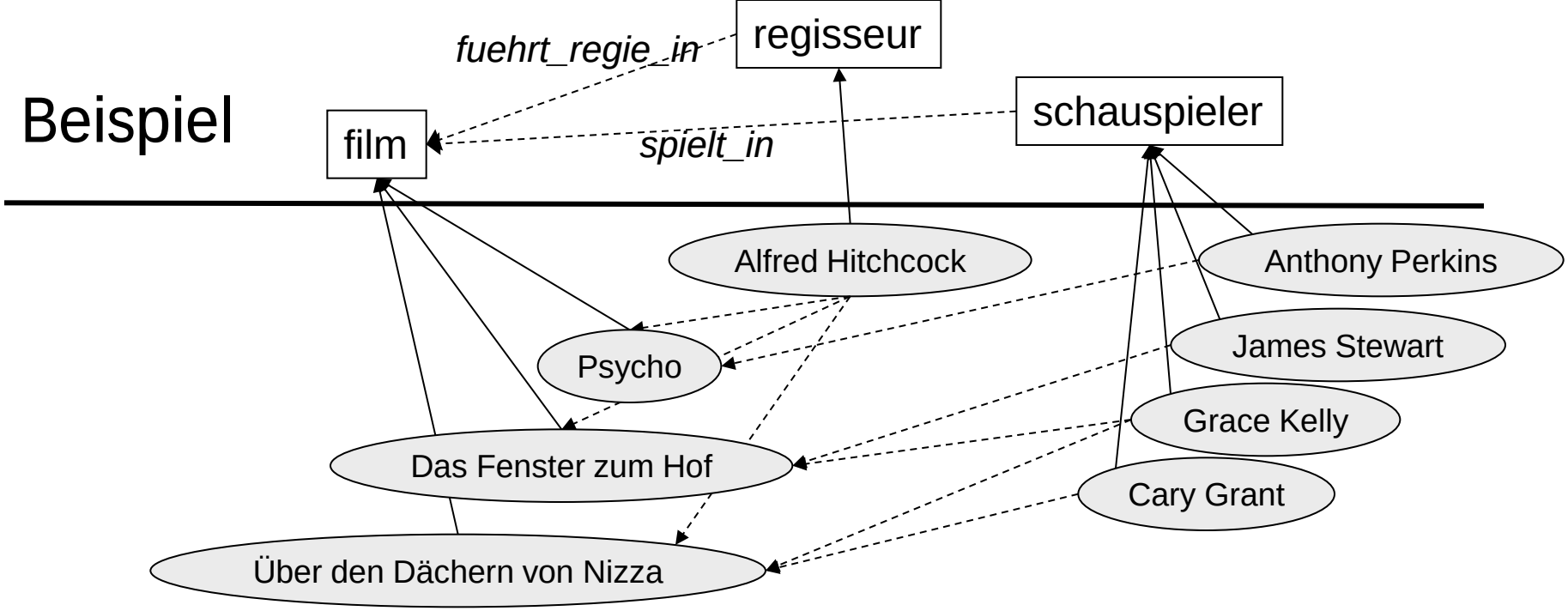
Anfragesprachen für RDF

- Historisch sind eine ganze Reihe von Sprachen entstanden
  - RQL, RDQL, SesamQL, ...
- W3C standard: SPARQL
  - ***S**PARQL **P**rotocol **a**nd **R**DF **Q**uery **L**anguage*
  - Wir behandeln nur den „QL“ Teil
  - Eine SPARQL Anfrage Q ist im Kern ein Graphmuster aus Knoten und Kanten, beschriftet mit Konstanten oder Variablen
  - Q auf RDF-Datenbasis D: Alle zu Q isomorphen Subgraphen
- Grundkonzept: Anfragetripel (X Y Z)
  - X,Y,Z können Literale/URI‘s oder Variable sein
  - Anfragetripel Q matched ein RDF-Datentripel R, wenn es eine Funktion s gibt, die Konstante auf Konstante und Variable auf Konstante abbildet und für die gilt: s(Q)=R

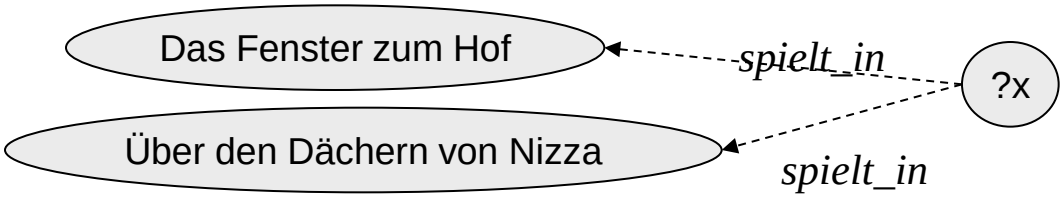
SPARQL Grundaufbau

- Eine einfache SPARQL Anfrage ist eine Menge von Anfragetripeln
- Semantik
  - Anfragetripel werden an alle matchenden Datentripel gebunden
  - Bilde kartesisches Produkt aller Bindungen
  - Streiche alle Elemente, in denen eine Variable an verschiedene Werte gebunden wird
    - Gleiche Variable in verschiedenen Tripeln erzeugen also Joins
  - Alle übrigen Elemente bilden das Result Set der Anfrage

Beispiel



- **SELECT ?X**
  WHERE (?X spielt_in „Über den Dächern von Nizza“
  ?X spielt_in „Das Fenster zum Hof“)
- Als Graph



- Berechnet: „Grace Kelly“

Erweiterungen

- WHERE Klausel
  - Optionals: Optionale Tripel
    - Wichtig wegen der fehlenden Strukturierung von RDF Daten
    - Vergleichbar Outer-Join (bzw. der Union von zwei Anfragen)
  - Filter für Wertebedingungen (=, <, >, REGEXP, ...)
  - Union: Logisches ODER
- SELECT Klausel
  - Ausgabe von Variablenbindungen oder Tripelmengen
  - Sortierung der Ergebnisse
- FROM Klausel
  - Implizite Annahme einer Default RDF Datenbasis
  - Named Graphs – Queries über Tripel verschiedener Datenbasen

SparQL/RDF und Informationsintegration

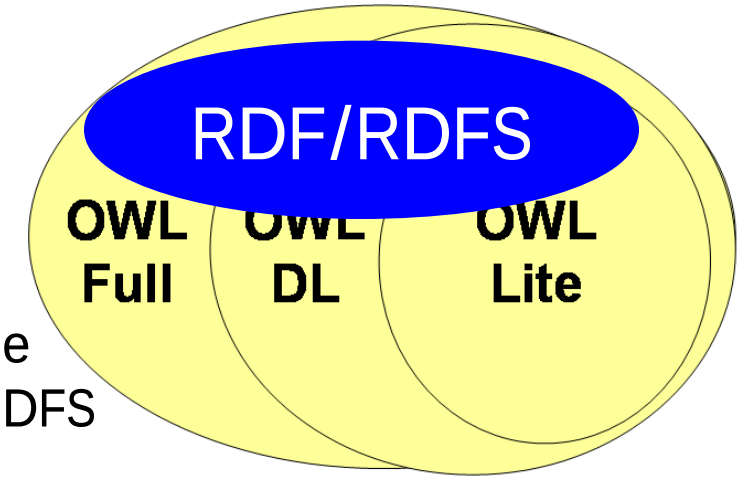
- Sehr flexibles Datenmodell
  - Subsumiert Tabellen (relational) und Bäume (XML)
  - Basiert auf RDF, nicht auf RDFS (oder OWL)
- Named Graphs
  - Verknüpfung mehrerer Datenquellen in einer Anfrage (Multi-DB-sprache)
  - Keine Vorkehrungen zur prädikatbasierten Identifikation von Quellen
- Schematische Heterogenität gibt es nicht (Graph)
- Optional und Union: Für strukturell heterogene Daten
- Keine Gruppierung oder Aggregation
  - Schlecht für Duplikaterkennung und Fusion
- Tools vorhanden

Inhalt dieser Vorlesung

- Grundidee des Semantic Web
- Layer Cake
- RDF und RDFS
- SparQL
- Die OWL Sprachfamilie

Ontology Web Language: OWL

- Historisch: DAML + OIL → DAML+OIL → OWL
- Formal basierend auf der DL SHIQ
  - DL mit transitiven und inversen Rolleneigenschaften sowie numerischen Kardinalitätseinschränkungen
- „Eigentlich“ fußt OWL auf RDF und RDFS
  - Erweiterung um Inferenz auch außerhalb der subClass-Beziehung
  - Aber nicht durchgehalten
- Drei Stufen
  - OWL Lite – Einfache Sprache für Taxonomien plus Constraints
  - OWL DL – Subsumption entscheidbar
  - OWL Full – Unentscheidbar, als einzige Sprache abwärtskompatibel zu RDF/RDFS



OWL Lite

- Trennung von Klassen, Werten und Instanzen
  - Die macht RDFS nicht
- Großteil der RDFS-Elemente
  - class, subClassOf, property, range, domain, ...
- Verhältnisse von Konzepten (class) und Rollen (property)
  - Zwischen Klassen: intersectionOf
  - Zwischen Klassen oder zwischen Beziehungen: equivalent
  - Zwischen Objekten: sameAs, differentFrom
- Eigenschaften von Rollen
  - inverseOf, transitive, symmetric, functional
- Rolleneinschränkungen
  - allValuesOf, someValuesOf, max/minCardinality (0 oder 1)
- Es fehlen z.B. $\sqcup$, $\neg$

OWL Full

- Abwärtskompatibel zu RDF
- Vermischung von Klassen, Instanzen, Rollen und Werten
  - Klassen können Instanzen anderer Klassen sein
  - Das wird von RDFS „geerbt“
- Weitere Sprachelemente
  - disjointWith: Schnitt zweier Klassen muss leer sein
  - unionOf, complementOf, intersectionOf für Klassen
  - ...
- Subsumption ist unentscheidbar
  - Man kann Antinomien formulieren: „Die Klasse K aller Dinge, die nicht zu K gehören“ (aus [HPvH03])

OWL DL

- Gegenüber OWL Full
  - Trennung von Klassen, Instanzen, Rollen und Werten
  - Diverse Einschränkungen
- Gegenüber OWL Lite
  - Neue Sprachelemente (Klassendefinitionen, Kardinalitäten, ...)
- Entscheidbare Sprache
  - „So ausdrucksstark wie gerade noch möglich“

Zusammenfassung Semantic Web

- Einige Technologien sind da, das Ziel bleibt Vision
- Stärke: Relativ geschlossener Framework zur Beschreibung komplexer Daten und Hintergrundwissen über diese Daten
- Hindernisse
  - Welche Benutzer können Sprachen wir OWL lernen und einsetzen?
  - Wer soll all die Ontologien schreiben?
  - Ontologieheterogenität als neue Art Heterogenität?
  - Welchen Vorteil bietet es für einen Webseitenbetreiber, Daten als RDF zu publizieren?
  - Wie integriert man 100 Millionen verteilte RDF Datenquellen?