

Informationsintegration

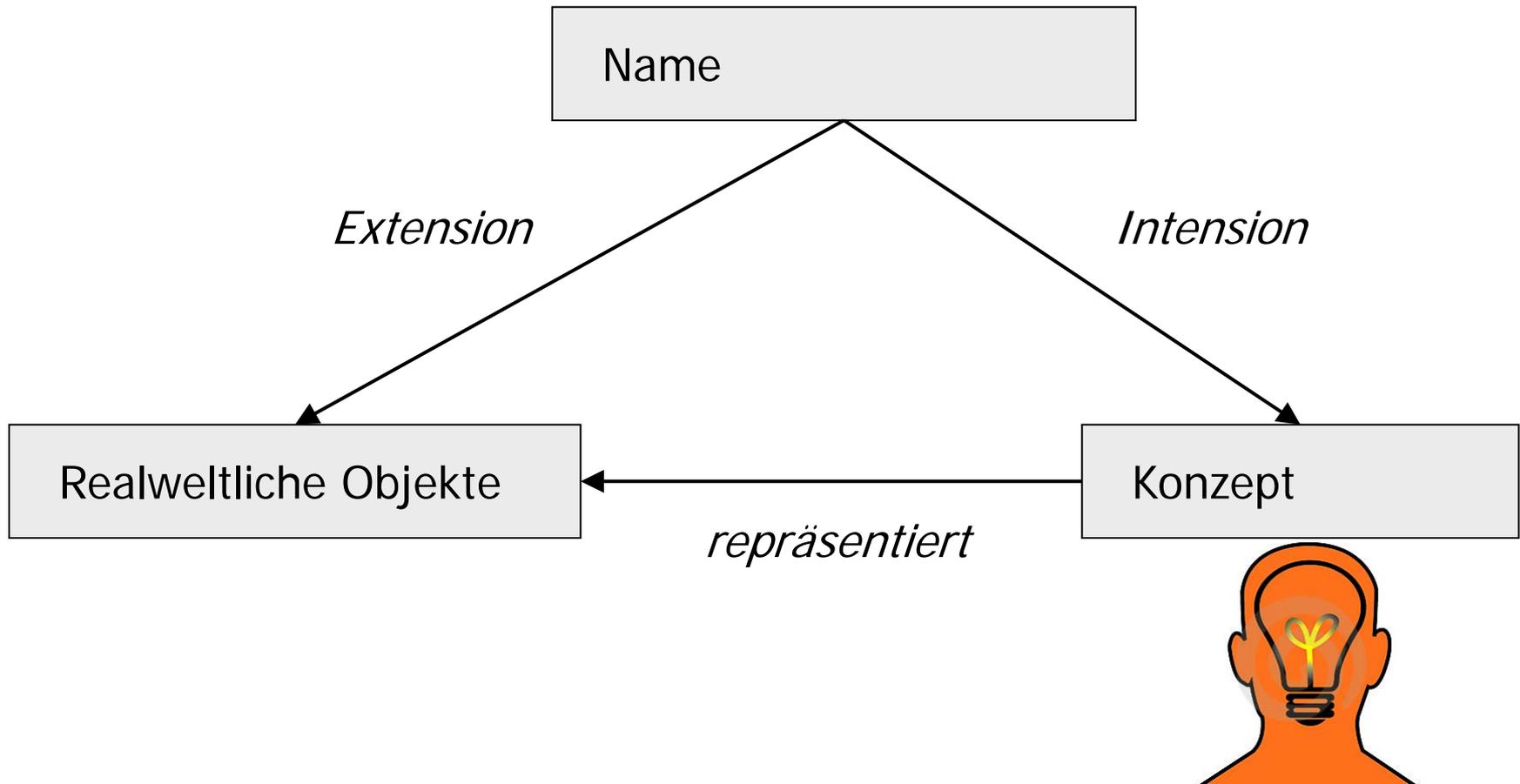
Semantische Integration

Ulf Leser

Inhalt dieser Vorlesung

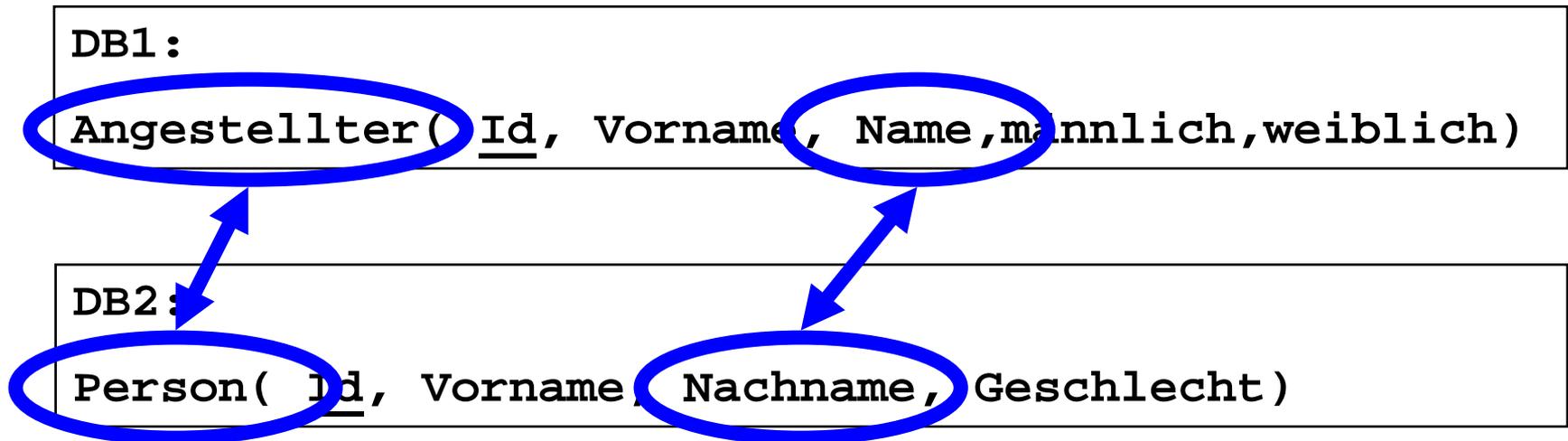
- Wdh: Semantische Heterogenität
- Semantische Integration
- Ontologien
- Beschreibungslogiken
- Ontologiebasierte Integration

Semantik von Namen



Synonyme

- Verschiedene Worte für „dasselbe“ Konzept
 - Immer im Kontext der Anwendung



Homonyme

- Gleiche Worte verschiedener Bedeutung
 - Treten oft bei Überschreitung von Domänengrenzen auf

Sekr., Sachbearbeiter,
Bereichsleiter, etc.

DB1:

Angestellter(Id, Vorname, Name, m, w, **Funktion**)

DB2:

Protein(Id, Sequenz, organismus, **Funktion** ...)

Transport, Katalyse, Signal, ...

Probleme

- Mögliche Beziehungen zwischen den Mengen realweltlicher Objekte, die Konzepte repräsentieren
 - $A=B$ (Äquivalenz): „semantische“ (echte) Synonyme
 - Kreditinstitut, Bank (?)
 - Gibt es echte Synonyme?
 - $A \subseteq B$ (Inklusion): B ist **Hyperonym** (Oberbegriff) zu A; A ist **Hyponym** zu B
 - Tochter \subseteq Kind
 - $A \cap B \neq \emptyset \wedge A \neq B$ (Überlappung): Schwierigster Fall
 - Küche-Kochnische; Haus-Gebäude; Regisseur-Schauspieler
 - $A \cap B = \emptyset$ (Disjunktheit): nicht verwandte Begriffe (häufigster Fall)
 - Dose-Lohnsteuerjahresausgleich

Semantische Integration

- Wie haben wir bisher semantische Heterogenität gelöst?

Semantische Integration

- Bisherige Lösung
 - Korrespondenzen definieren **semantische Beziehungen** zwischen Attributen, Relationen, Anfragen
 - Die kann man herleiten (Schema Matching) oder manuell festlegen
 - Lösung bisher auf **Schemaebene** beschränkt
 - Siehe Duplikaterkennung
- Die folgende Anfrageplanung ist semantikkfrei
 - Benutzt Äquivalenz- und Inklusionsbeziehungen

Software und Semantik

- Natürlichsprachige Webseite für eine Maschine

林克昌 根留台灣 可能增高

在愛戴者熱心奔走之下，華裔名指揮家林克昌根留台灣的可行性又提升了幾分。兩廳院主任李炎、國家音樂廳樂團副團長黃奕明日前親赴林克昌、石聖芳寓所拜會，並提出多場客席邀約。此外，台灣省立交響樂團團長陳澄雄也早早「下訂」，邀請林克昌赴台中霧峰，從八月十日起訓練省交，為期長達一個月。

在台灣諸多公家樂團中，陳澄雄是以實際行動表達對林克昌肯定的樂界人士之一，曾多次公開表示對林克昌指揮才華的欽佩，而且幾乎每個樂季都邀請林克昌客席演出。

此外，林克昌上個月赴俄羅斯與頂尖的「俄羅斯國家管絃樂團」灌錄了柴可夫斯基晚期三大交響曲以及「羅密歐與茱麗葉」、「斯拉夫進行曲」、「義大利隨想曲」，最後的DAT母帶也在前兩天寄回台灣。製作人楊忠衡與林克昌試聽之後，都對錄音效果—尤其音質表現感到相當滿意，楊忠衡估計呈現了七分林克昌指揮神韻。

俄羅斯國家管絃樂團首席布魯尼日前也讚譽林克昌的指揮藝術有三大特點：一是控制自如的彈性速度；二是強烈的動態對比；三是宛如呼吸歌唱的旋律處理。這些對錄音師而言都構成很大挑戰。俄國錄音師雖然採用多軌混音，但定位、場面都有可觀之處。

Quelle: [HenC

XML ≠ maschinenlesbare Bedeutung

- XML Dokument für eine Maschine

林克昌 根留台灣 可能增高

</U_>:

<&%GGTDR>

<NH&&%\$D>

<90(IU>

<4Rt5\$\$>

在愛戴者熱心奔走之下，華裔名指揮家林克昌根留台灣的可行性又提升了幾分。兩廳院主任李炎、國家音樂廳樂團副團長黃奕明日前親赴林克昌、石聖芳寓所拜會，並提出多場客席邀約。此外，台灣省立交響樂團團長陳澄雄也早早「下訂」，邀請林克昌赴台中霧峰，從八月十日起訓練省交，為期長達一個月。

在台灣諸多公家樂團中，陳澄雄是以實際行動表達對林克昌肯定的樂界人士之一，曾多次公開表示對林克昌指揮才華的欽佩，而且幾乎每個樂季都邀請林克昌客席演出。

此外，林克昌上個月赴俄羅斯與頂尖的「俄羅斯國家管絃樂團」灌錄了柴可夫斯基晚期三大交響曲以及「羅密歐與茱麗葉」、「斯拉夫進行曲」、「義大利隨想曲」，最後的DAT母帶也在前兩天寄回台灣。製作人楊忠衡與林克昌試聽之後，都對錄音效果—尤其音質表現感到相當滿意，楊忠衡估計呈現了七分林克昌指揮神韻。

俄羅斯國家管絃樂團首席布魯尼日前也讚譽林克昌的指揮藝術有三大特點：一是控制自如的彈性速度；二是強烈的動態對比；三是宛如呼吸歌唱的旋律處理。這些對錄音師而言都構成很大挑戰。俄國錄音師雖然採用多軌混音，但定位、場面都有可觀之處。

Quelle: [HenC

Schemata

Gleiche Schemata helfen, weil sie intensional gleiche Elementen festlegen

林克昌 橫留台灣 可能增加

< εδυσχαιον >

< ωορκ >

< πριωατ >

< ναμε >

< &%TT >

在歐戰熱心奔走之下，華裔指揮家林克昌橫留台灣的可行性又提升了幾分。兩院院士李炎、國家音樂廳樂團團長黃奕明日前親赴林克昌、石叻力黨所拜會，並提出多場客席邀約。此外，台灣省立交響樂團團長陳澄達也早早「下訂」，邀請林克昌赴台中露峰，從八月十日起加練五天，為期長達一個月。

在台灣諸多公家樂團中，陳澄達是以實際行動表達對林克昌肯定的樂界人士之一。曾多次公開表示對林克昌指揮才華的讚賞，而且幾乎每個樂季都邀請林克昌客席演出。

此外，林克昌上個月赴俄羅斯與頂尖的「俄羅斯國家管絃樂團」灌錄了柴可夫斯基時期三大交響曲以及「羅密歐與朱麗葉」、「斯壯次後行曲」、「義大利總協曲」。最後的DAT母帶也在前兩天寄回台灣。製作人楊忠康與林克昌試聽之後，都對錄音效果—尤其音樂表現感到相當滿意，楊忠康估計呈現了七分林克昌指揮神韻。

俄羅斯國家管絃樂團首席帝魯尼日前也讚譽林克昌的指揮藝術有三大特點：一是控制自如的彈性速度；二是強烈的動態對比；三是宛如呼吸歌過的旋律處理。這些對錄音師而言都構成重大挑戰。俄國錄音師雖然採用多軌混音，但定位、場面都有可觀之處。

林克昌 橫留台灣 可能增加

< εδυσχαιον >

< ωορκ >

< πριωατ >

< ναμε >

< &%TT >

在歐戰熱心奔走之下，華裔指揮家林克昌橫留台灣的可行性又提升了幾分。兩院院士李炎、國家音樂廳樂團團長黃奕明日前親赴林克昌、石叻力黨所拜會，並提出多場客席邀約。此外，台灣省立交響樂團團長陳澄達也早早「下訂」，邀請林克昌赴台中露峰，從八月十日起加練五天，為期長達一個月。

在台灣諸多公家樂團中，陳澄達是以實際行動表達對林克昌肯定的樂界人士之一。曾多次公開表示對林克昌指揮才華的讚賞，而且幾乎每個樂季都邀請林克昌客席演出。

此外，林克昌上個月赴俄羅斯與頂尖的「俄羅斯國家管絃樂團」灌錄了柴可夫斯基時期三大交響曲以及「羅密歐與朱麗葉」、「斯壯次後行曲」、「義大利總協曲」。最後的DAT母帶也在前兩天寄回台灣。製作人楊忠康與林克昌試聽之後，都對錄音效果—尤其音樂表現感到相當滿意，楊忠康估計呈現了七分林克昌指揮神韻。

俄羅斯國家管絃樂團首席帝魯尼日前也讚譽林克昌的指揮藝術有三大特點：一是控制自如的彈性速度；二是強烈的動態對比；三是宛如呼吸歌過的旋律處理。這些對錄音師而言都構成重大挑戰。俄國錄音師雖然採用多軌混音，但定位、場面都有可觀之處。

< &%TT >

Quelle: [Hen02]

Inhalt dieser Vorlesung

- Semantische Heterogenität
- **Semantische Integration**
- Ontologien
 - Einführung
 - Thesauri und semantische Netze
- Beschreibungslogiken
- Ontologiebasierte Integration

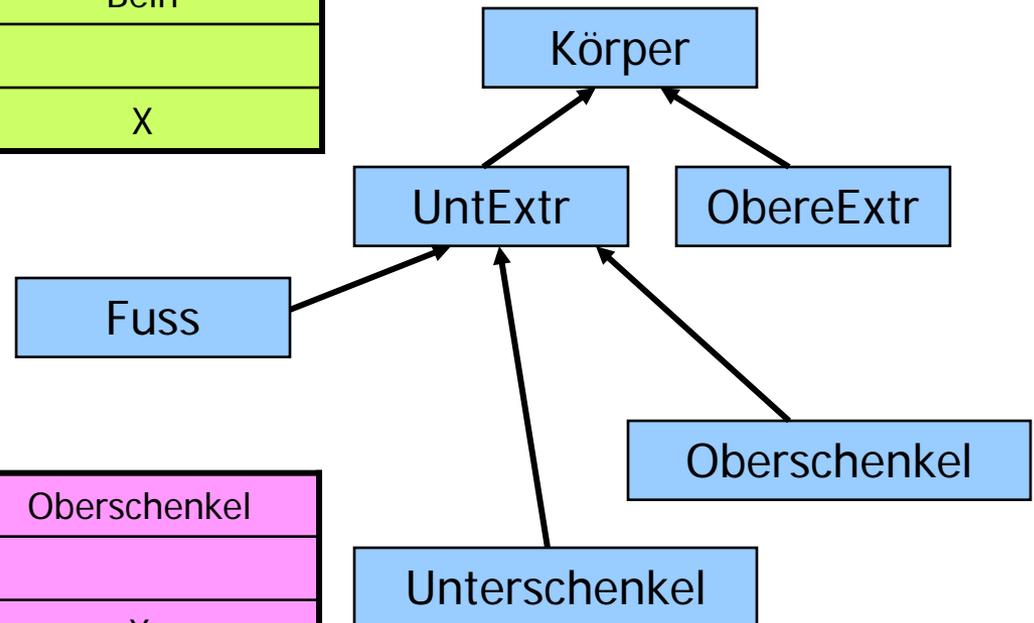
Andere Idee: Ontologien

- Ziel: Computer sollen **Schemata „verstehen“**
 - Dann könnten Korrespondenzen automatisch abgeleitet werden
- Problem: Computer „verstehen“ nicht
- **Ontologie-basierte Integration**
 - Logische Definition aller Begriffe in einer **Ontologie**
 - Verwendung einer speziell geeigneten **Beschreibungslogik**
 - Zur Beschreibung von Konzepten und deren Beziehungen zueinander
 - Beziehungen zwischen Konzepten (verschiedener Schemata) durch **logische Inferenz** herleiten

Ontologiebasierte Integration

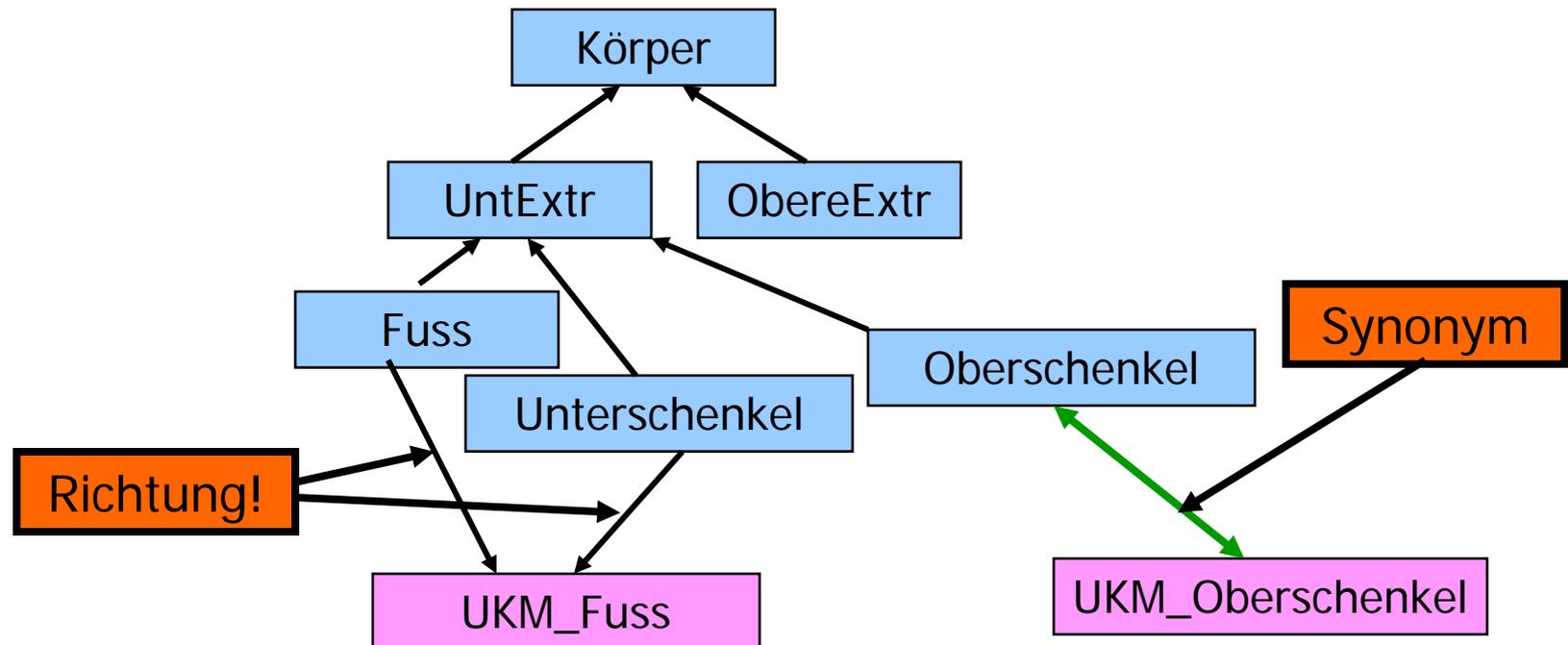
KK Hamburg	Fuss	Bein
Pat1	X	
Pat2		X

UK München	Fuss	Oberschenkel
Pat1	X	
Pat2		X



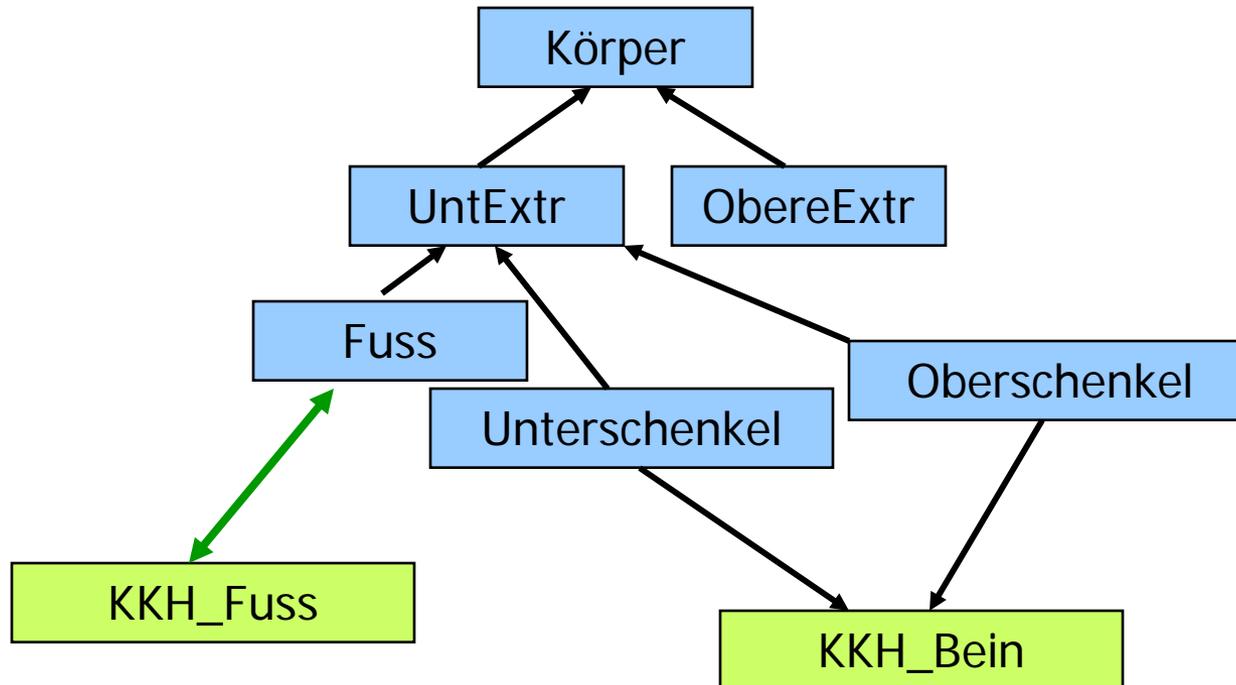
Quelle UK München

UK München	Fuss	Oberschenkel
Pat1	X	
Pat2		X

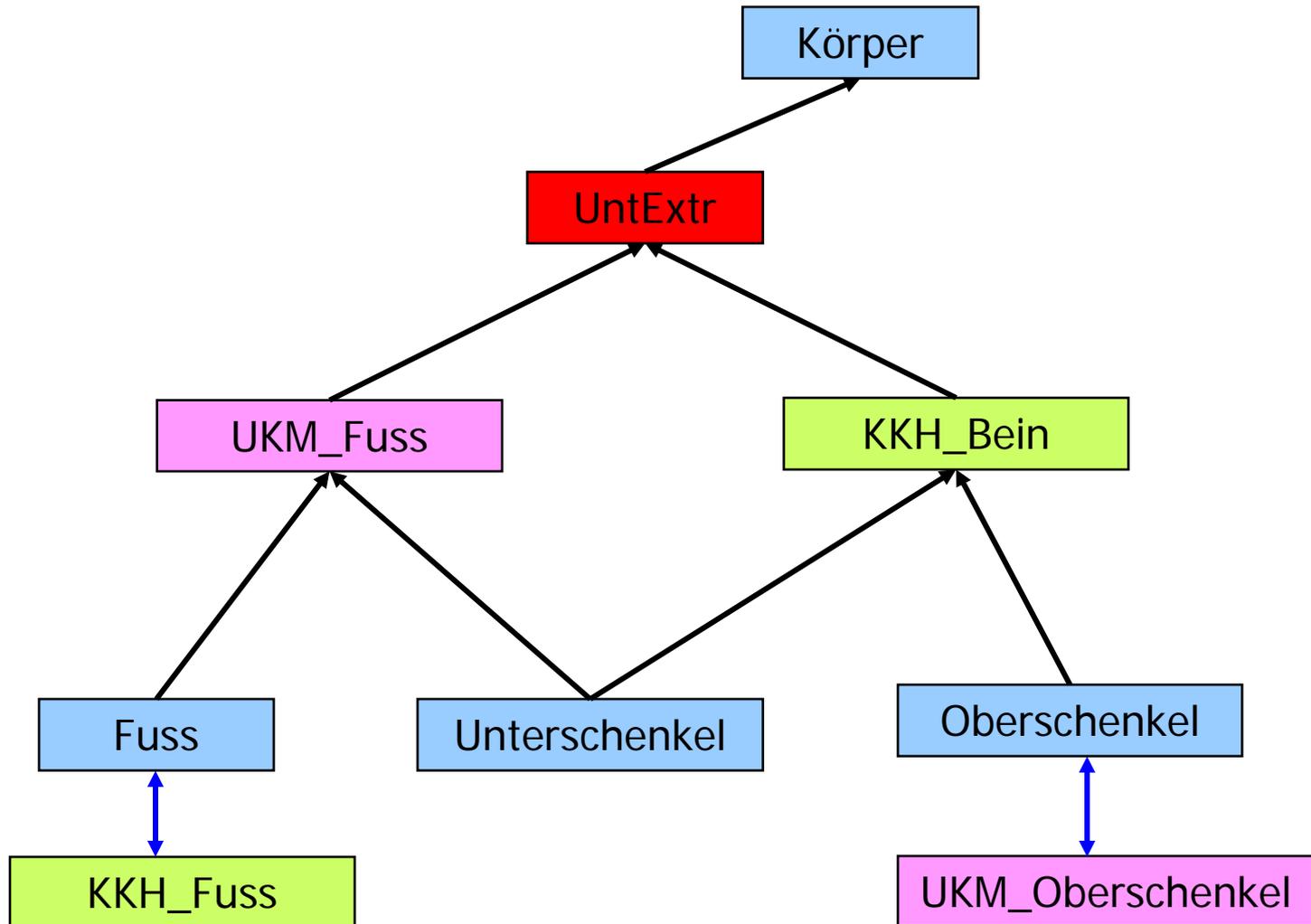


Quelle KK Hamburg

KK Hamburg	Fuss	Bein
Pat1	X	
Pat2		X

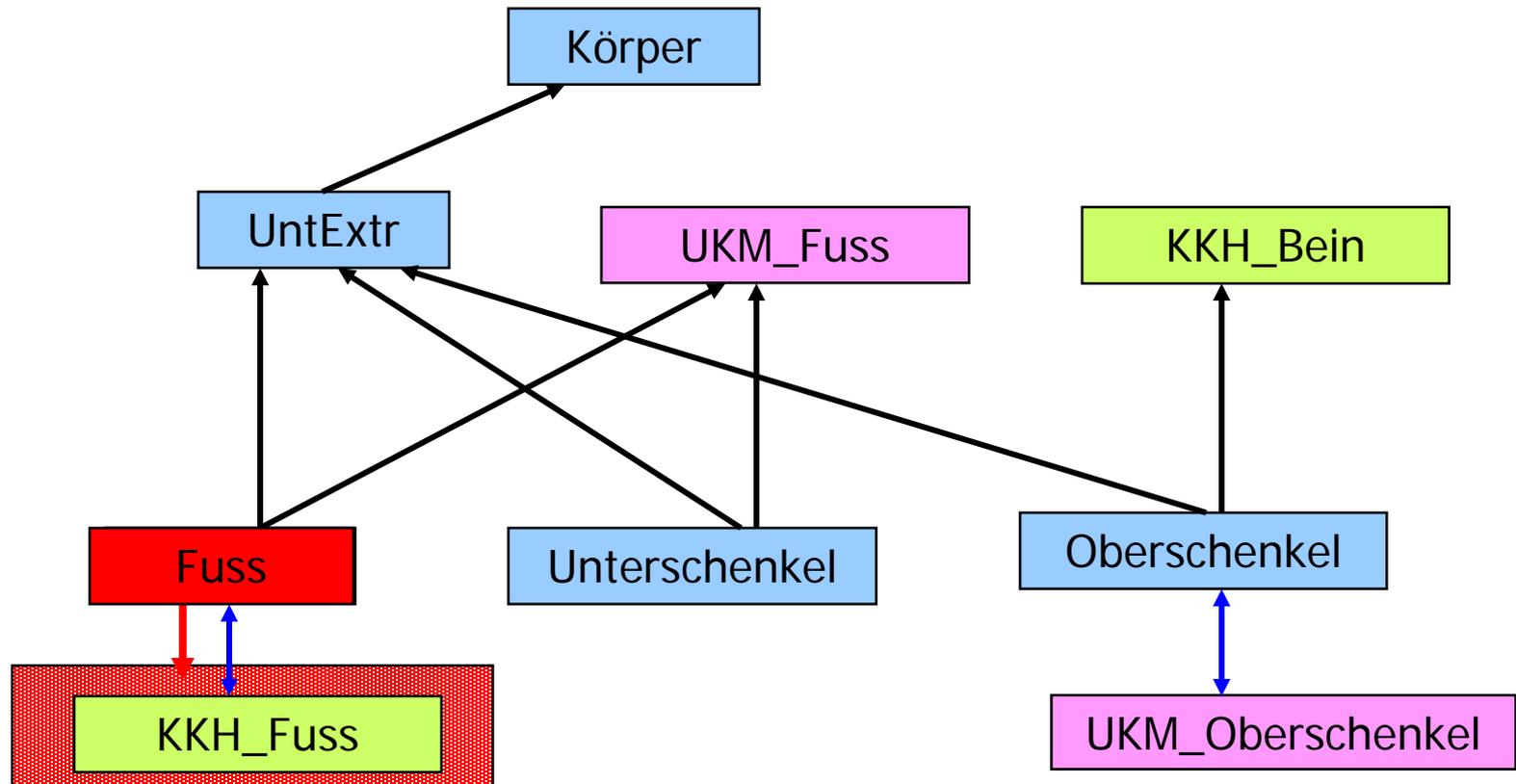


Bessere Modellierung



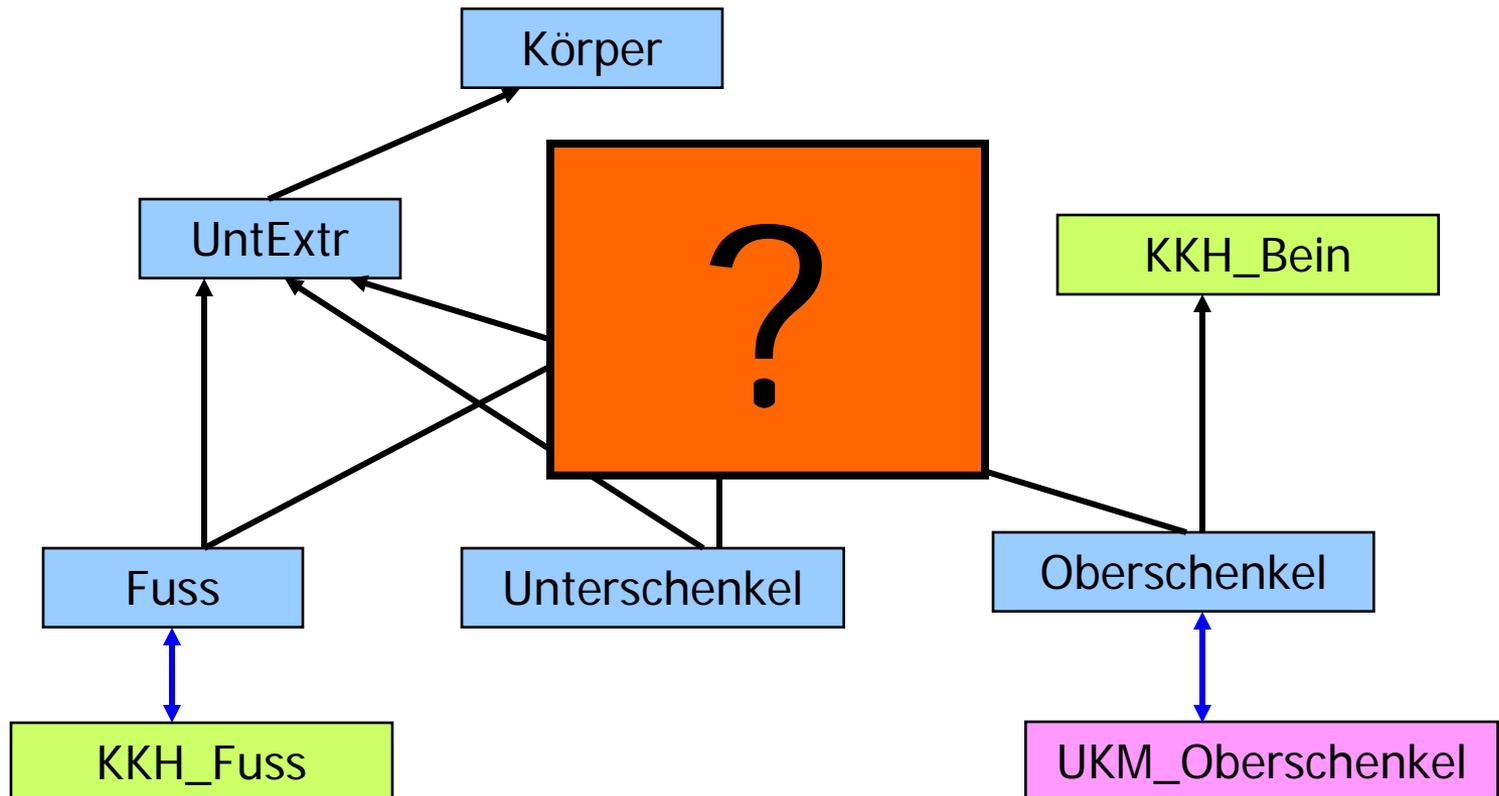
Anfrage 2

- „Fußverletzungen“



Anfrage 3

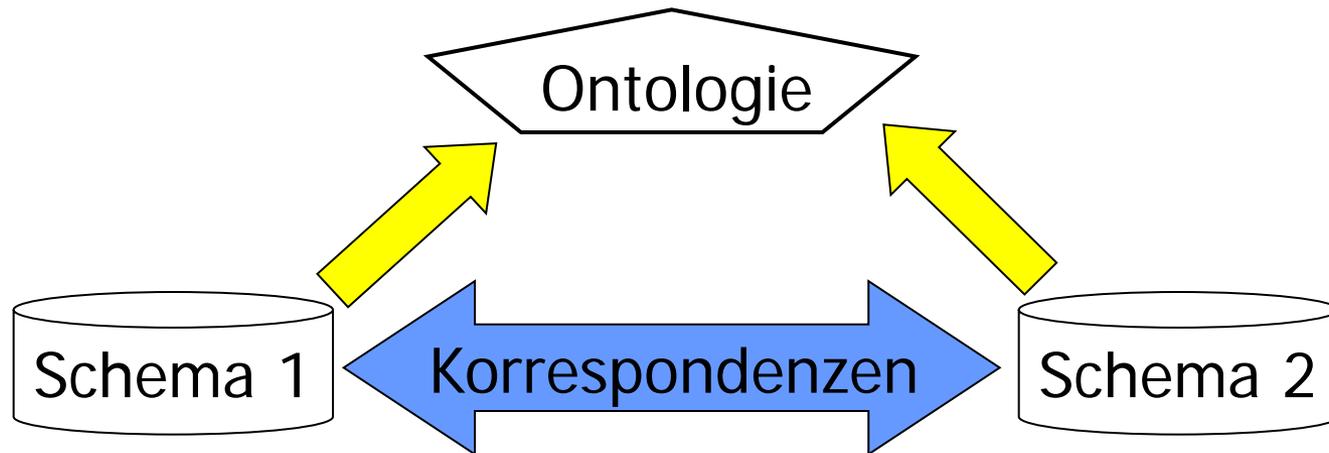
- Unterschenkel?
- Knie?



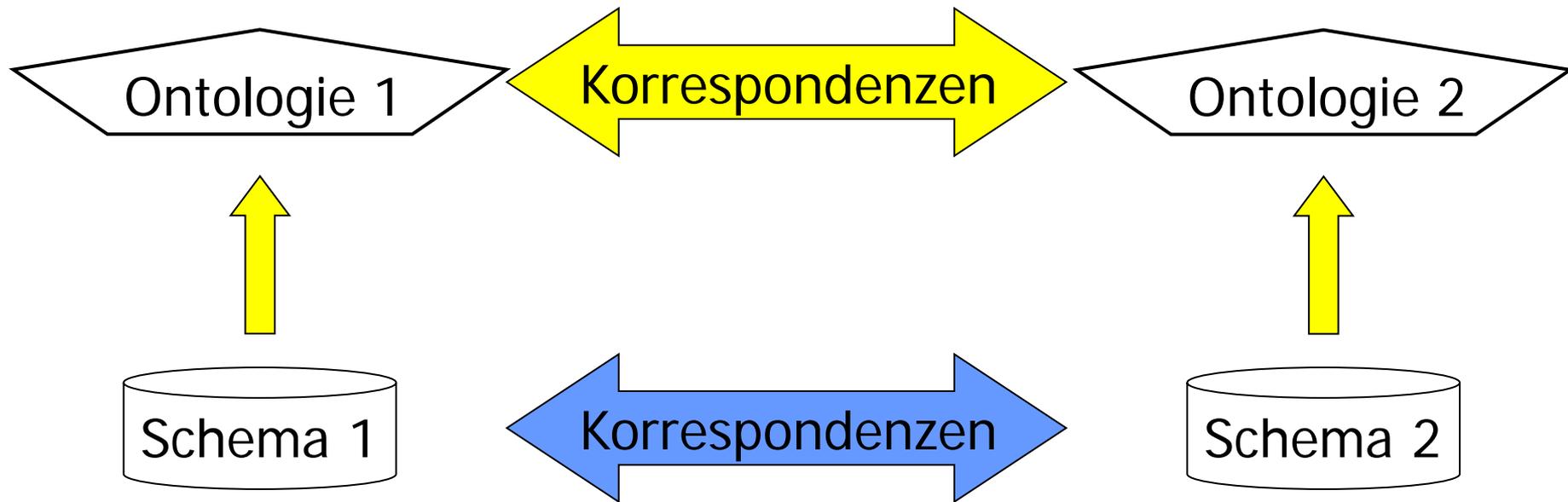
Reality Check

- I.A. verwenden verschiedene Quellen **verschiedene Ontologien**
 - `schauspieler` \equiv `person` \sqcap \exists `spielt_in.film`
 - `schauspieler` \equiv `mann` \sqcap \exists `beruf.schauspiel`
 - `schauspieler` \equiv `mensch` \sqcap `schummler`
 - Sind das die selben Klassen von Personen?
- **Ontologieintegration** statt Schemaintegration
- **Ontologiealignment** statt Schema Mapping

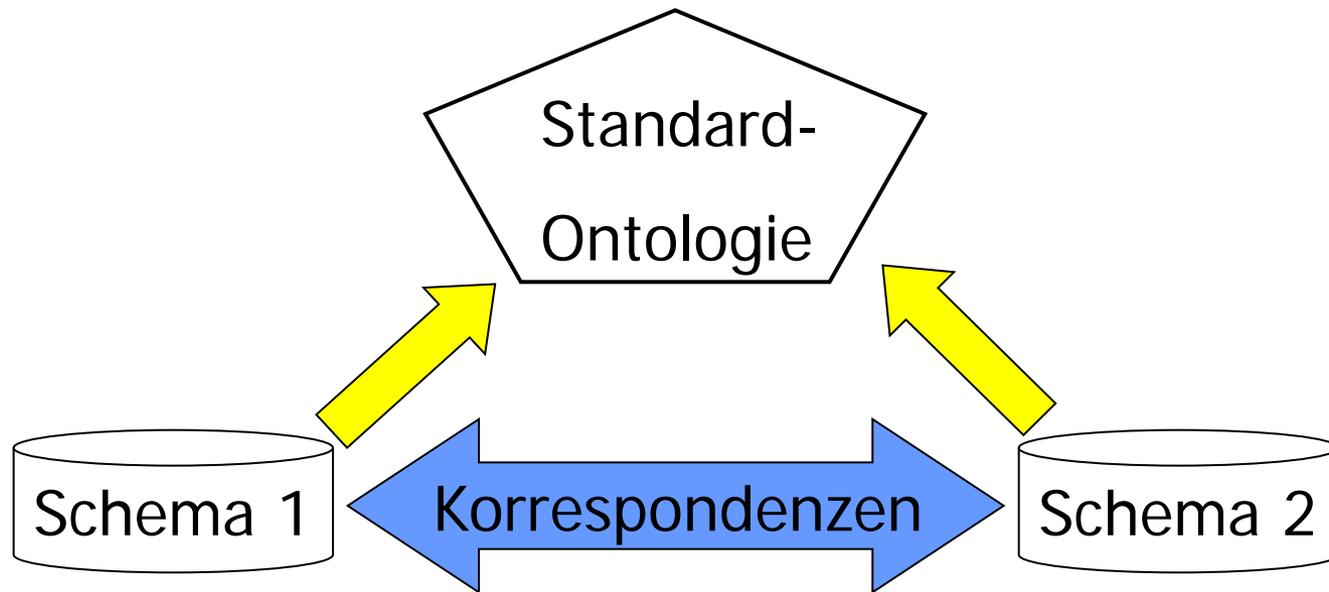
Hoffnung



Normalfall



Standards



Ontologien und Standards

- **Proprietäre Ontologien** helfen nicht bei Integration
 - Beschreibungslogiken ausdrucksstärker als Schemasprachen
 - Z.B: Verschiedene Beziehungstypen
 - Integrationsproblem wird schwieriger
- Besser: **Standardontologie** zur Beschreibung einer Domäne
 - Quellen mappen ihre Schemata in diese Ontologie
- Probleme
 - Standards werden **nicht eingehalten**
 - **Hoher Aufwand** zur Erstellung umfassender und akzeptierter Ontologien
 - Web2.0: Folksonomie, freies Tagging statt formale Definitionen

Schema und Daten

- Ontologien setzen **Konzepte** in Beziehung
 - Schemata: Begriffe und deren Struktur
 - Daten: Nur Begriffe
- Beschreibung von Schemaelementen und von Daten
 - Strukturierte „**Controlled Vocabularies**“
- Beschreibung von Schema eher schwierig
 - Ontologien kennen keine Relationen etc.
 - Schwierigkeiten bei struktureller Heterogenität

Inhalt dieser Vorlesung

- Semantische Heterogenität
- Semantische Integration
- **Ontologien**
 - Einführung
 - Thesauri und semantische Netze
- Beschreibungslogiken
- Ontologiebasierte Integration

Was ist Ontologie?

- Philosophisch: Ontologie = **Lehre vom Sein** und den Bedingungen des menschlichen Seins
- Beantwortung von Fragen wie [wikipedia]
 - What constitutes the **identity of an object**?
 - What is a physical object?
 - What features are the essential, as opposed to merely accidental, attributes of a given object?
 - What are an object's properties or relations and how are they related to the object itself?
 - When does an object go out of existence, as opposed to merely changing?
- Beispiel
 - Wenn ein Mensch stirbt – hört er auf zu existieren?

Was ist eine Ontologie für uns?

- „An ontology is a data model that represents a set of **concepts** within a domain and the **relationships** between those concepts. It is used to reason about the objects within that domain.“ [Wikipedia.org]
- Eine Ontologie ist „**an explicit specification of a conceptualisation**“ [Gru93]
 - ‚Konzeptionalisierung‘: **Abstraktes Modell** von Phänomenen der wirklichen Welt durch Identifikation der relevanten Konzepte
 - ‚Explizit‘: Konzepte und Beziehungen sind explizit definiert
 - Spezifikation: Verwendung einer formalen Sprache
- „Ontology is not about peoples’ conceptions or interpretations, **but about the world.**“ [OntologysWorks.com]
 - Was unterscheidet sonst eine schlechte Ontologie von einer guten?
 - Was ist objektiv wahr, was ist subjektiv?

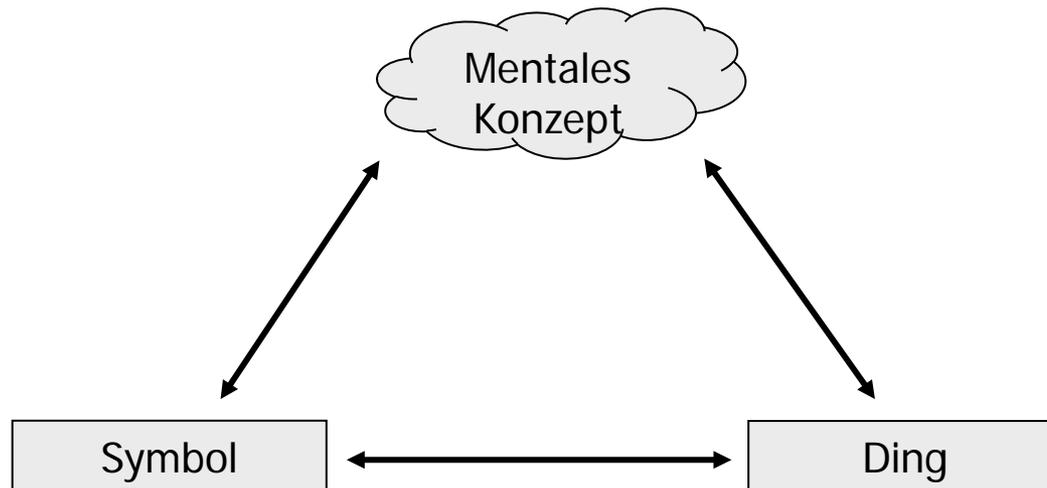
Domäne: Verwandtschaft

$$\begin{aligned} \textit{frau} &\equiv \textit{person} \sqcap \textit{weiblich} \\ \textit{mann} &\equiv \textit{person} \sqcap \neg \textit{weiblich} \\ \textit{mutter} &\equiv \textit{frau} \sqcap \exists \textit{hat_kind}.\textit{person} \\ \textit{vater} &\equiv \textit{mann} \sqcap \exists \textit{hat_kind}.\textit{person} \\ \textit{elternteil} &\equiv \textit{vater} \sqcup \textit{mutter} \\ \textit{grossmutter} &\equiv \textit{frau} \sqcap \exists \textit{hat_kind}.\textit{elternteil} \end{aligned}$$

- Beantwortung von Fragen wie
 - Ist jede Großmutter eine Person?
 - Hat jede Großmutter ein Kind?
 - Ist jede Großmutter ein Elternteil?
 - Wie nenne ich jemanden, der mit dem Schwager meiner Schwester verheiratet ist und keine Kinder hat?
 - ...

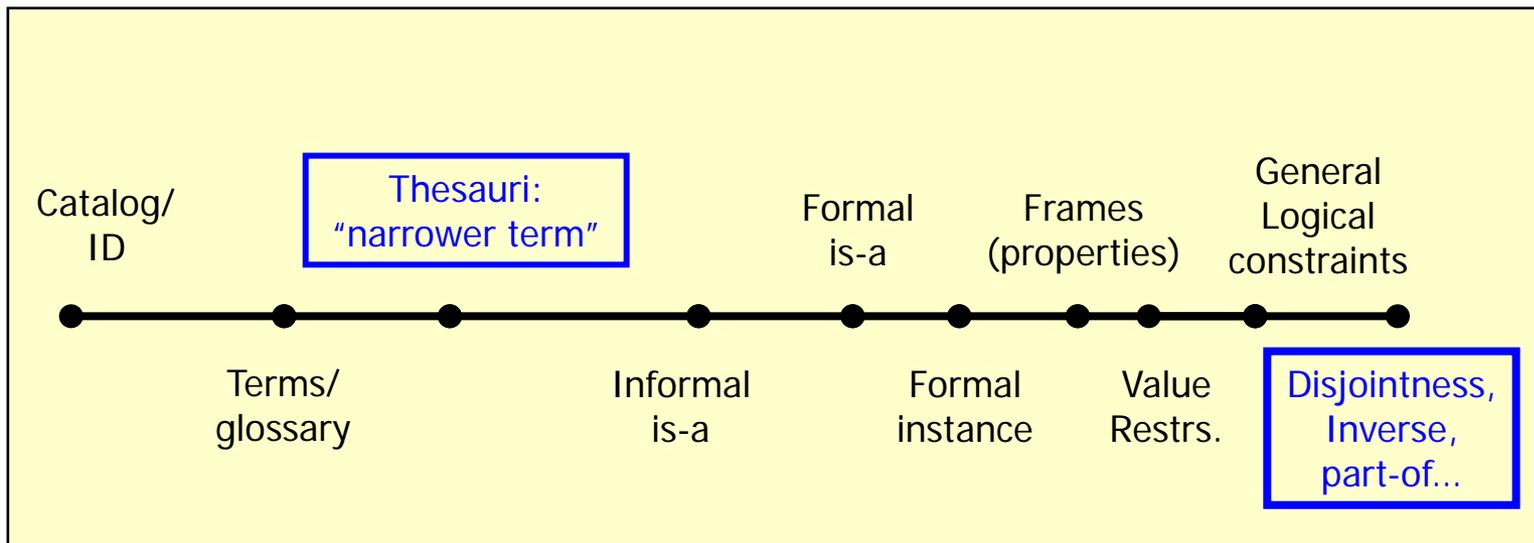
Konzeptualisierung

- Konzeptualisierung einer Domäne
 - Anwendung oder **Ausschnitt der Welt**
 - Erfassung aller **relevanten Konzepte**
 - Erfassung der Beziehungen zwischen diesen Konzepten
 - Insbesondere Subkategorie-Beziehungen
- Festlegung des **Vokabulars**, über das man sprechen kann



Explizite Spezifikation

- **Formalen Sprache zur Spezifikation**
 - Wissensrepräsentationssprache
- **Unterschiedliche Ausdrucksstärken**



Source: Robert Stevens, From Building and Using Ontologies, U. of Manchester

Kommunikation

- Ontologien wurden ursprünglich zur **Kommunikation** zwischen Softwareagenten konzipiert
 - Nachricht: Sachverhalt + Ontologie
- Ziel: Unterstützung der Kommunikation in einer Gruppe von ... (Personen, Programmen, Organisationen, ...)
 - = Festlegung von Standards
- Ohne den **Austausch von Daten** machen Ontologien wenig Sinn
 - Zu viel Aufwand
 - Semantik von Begriffen muss dann nicht expliziert werden, sondern kann verborgen im Code bleibe

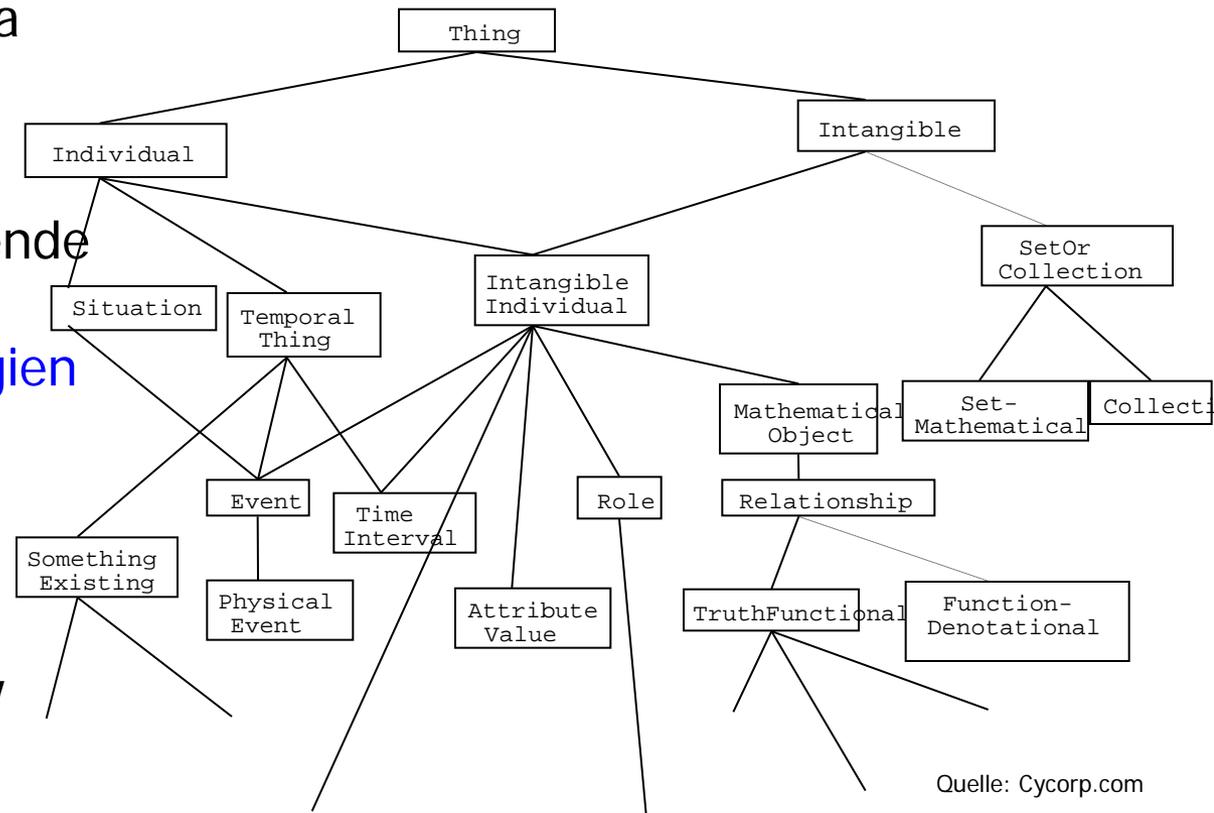
Arten von Ontologien

- Domänen- oder
anwendungsspezifische
Ontologien

- Das ist unser Thema

- Top-Level
Ontologien

- Domänenübergreifende
Sachverhalte
- Verknüpfen Ontologien
miteinander
- Arbeitersparnis
- SUMO: Suggested
Upper Ont. (IEEE)
- Cyc Upper Ontology



Quelle: Cycorp.com

Inhalt dieser Vorlesung

- Semantische Heterogenität
- Semantische Integration
- Ontologien
 - Einführung
 - [Thesauri und semantische Netze](#)
- Beschreibungslogiken
- Ontologiebasierte Integration

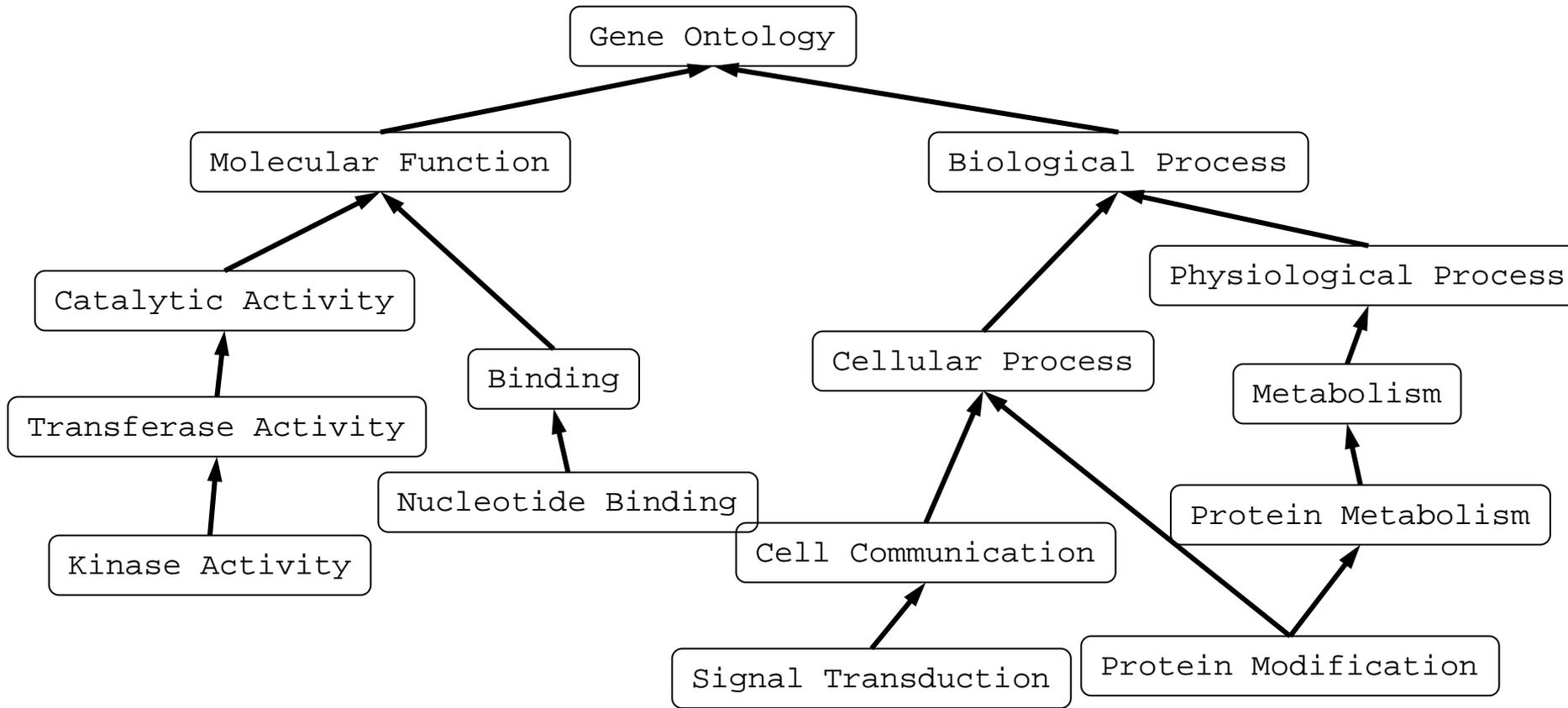
Thesaurus

- Einfachste Form einer Ontologie
- Elemente
 - **Begriffe** (inkl Dokumentation)
 - **Beziehungen**: ISA, SYNONYM_OF, PART_OF
 - Haben Eigenschaften (Symmetrie, Transitivität, ...)
 - Informeller: RELATED_TO, SIMILAR_TO, USED_FOR
 - Graph der Begriffe und Beziehungen muss **zyklenfrei** sein (DAG)
 - Bezüglich der ISA und PART-OF Beziehung
- **Anwendung: (Ein)ordnung von Dingen**
 - Zuordnung von Schlagwörter / Kategorien zu Objekten
 - Wichtig für die Suche: **Transitivität der ISA** Beziehung
 - Schwieriger: PART_OF
 - Ist ein Embryo Teil des Körpers der Mutter?
 - Physikalisch umschlossen (wann)? Konstituierendes Teil? Notwendiges Teil?

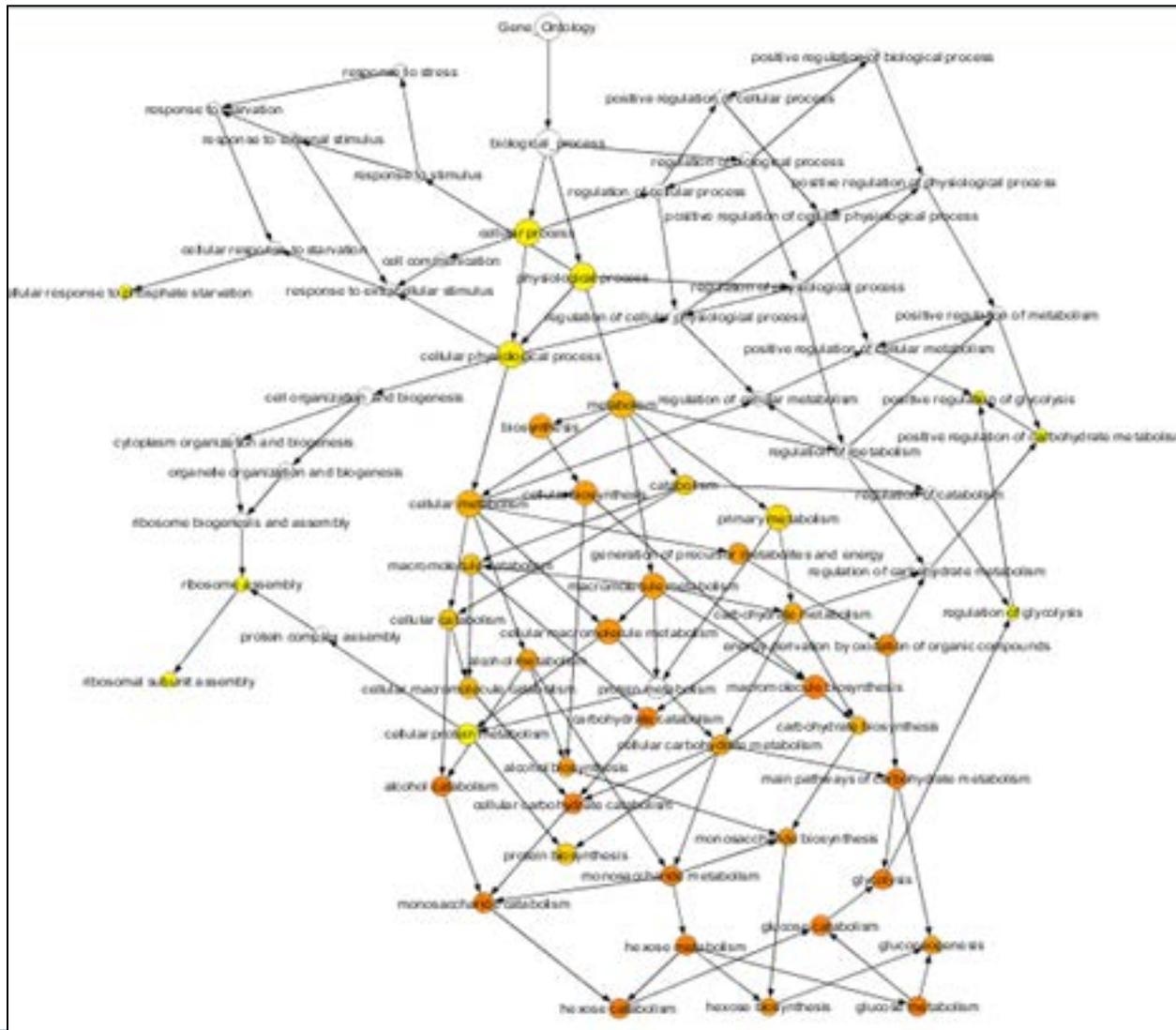
Example: Gene Ontology

- Goals of the GO Consortium
 - Development of a **structured vocabularies** describing certain aspects of molecular biology
 - Use of these vocabularies to **annotate genes**
 - Develop tools for editing and using the GeneOntology
- Three vocabularies
 - Biological processes
 - Molecular function
 - Intra-cellular location
- Created an ontology “industry” (OBO Foundry)
 - OBO: Open Biomedical Ontology
 - Guidelines and best practice (and license)
 - ~ 60 ontologies (1/2009), ~120 (6/2016)

Small Fraction of GO



A Large Ontology



Database Annotation InterPro

Reset View InterProEntry

This entry is from: [INTERPRO](#)

Save Link Printer Friendly

Glucose-methanol-choline oxidoreductase

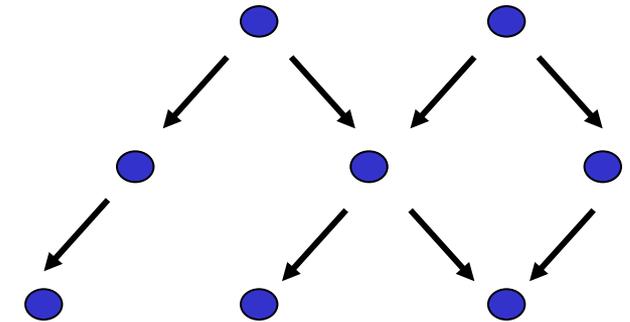
Accession	IPR000172; (GMC_oxred) matches 174 proteins
FullName	Glucose-methanol-choline oxidoreductase
Type	Family
Signatures	PROSITE: PS00623 GMC_OXRED_1 PROSITE: PS00624 GMC_OXRED_2 PFAM: PF00732 GMC_oxred
Biological Process	electron transport (GO:0006118)
Molecular Function	electron transfer flavoprotein (GO:0008246)
Abstract	The glucose-methanol-choline (GMC) oxidoreductase oxidoreductases are FAD flavoproteins oxidoreductases [1, 5]. These enzymes include a variety of proteins; choline dehydrogenase (CHD), methanol oxidase (MOX) and cellobiose dehydrogenase [EC:1.1.5.1] [6] which share a number of regions of sequence similarities. One of these regions, located in the N-terminal section, corresponds to the FAD ADP- binding domain. The function of the other conserved domains is not yet known.
Examples	<ul style="list-style-type: none">• P22637 Cholesterol oxidase (CHOD) () from Brevibacterium sterolicum and Streptomyces strain SA-COO.• P13006 Glucose oxidase () (GOX) from Aspergillus niger.• O50048 (R)-mandelonitrile lyase () (hydroxynitrile lyase) from plants [PUB00004524].• P54223 Choline dehydrogenase () (CHD) from bacteria.• P18173 Glucose dehydrogenase (GLD) () from Drosophila.

Document: Done (2.794 secs)

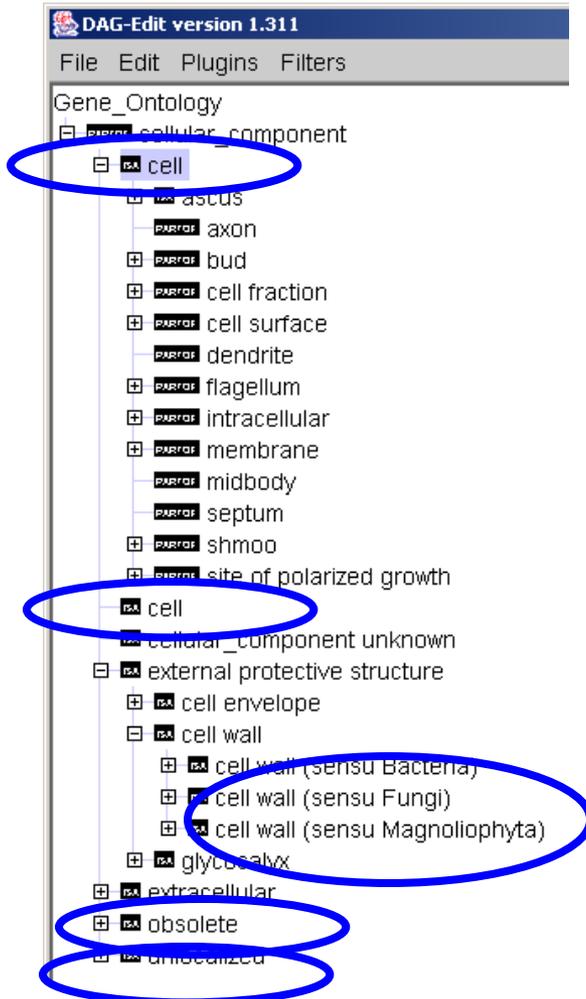
- Used by many databases
- Allows cross-database search
- Provides fixed meaning of terms
 - As informal textual description, not as formal definitions

GO model

- A GO ontology is
 - A collection of terms
 - Each term has a (free text) description
 - Each term has a unique ID (GO:XXXX)
 - Terms may have synonyms and DB-Xrefs
 - IS-A relationships
 - PART-OF relationships
 - IS-A and PART-OF relationships form (each) an **acyclic graph**
 - True Path rule: „Every path from a node back to the root must be biologically accurate“
- Support for versioning
 - Obsolete terms



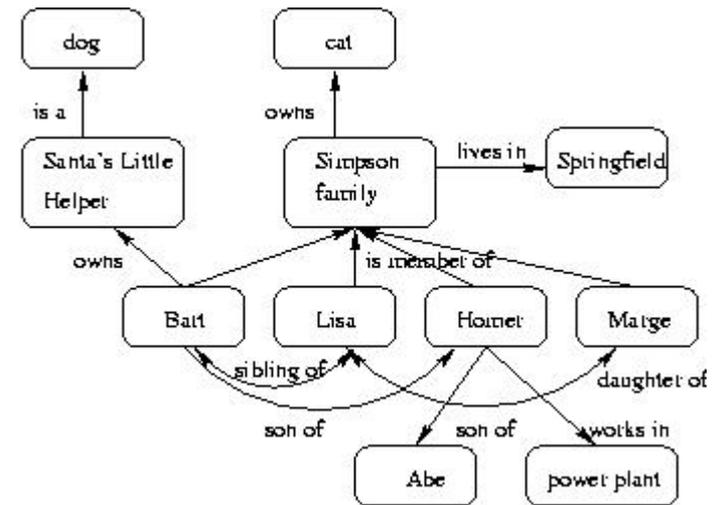
Example



- „Little semantic rigor“
 - „Obsolete“?
 - Species specificity?
 - Equal names?
 - „... unlocalized is a ... „

Semantische Netze

- Aufgaben der Forderungen nach
 - Zyklenfreiheit
 - Begrenzter Zahl von Beziehungstypen
- Entwickelt als **Wissensrepräsentationsmechanismus** in der künstlichen Intelligenz
- Kanten können beliebige Label tragen
 - Sehr flexibel
 - Aber: Keine logische **Inferenz**
 - Da keine formalen Eigenschaften definiert sind
 - Heuristische Inferenz: Zwei Begriffe sind sich **semantisch näher**, je kürzer der kürzeste Pfad zwischen ihnen



Inhalt dieser Vorlesung

- Wdh: Semantische Heterogenität
- Semantische Integration
- Ontologien
- **Beschreibungslogiken**
- Ontologiebasierte Integration

Wissensrepräsentationssprachen

- Entwickelt in den 80ziger
 - Urvater ist [KL-ONE](#) [BS85]
 - Beschreibungslogiken, [Description Logics \(DL\)](#), Terminologische Logiken
 - Vertreter: Classic, Loom, DAML, DAML+OIL, OWL, ...
 - Ein weites Feld, das wir nur anreißen
- Formale Sprachen zur Definition von Konzepten und deren Beziehungen untereinander
- Zentrale Begriffe: [Konzepthierarchien](#) und Klassenzugehörigkeit
 - DL: Sag mir die Eigenschaften, und ich [sage dir die Klasse](#)
 - OO: Sag mir die Klasse, und ich sage dir die Eigenschaften

Description Logic

- Elemente einer DL
 - Atomarer Konzepte
 - Uninterpretierte, eindeutige Namen
 - Konzept ~ Basisklasse
 - Atomare Rollen
 - Uninterpretierte, eindeutige Namen
 - Rolle ~ Basisbeziehung
 - Sprache: Steht ein Konzept C in der Beziehung R zu einem Konzept D, so „füllt“ C die Rolle R von D
 - Abgeleitete Konzepte und abgeleitete Rollen
 - Abgeleitet unter Verwendung einer je nach DL unterschiedlichen Menge von Operationen
 - Axiome

Abgeleitete Konzepte

- Definition

*Seien C, D (atomare oder abgeleitete) Konzepte und R eine atomare Rolle. Dann können **abgeleitete Konzepte** wie folgt gebildet werden*

- $C \cap D$

- *Das Konzept, das alle Individuen in der Schnittmenge von C und D repräsentiert*

- $C \cup D$

- *..., das alle Individuen in der Vereinigungsmenge von C und D repräsentiert*

- $\neg C$

- *..., das das Komplement der Menge der Individuen in C repräsentiert*

- $\forall R.C$

- *..., alle Individuen, deren Rollenfüller für R Individuen des Konzeptes C sind*

- $\exists R.C$

- *..., alle Individuen, die mindestens einen Rollenfüller für R haben, der ein Individuum des Konzeptes C ist*

- $\geq nR$

- *..., alle Individuen, die höchstens n Rollenfüller für R haben (\sim Kardinalität)*

Axiome

- Definition

*Seien C und D (atomare oder abgeleitete) Konzepte. Dann können **Axiome** die folgende Form haben*

- $C \equiv D$

- *Alle Individuen von C sind auch Individuen von D und umgekehrt*

- $C \sqsubseteq D$

- *Alle Individuen von C sind auch Individuen von D*

- Bemerkung

- Mit einem Axiom **postuliert man eine Subsumptionsbeziehung**

- Klassifikation leitet eine Subsumptionsbeziehung ab

- Wir werden später noch weitere Axiomarten kennen lernen

- Siehe OWL Full

Ontologie

- Definition

*Eine **formale Ontologie** besteht aus*

- *einer Menge von atomare Konzepten,*
- *einer Menge von atomaren Rollen,*
- *einer Menge von abgeleiteten Konzepten und*
- *einer Menge von Axiomen über den Konzepten/Rollen.*

- Beispiel

- Atomare Konzepte:
person, weiblich

frau \equiv *person* \sqcap *weiblich*

mann \equiv *person* \sqcap \neg *weiblich*

- Atomare Rollen:
hat_kind

mutter \equiv *frau* \sqcap \exists *hat_kind.person*

vater \equiv *mann* \sqcap \exists *hat_kind.person*

- Abgeleitetes Konzept:
person \sqcap weiblich

elternteil \equiv *vater* \sqcup *mutter*

- Axiome:

grossmutter \equiv *frau* \sqcap \exists *hat_kind.elternteil*

Subsumption

- Definition
Ein Konzept C subsumiert ein Konzept D , $D \sqsubseteq C$, wenn alle Individuen von C auch Individuen von D sind
 - C ist generischer, abstrakter, allgemeiner als D
- Inferenz in DL bedeutet im wesentlichen das **Beweisen von Subsumptionsbeziehungen** zwischen Konzepten
- Deren Entscheidbarkeit hängt von den erlaubten Operationen zur Ableitung von Konzepten/Rollen und den erlaubten Axiomen ab
 - In der hier verwendeten Sprache ist Subsumption im Allgemeinen schon unentscheidbar

Konsistenz und Erfüllbarkeit

- Definition

Gegeben ein Konzept C und eine Ontologie O mit $C \in O$.

- *C heißt **erfüllbar**, wenn es Individuen geben kann, die zur Menge der von C repräsentierten Individuen gehört*
- *O heißt **widerspruchsfrei** (oder **konsistent**), wenn alle Konzepte in O erfüllbar sind*

- Bemerkung

- Erfüllbarkeit kann auf Subsumption zurückgeführt werden
- Ist also Subsumption in einer DL entscheidbar, kann man für jede Ontologie in diese DL Widerspruchsfreiheit testen

Klassifikation

- Definition

*Das **Klassifikationsproblem** für ein Konzept C bzgl. einer Ontologie O berechnet alle C subsumierenden bzw. von C subsumierten Klassen in O .*

- Bemerkung

- Offensichtlich ist Klassifikation lösbar, wenn Subsumption entscheidbar ist
- Durch Klassifikation wird ein (neues) Konzept in die **Konzepthierarchie** von O eingeordnet
- Klassifikation basiert auf der Beschreibung der Eigenschaften von C durch abgeleitete Konzepte und Axiome

Beispiel

<i>frau</i>	\equiv	<i>person</i> \sqcap <i>weiblich</i>
<i>mann</i>	\equiv	<i>person</i> \sqcap \neg <i>weiblich</i>
<i>mutter</i>	\equiv	<i>frau</i> \sqcap \exists <i>hat_kind.person</i>
<i>vater</i>	\equiv	<i>mann</i> \sqcap \exists <i>hat_kind.person</i>
<i>elternteil</i>	\equiv	<i>vater</i> \sqcup <i>mutter</i>
<i>grossmutter</i>	\equiv	<i>frau</i> \sqcap \exists <i>hat_kind.elternteil</i>

- **Ableitbare Subsumptionsbeziehungen**

- **frau \sqsubseteq person**
- **frau \sqsubseteq weiblich**
- **mutter \sqsubseteq frau \sqsubseteq person**
- **grossmutter \sqsubseteq frau**
- **grossmutter \sqsubseteq elternteil**
 - Sei **$o \in$ grossmutter**
 - Es folgt: **$o \in$ frau** und **$o \in$ \exists hat_kind.elternteil**
 - Es folgt: **$o \in$ \exists hat_kind.vater** oder **$o \in$ \exists hat_kind.mutter**
 - Es folgt: **$o \in$ \exists hat_kind.person**
 - Es folgt: **$o \in$ mutter**
 - Es folgt: **$o \in$ elternteil**

- **Inferenzalgorithmen** lösen solche Probleme automatisch
 - Z.B. Racer, FaCT, ...

Erfüllbarkeit

<i>frau</i>	\equiv	<i>person</i> \sqcap <i>weiblich</i>
<i>mann</i>	\equiv	<i>person</i> \sqcap \neg <i>weiblich</i>
<i>mutter</i>	\equiv	<i>frau</i> \sqcap \exists <i>hat_kind.person</i>
<i>vater</i>	\equiv	<i>mann</i> \sqcap \exists <i>hat_kind.person</i>
<i>elternteil</i>	\equiv	<i>vater</i> \sqcup <i>mutter</i>
<i>grossmutter</i>	\equiv	<i>frau</i> \sqcap \exists <i>hat_kind.elternteil</i>

- Wir erweitern unsere Axiome und Konzepte
 - **`keine_grossmutter`** \equiv **`frau`** \sqcap **`\neg hat_kind.person`**
 - **`grossmutter`** \sqsubseteq **`keine_grossmutter`**
- Nun haben wir ein **unerfüllbares Konzept**
 - Sei $o \in$ **`grossmutter`**
 - Es folgt: $o \in$ **`hat_kind.elternteil`**
 - Es folgt: $o \in$ **`hat_kind.person`**
 - Es folgt: $o \notin$ **`\neg hat_kind.person`**
 - Es folgt: $o \notin$ **`keine_grossmutter`**
 - Also kann es kein Element von **`grossmutter`** geben
 - Das Konzept **`grossmutter`** ist unerfüllbar
 - Unsere Ontologie ist **inkonsistent**

Konzepthierarchie

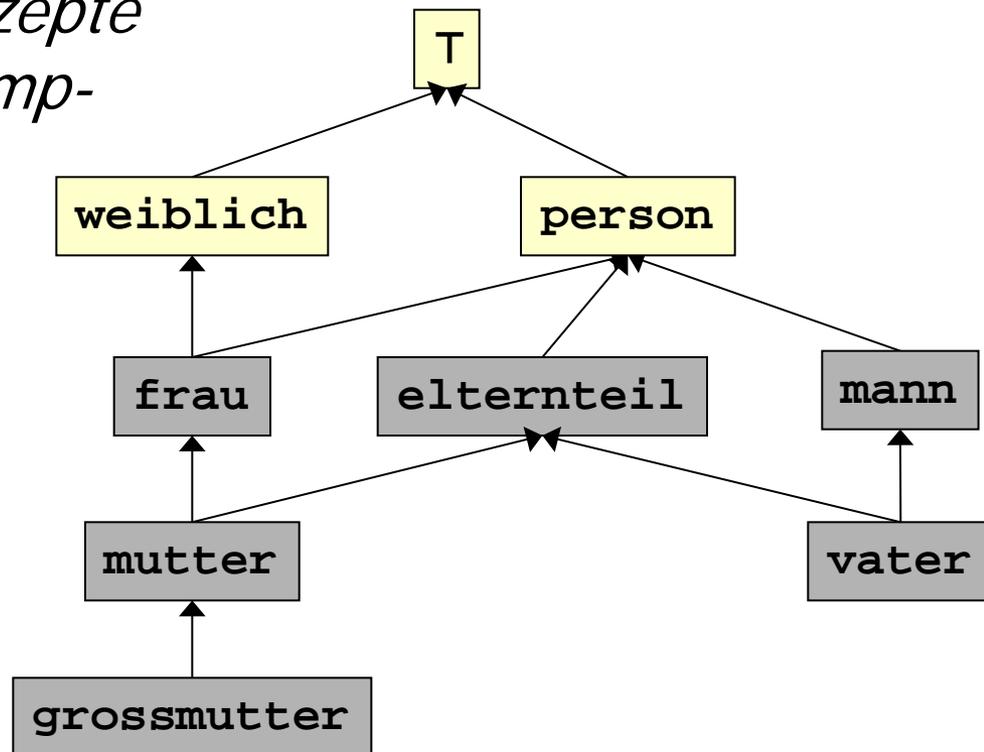
- Sei T die Menge aller Individuen (das allgemeinste Konzept)

- Definition

Die *Konzepthierarchie* einer Ontologie O ist ein Graph, der alle Konzepte als Knoten und alle Subsumptionsbeziehungen als Kanten enthält.

- Lemma

Ist eine Ontologie O widerspruchsfrei, so ist ihre *Konzepthierarchie* ein DAG mit T als Wurzel.

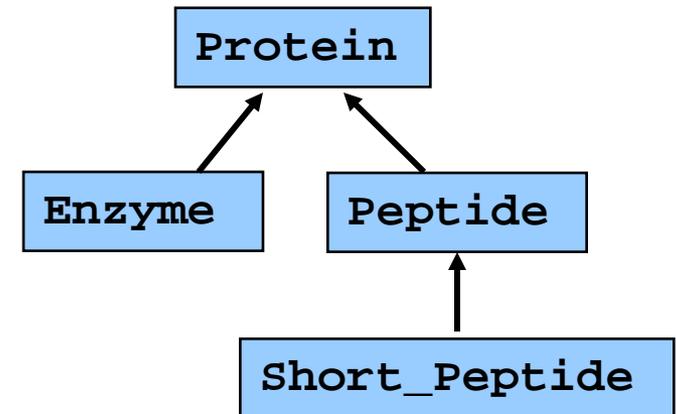


Ein weiteres Beispiel

- `enzyme` \equiv `protein` \wedge \exists `catalyses.reaction`
- `peptide` \equiv `protein` \wedge `<100 aminoacids`
- `short_peptide` \equiv `Protein` \wedge `<50 aminoacids`
- `long_peptide` \equiv `protein` \wedge `>100 aminoacids`
- `long_peptide` \sqsubseteq `peptide`

- Was folgt?

- `enzyme` \sqsubseteq `protein`
- `peptide` \sqsubseteq `protein`
- `long_peptide` \sqsubseteq `protein`
- **Unerfüllbar:** `long_peptide`

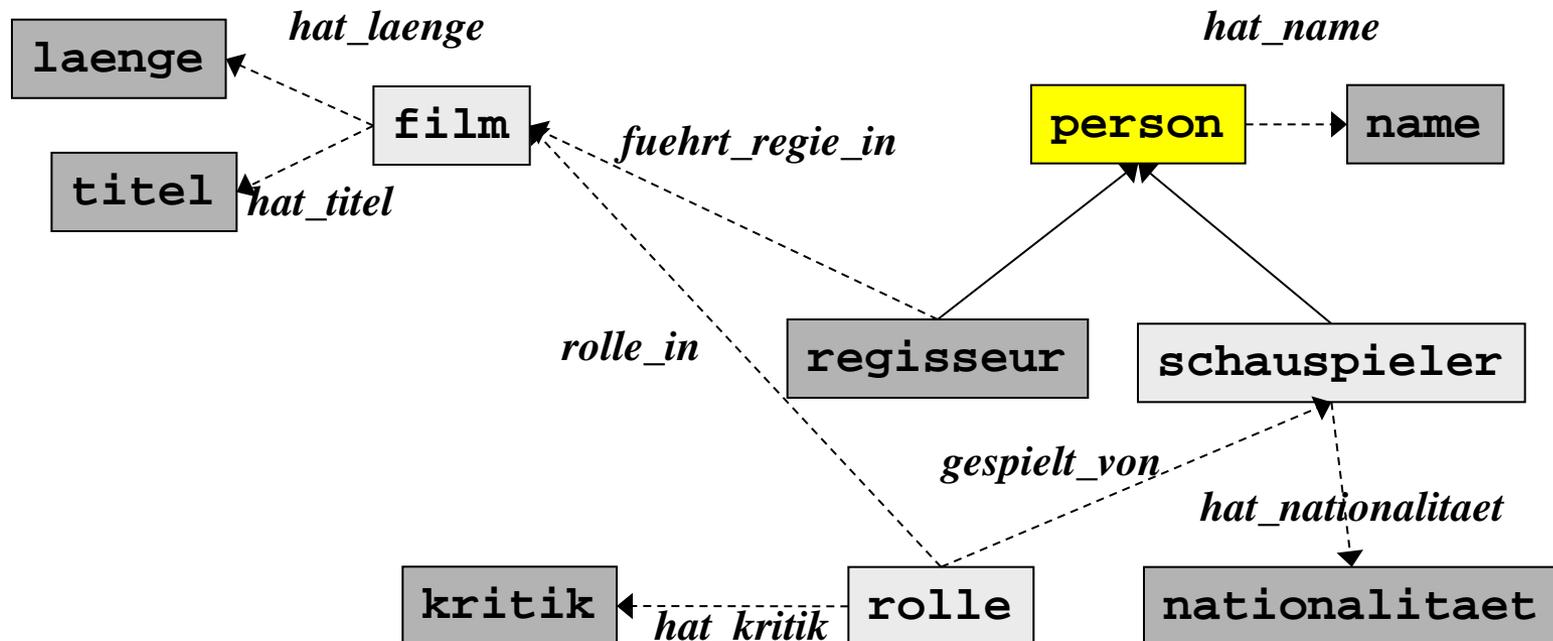


DL und relationales Model / OO

- Keine Attribute
 - Attribute werden als eigene Konzepte definiert
 - Ob ein Konzept C ein Attribut A hat, wird über eine Rolle definiert
 - Beispiel: `person hatName name`
- Beziehungen werden zu Rollen
- Rollen und Attributkonzepte leben unabhängig von Klassen
 - Beispiel: `organization hatName name, person hatName name`
 - `hatName (name)` ist eine (globale) Rolle (Konzept) mit einmal definierten Eigenschaften
 - Ungleich dem [Scope-Konzept](#) in OO oder Programmiersprachen

Das Filmbeispiel

```
film(titel, typ, regisseur, laenge);  
schauspieler(schauspieler_name, nationalitaet);  
spielt(titel, schauspieler_name, rolle, kritik);
```



Inhalt dieser Vorlesung

- Semantische Heterogenität
- Semantische Integration
- Ontologien
- Beschreibungslogiken
- **Ontologiebasierte Integration**
 - Quellenkataloge
 - Integration durch Subsumption

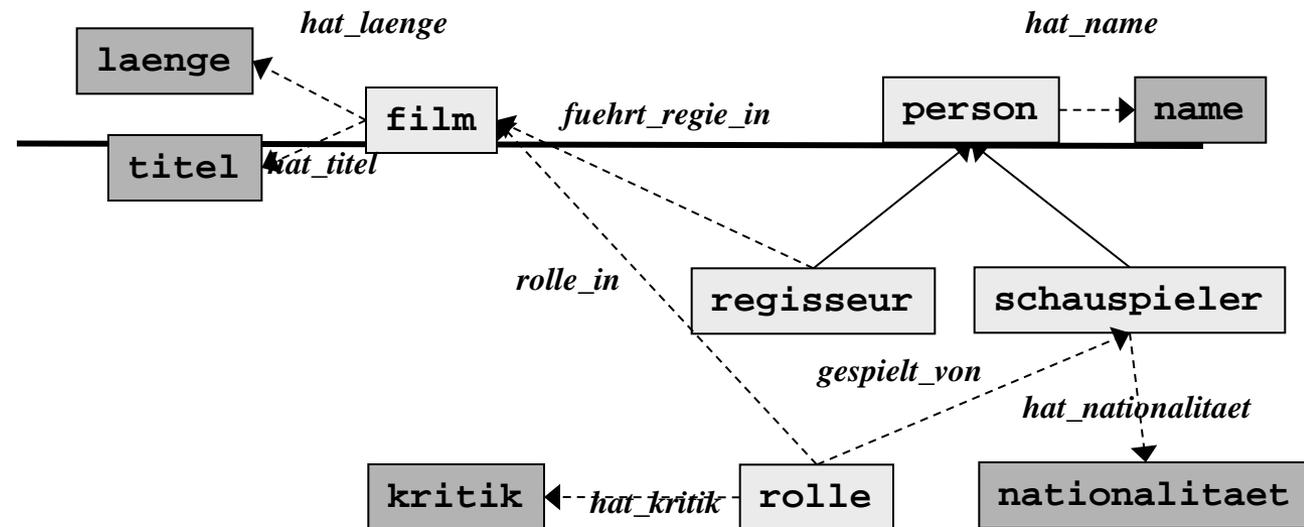
Quellenkataloge

- Die einfachste Art der Integration vieler Quellen
 - So einfach, dass wir sie nicht unter Architekturen aufgeführt haben
- Quellenkataloge
 - Modelliere eine Ontologie (oder Thesaurus) der Domäne
 - Annotiere **Quellen mit den Konzepten**, die ihren Inhalt beschreiben
 - Benutzer verwenden die Ontologie, um **passende Quellen** zu finden
 - Sehr populär (Web-Kataloge, DBCat, Yahoo Directories, ...)
- Pro und Contra
 - Bietet eine schnelle Übersicht über „Was ist da“
 - Einfach, schnell, billig, non-intrusive
 - Quellen müssen **nicht kooperieren**
 - Keine Überwindung von Heterogenität, keine Transparenz
 - **Keine Integration** in unserem Sinne

Integration durch Subsumption

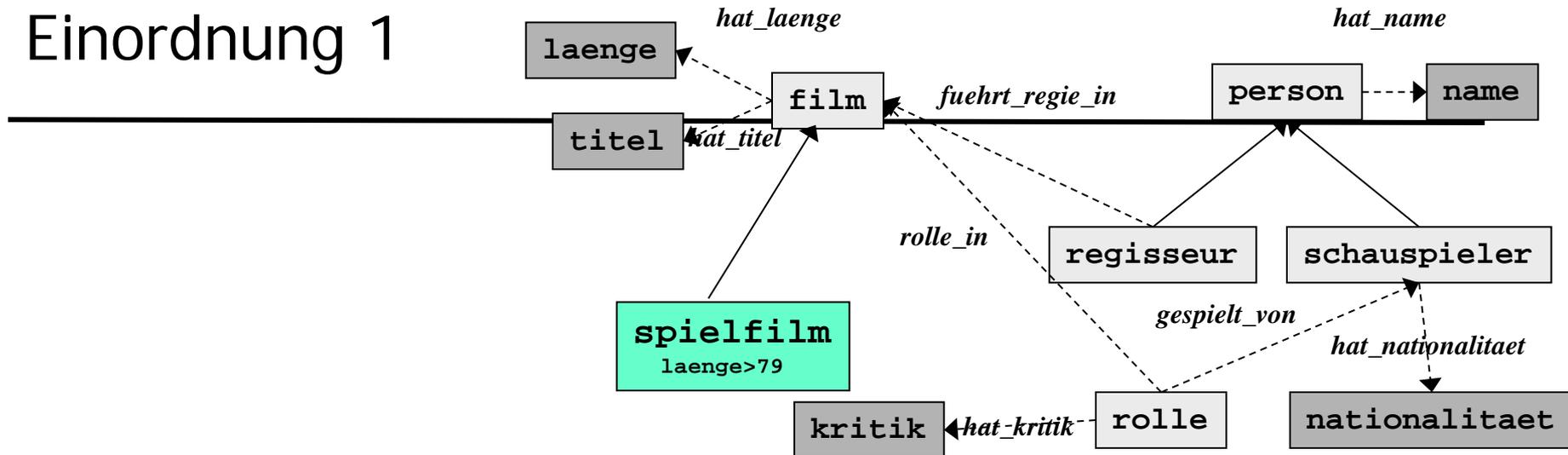
- Fokus semantische Heterogenität
 - Bedeutung von Schemaelementen
 - Klassische Ontologiesprachen tun sich schwer mit struktureller Heterogenität
 - Aber es gibt Erweiterungen; Kombinationen von DL und Datalog
- Vorgehen
 - Angelehnt an das SIMS Projekt [AHK96]
 - Eine Domänenontologie O fungiert als globales Schema
 - Exportrelationen der Quellen werden als Konzepte in O definiert
 - Einordnung in O erfolgt durch Subsumption
 - Was kann in dieser Welt eine Query sein?

Filmquellen und globale Ontologie



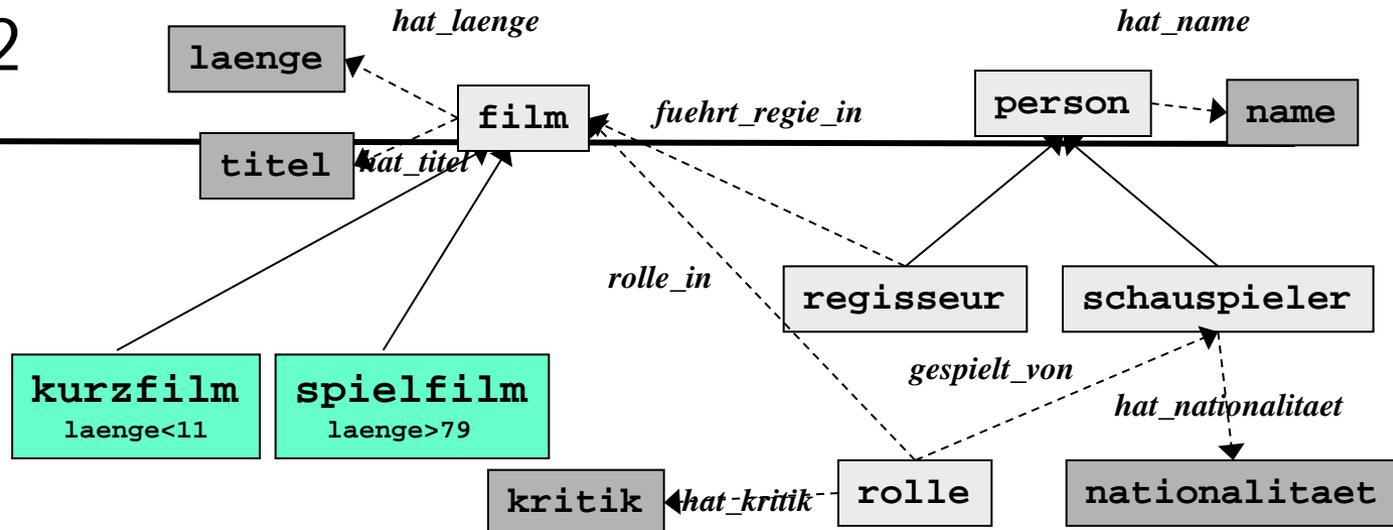
Datenquelle	Beschreibung
spielfilme(titel, regisseur, laenge)	Informationen über Spielfilme, die mindestens 80 Minuten Länge haben.
kurzfilme(titel, regisseur)	Informationen über Kurzfilme. Kurzfilme sind höchstens 10 Minuten lang.
filmkritiken(titel, regisseur, schauspieler, kritik)	Kritiken zu Hauptdarstellern von Filmen
us_spielfilme(titel, laenge, schauspieler_name)	Spielfilme mit US-amerikanischen Schauspielern
spielfilm_kritiken(titel, rolle, kritik)	Kritiken zu Rollen in Spielfilmen
kurzfilm_rollen(titel, rolle, schauspieler_name, nationalitaet)	Rollenbesetzungen in Kurzfilmen

Einordnung 1



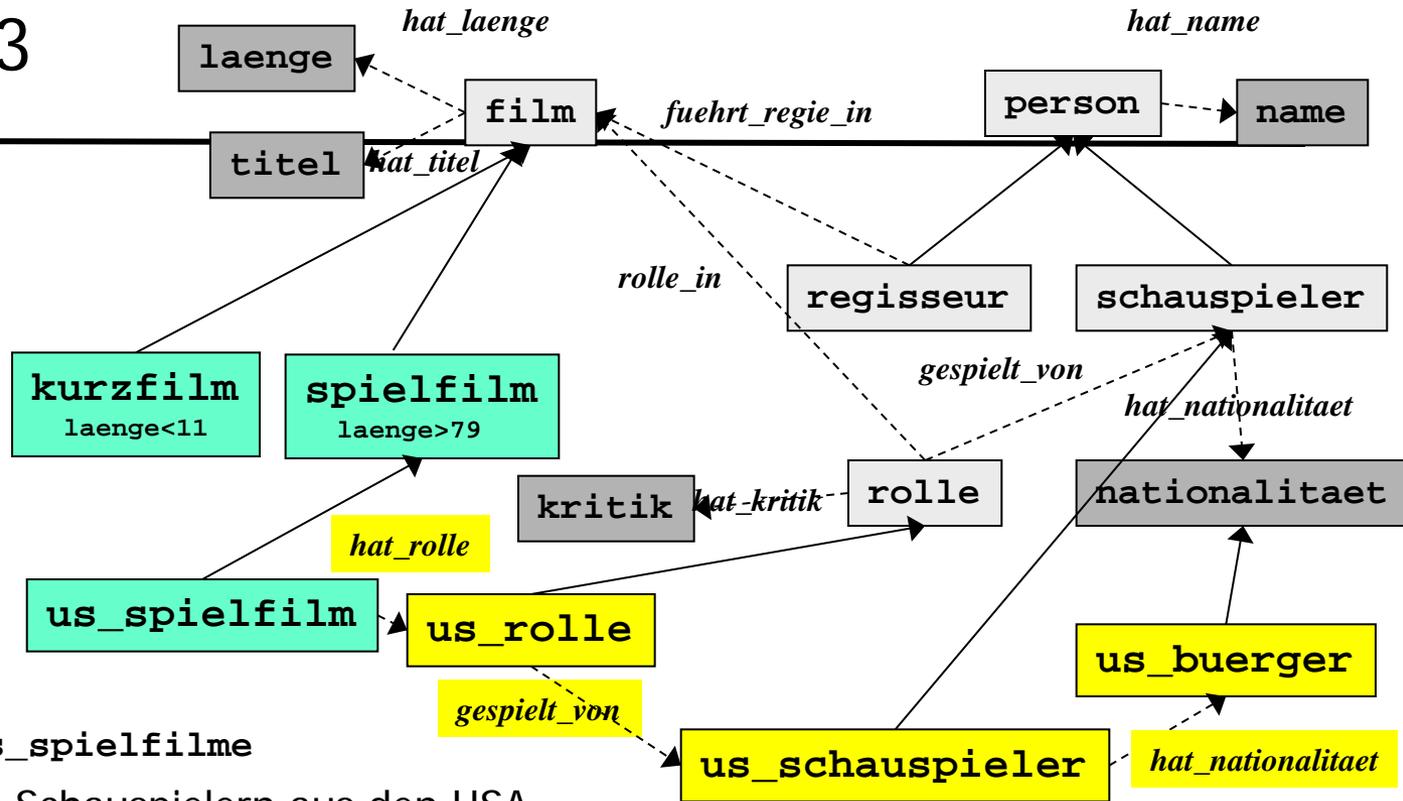
- Datenquelle **spielfilme**
 - Filme mit einer Länge über 79 Minuten
 - **spielfilm** \equiv **film** \cap \forall laenge > 79
 - „ \forall laenge < 79“ ist nicht in unserer bisherigen Sprache ausdrückbar
 - Umwege sind möglich (sparen wir uns)

Einordnung 2



- Datenquelle **kurzfilme**
 - Filme mit einer Länge unter 11 Minuten
 - $kurzfilm \equiv film \cap \forall laenge < 11$

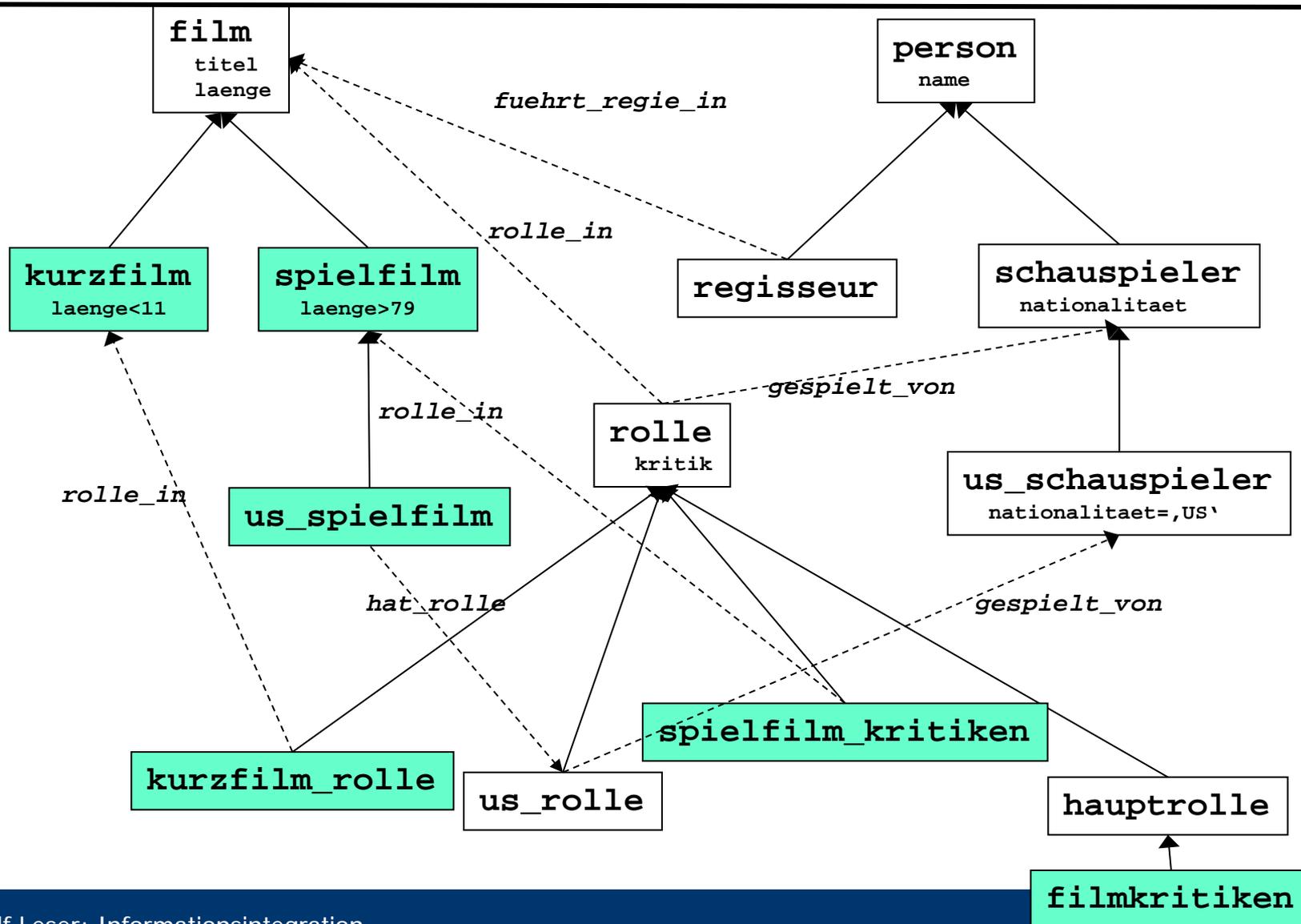
Einordnung 3



- Datenquelle **us_spiel filme**

- Spielfilme mit Schauspielern aus den USA
- Wir brauchen zunächst amerikanische Schauspieler und Rollen
 - $us_buerger \sqsubseteq nationalitaet$
 - $us_schauspieler \equiv schauspieler \sqcap \forall hat_nationalitaet.us_buerger$
 - $us_rolle \equiv rolle \sqcap \forall gespielt_von.us_schauspieler$
- $us_spiel film \equiv spielfilm \sqcap \forall hat_rolle.us_schauspieler$
 - *hat_rolle* definieren als inverse Rolle zu *rolle_in*

Zusammen

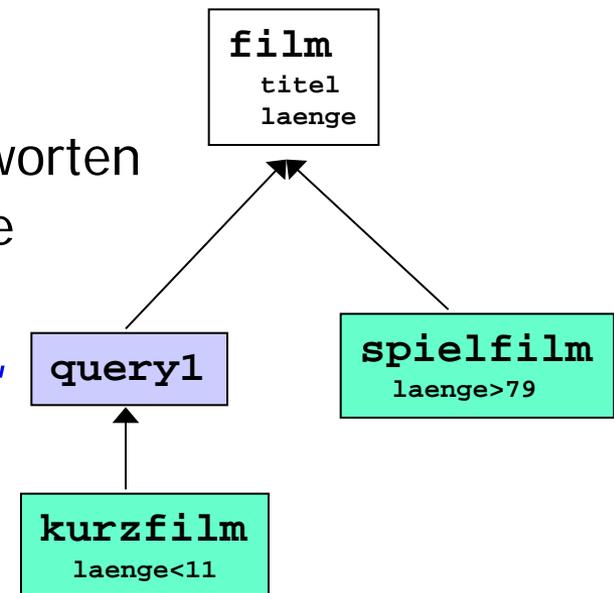


Anfragebearbeitung

- Was kann nun eine Query sein?

Anfragebearbeitung

- Ein Konzept
 - Alle Filme, die kürzer als 20 Minuten sind
 - $\text{query1} \equiv \text{film} \sqcap \forall \text{laenge} < 20$
- Können wir per **Subsumption in die Konzepthierarchie** einordnen
 - $\text{kurzfilm} \sqsubseteq \text{query1} \sqsubseteq \text{film}$
- Wo finden wir Antworten?
 - Alle spezielleren Konzepte sind sichere Antworten
 - Alle allgemeineren Konzepte sind potentielle Antworten
 - Ergebnis: Inhalt aller **spezielleren Konzepte**, die **Datenquellen** entsprechen



Inhalt dieser Vorlesung

- Semantische Heterogenität
- Semantische Integration
- Ontologien
- Beschreibungslogiken
- Ontologiebasierte Integration
 - Quellenkataloge
 - Integration durch Subsumption
- **Bewertung und Einordnung**

Anfrageplanung mit/ohne Ontologien

Föderiertes System	Ontologiebasierte Integration
Globales Schema Relationen und Attribute	Globale Ontologie Klassen und Rollen, Axiome
Korrespondenzen Explizite Spezifikation der Beziehungen	Quellen als Konzepte Definition mit gemeinsamem Vokabular und automatische Ableitung der Beziehungen
Anfrageplanung Global-as-View, Local-as-View, Query Containment	Subsumption Anfrage als Konzepte, Subsumption

Bewertung: Ontologiebasierte Integration

- Ausdrucksstark zur Definition **semantischer Unterschiede**
- Schwach bei struktureller Heterogenität
- Basiert auf einem globalen Schema
 - Benutzer müssen ein komplexes Modell neu lernen
- **Ontologiedesign** ist kompliziert
 - Je komplexer, je ausdrucksstärker die Modellierungssprache
 - **Ontology Engineering** – wie geht man vor? Wann hört man auf?
- Wesentliche Erleichterung wird erreicht, wenn **Ontologien als Standard** akzeptiert werden
 - Durch die semantische Ausdrucksstärke sind sie gut geeignet zur Definition von Standards
- Kein „silver bullet“

Literaturhinweise

- [AKS96] Arens, Y., Knoblock, C. A. and Shen, W.-M. (1996). "Query Reformulation for Dynamic Information Integration." *Journal of Intelligent Information Systems - Special Issue on Intelligent Information Integration* 6(2/3): 99-130.
- [BBBG+98] Baker, P. G., Brass, A., Bechhofer, S., Goble, C., Paton, N. and Quinn, M. (1998). "Transparent Access to Multiple Biological Information Sources: An Overview", University of Manchester.
- [BS85] Brachman, R. J. and Schmolze, J. G. (1985). "An Overview of the KL-ONE Knowledge Representation System." *Cognitive Science* 9(2): 171-216.
- [Gru93] Gruber, T. R. (1993). "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5(2): 199-220.
- [HM01] Haarslev, V. and Möller, R. (2001). "Description of the RACER System and its Applications". *Description Logics 2001*, Stanford, CA.