Vortrag

MicroKosmos

Nach

Ontology Development for MT: Ideology and Methodology,

Kavi Mahesh, 1995

IfI Universität Leipzig

20. Juni 2002

Inhaltsverzeichnis

Was ist MikroKosmos (μK)?

- Aufbau von μK
- Probleme bei der Übersetzung

Die Ontologie

- Concepts
- Slots und fillers
- Relations
- Attribute
- Literale
- TMR als erweiterter Instanzengraph
- Beispiel

- Und-Graph

Prinzipien und Probleme der Ontologieentwicklung

- Anforderungen an die Ontologie
- Strukturelle Prinzipien
- Redundanz
- Erfassung von Concepts
- Integration anderer Ontologien

Richtlinien

- Welche Konzepte in die Ontologie
- Namenskonventionen

Qualitätssicherung

Resourcen

Was ist MikroKosmos (μK)?

- System zur automatischen Übersetzung von spanischen Nachrichtenartikeln ins Englische
- Wurde an 400 Seiten Nachrichtenmeldungen und Artikeln über Firmenaufkäufe und Firmenfusionen getestet.
- Entwickelt an der New Mexico State University in Zusammenarbeit mit dem US Department of Defense
- Konzeptuell sprachunabhängig

- Erweiterung um andere Sprachen (Lexikas für Japanisch und Russisch sind in Arbeit)
- Es sind keine Weiterentwicklungen im Internet nach 1997 zu finden.

Aufbau von μK

Syntaktische Analyse:	• Erster Schritt bei der Übersetztung • Speichert Zeitform, Satzbau, Geschlecht und Fall
Semantische Analyse:	• Löst Bezüge zwischen Sätzen auf • Bedeutung von Metaphern herausfinden
Ontologie:	• Notwendig um Bedeutungen zwischen den Sprachen zu übertragen. • Sprachgeneratoren der Zielsprache können Wissen teilen.

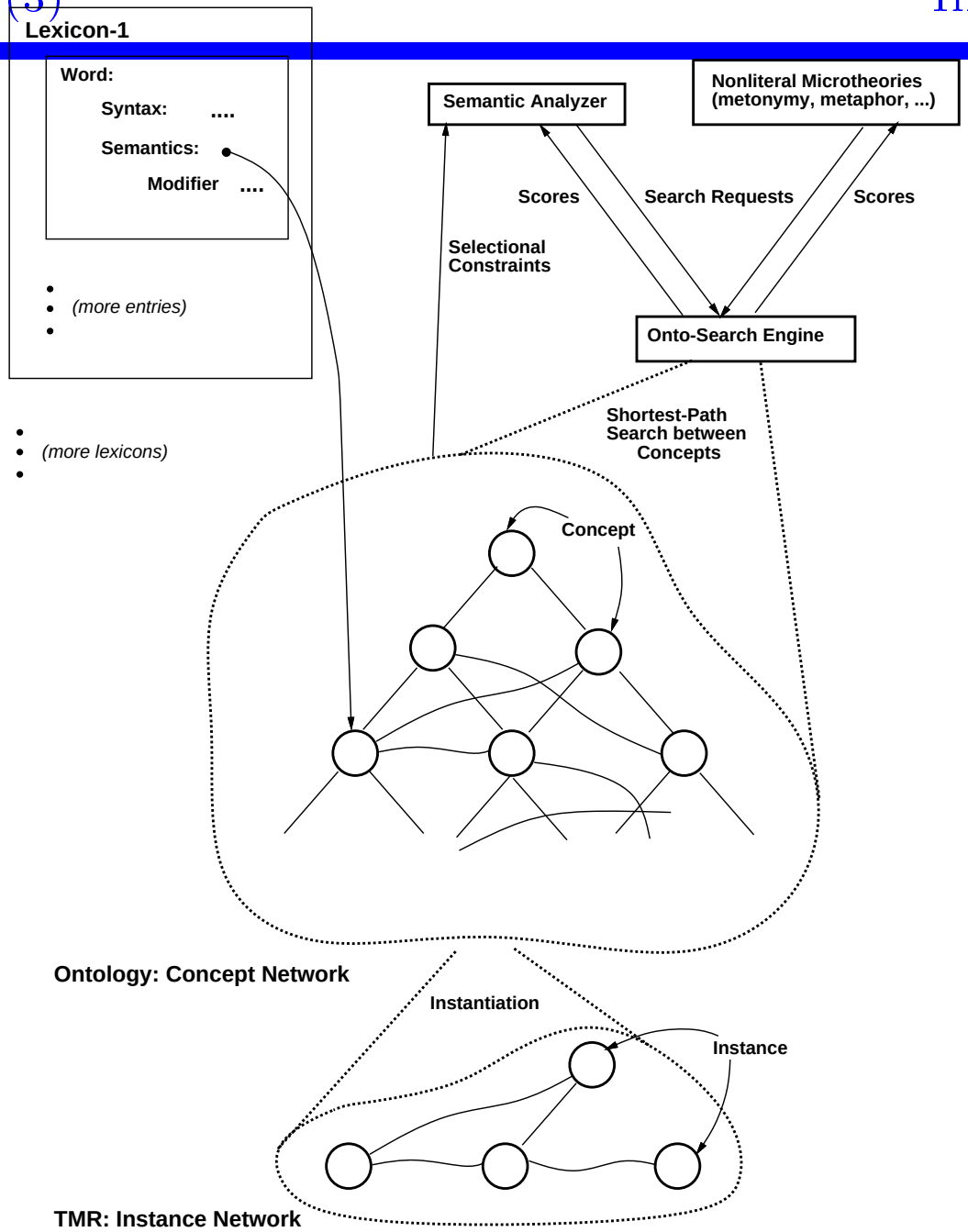
- Wegen Aufgabenstellung ist breite Abdeckung von Themen der realen Welt notwendig.
- Beinhaltet momentan 4000 Konzepte

Lexikas:

- Für die verschiedenen Sprache möglich.
- Bindeglied zwischen Ontologie und Text.
- 7000 im Spanischen

Sprachgeneratoren:

- Erzeugt aus TMRs (Text Meaning Representation) die Sätze in der Zielsprache.
- Bisher nur für Englisch


Lexicon-1
Word:
Syntax: ....
Semantics:
Modifier ....
(more entries)
(more lexicons)
Semantic Analyzer
Nonliteral Microtheories
(metonymy, metaphor, ...)
Scores
Search Requests
Scores
Selectional
Constraints
Onto-Search Engine
Shortest-Path
Search between
Concepts
Concept
Ontology: Concept Network
Instantiation
Instance
TMR: Instance Network

Probleme bei der Übersetzung

- Auflösen von Bezügen
- Auflösen von Mehrdeutigkeiten: Mit Hilfe von Ontosearch, das den kürzesten Pfad zwischen 2 Kategorien findet.

- Methapher richtig zuordnen: Die Bedeutung des Worts wird über die Kategorie ermittelt, die semantisch verwandt mit dem restlichen Kontext ist. Die semantische Verwandtschaft wird über die Struktur der Ontology herausgefunden.

- Lücken mit Defaults füllen: Bsp.: “John geht schwimmen, auch wenn das Drausen kalt ist” SUBSTRATE von SWIM ist WATER

Die Ontologie

Concepts

Concept als Datenstruktur

Bilden die Knoten des Graphen, deshalb ***node***s genannt.

- Namen (eindeutig)
- Menge von ***slot***s

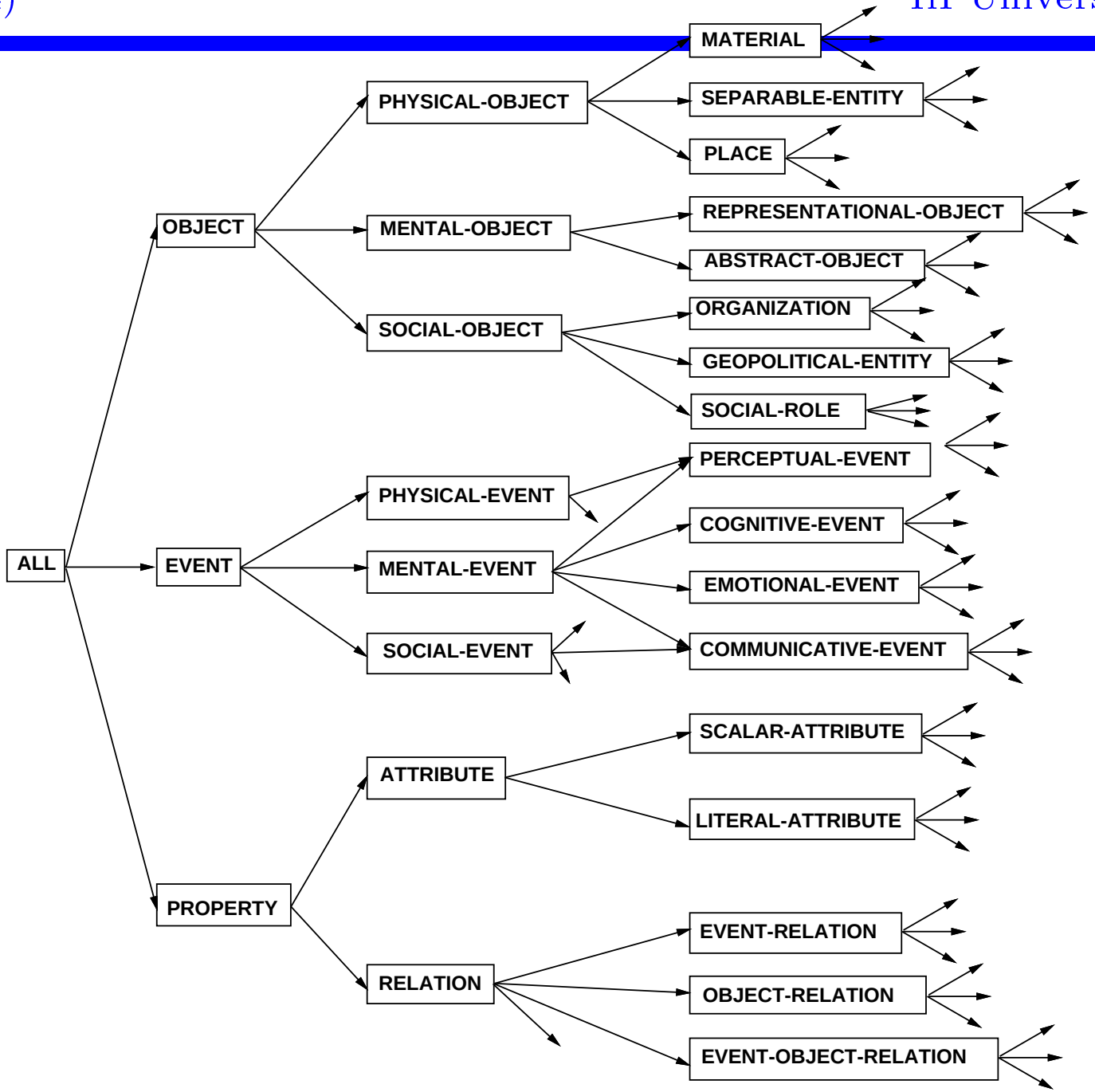
Folgende Strukturen sind ***node***s:

Object: Kategorie für ein Objekt. Kann alleine existieren und ist instanziierbar.

Event: Kategorie für ein Ergeignis. Kann auch alleine existieren und ist instanziierbar.

Property: Eine Eigenschaft ist nicht instanziierbar. Es wird unterschieden in:

- ***Relation***
- ***Attribute***


ALL
OBJECT
EVENT
PROPERTY
PHYSICAL-OBJECT
MENTAL-OBJECT
SOCIAL-OBJECT
MATERIAL
SEPARABLE-ENTITY
PLACE
REPRESENTATIONAL-OBJECT
ABSTRACT-OBJECT
ORGANIZATION
GEOPOLITICAL-ENTITY
SOCIAL-ROLE
PHYSICAL-EVENT
MENTAL-EVENT
SOCIAL-EVENT
PERCEPTUAL-EVENT
COGNITIVE-EVENT
EMOTIONAL-EVENT
COMMUNICATIVE-EVENT
ATTRIBUTE
RELATION
SCALAR-ATTRIBUTE
LITERAL-ATTRIBUTE
EVENT-RELATION
OBJECT-RELATION
EVENT-OBJECT-RELATION

Concept als ontologischer Begriff

- ***object***
- ***event***

Slots und fillers

- Ein ***slot*** ist entweder ein ***relation-slot*** oder ein ***attribute-slot***.
- Ein ***slot*** hat dann den Namen der ***relation*** oder des ***attribute***s. (Typisierung)
- Es gibt eine abgeschlossene Menge von Spezial-***slot***s.
- Spezielle ***slot***-Namen: sind nicht als ***property***s gespeichert.

Facets

Ein ***facet*** schreibt vor welche ***node***s diesen Teil-***slot*** belegen dürfen.

Value:	Wird normalerweise mit Instanz eines ***concept***s oder mit einem Skalar belegt.
Sem:	***Filler*** ist ein anderes ***concept*** oder ein skalarer Bereich
Default:	Wie ***value***. Wird benutzt bei Instanziierung wenn ***value-facet*** fehlt.

Measuring-Unit:	Quantisiert ***value***. Maßeinheit-***concept*** als ***filler***.
Salience:	Bezeichnet die Wichtigkeit eines ***slot***s für den ***node***.
Relaxable-to:	Wird nur in Lexikon benutzt. Bezeichnet wie hoch bei Metaphern gesucht werden soll. Sollte mit dem Top-***concept*** aller Möglichkeiten belegt werden.

Spezielle slots

Sie tauchen nicht als ***node*** in der Ontologie auf und haben meistens nur ein ***facet***.
Pflicht ***slot***s für alle ***node***s:

Definition:	Nur ***value-facet*** mit Englischer Text, der nur für den Menschen oder bestimmt ist.
Is-A:	Nur ***value-facet*** mit Liste der direkten Eltern. (Außer dem ALL ***concept***)
Subclasses:	Nur ***value-facet*** mit Liste der direkten Kinder. (Außer den Blättern)

Slots für ***property***s:

Domain: Pflicht für alle ***property***s. Hat ***sem-facet*** mit der Liste der ***concept***s.

Range: Pflicht für alle ***property***s .

- ***Relation***s: Liste der ***concept***s die als Wertebereich möglich sind.
- ***Attribute***s: Liste der Literale oder ein numerischer Wertebereich.

Inverse: Nur für ***relation***s. Hat nur ***value-facet*** mit der inversen Relation.

Weitere ***slot***s:

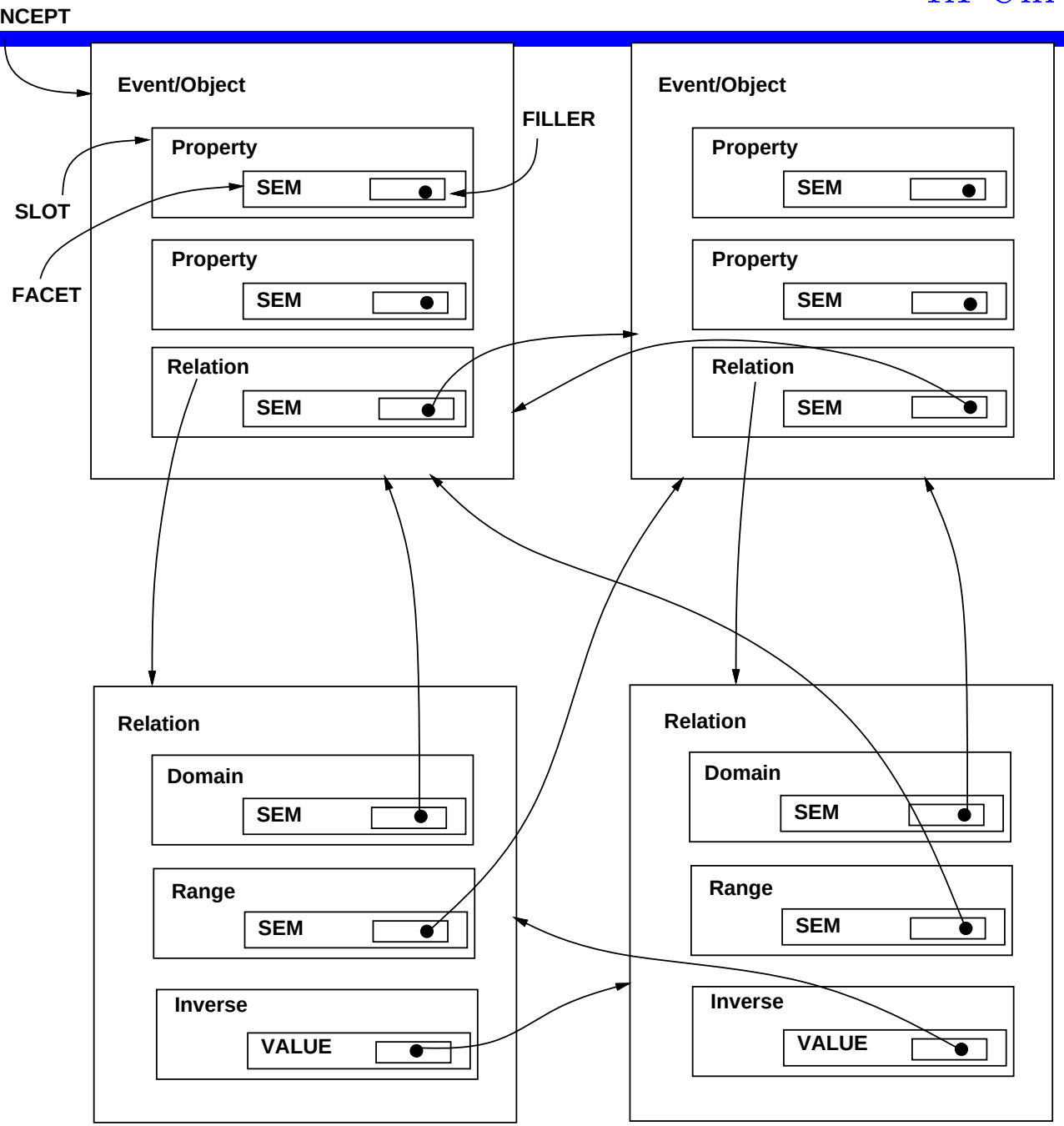
Instances:	Hat nur ***value-facet*** mit der Liste der Instanzen dieses ***concept***s
Instance-Of:	Pflicht für alle Instanzen. Hat nur ***value-facet*** mit dem parent ***concept***

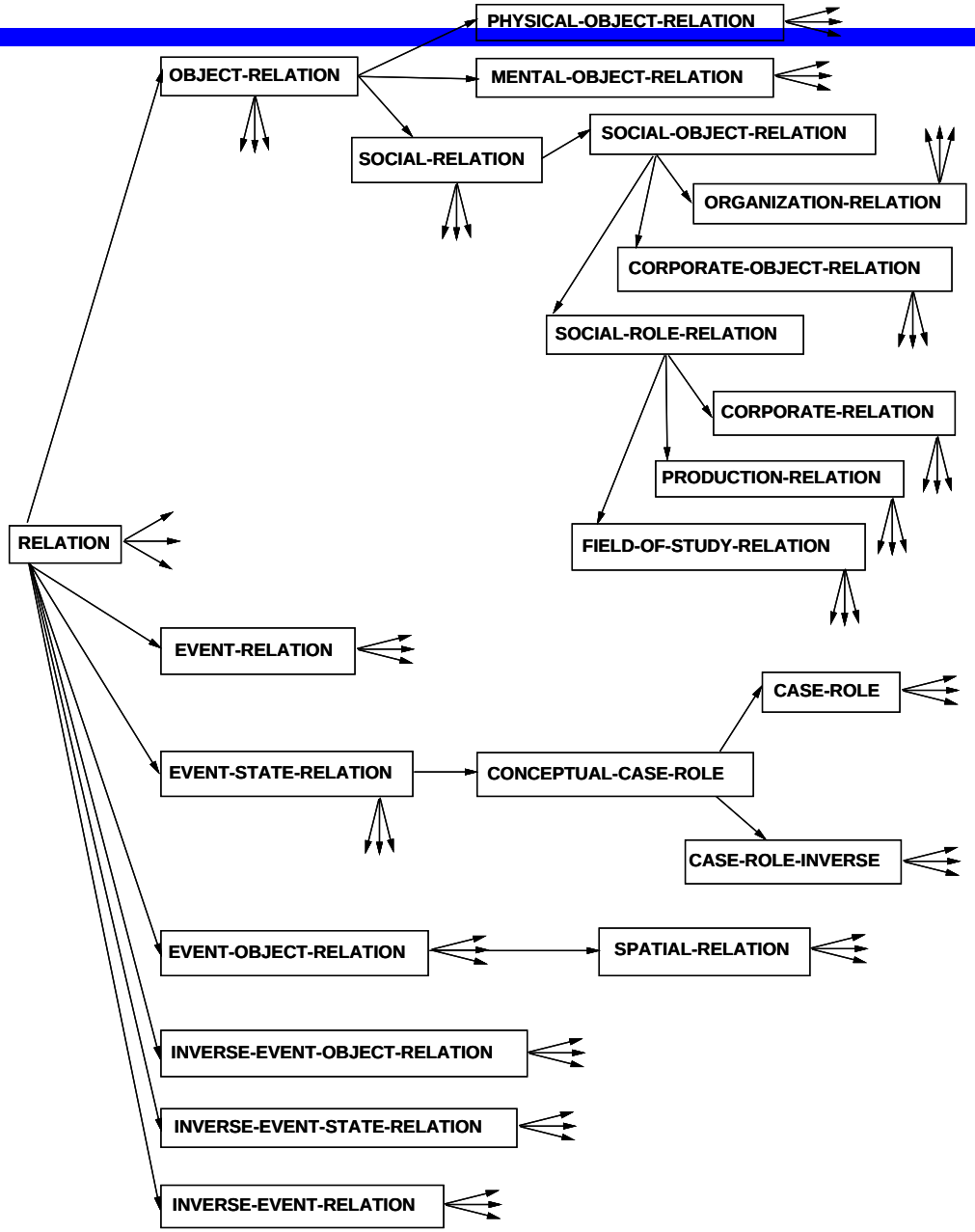
Kein ***facet*** ist Pflicht.

Relations

Jede ***relation*** ist als ***node*** in der Ontologie gespeichert.

Zu jeder Relation existiert eine inverse Relation.

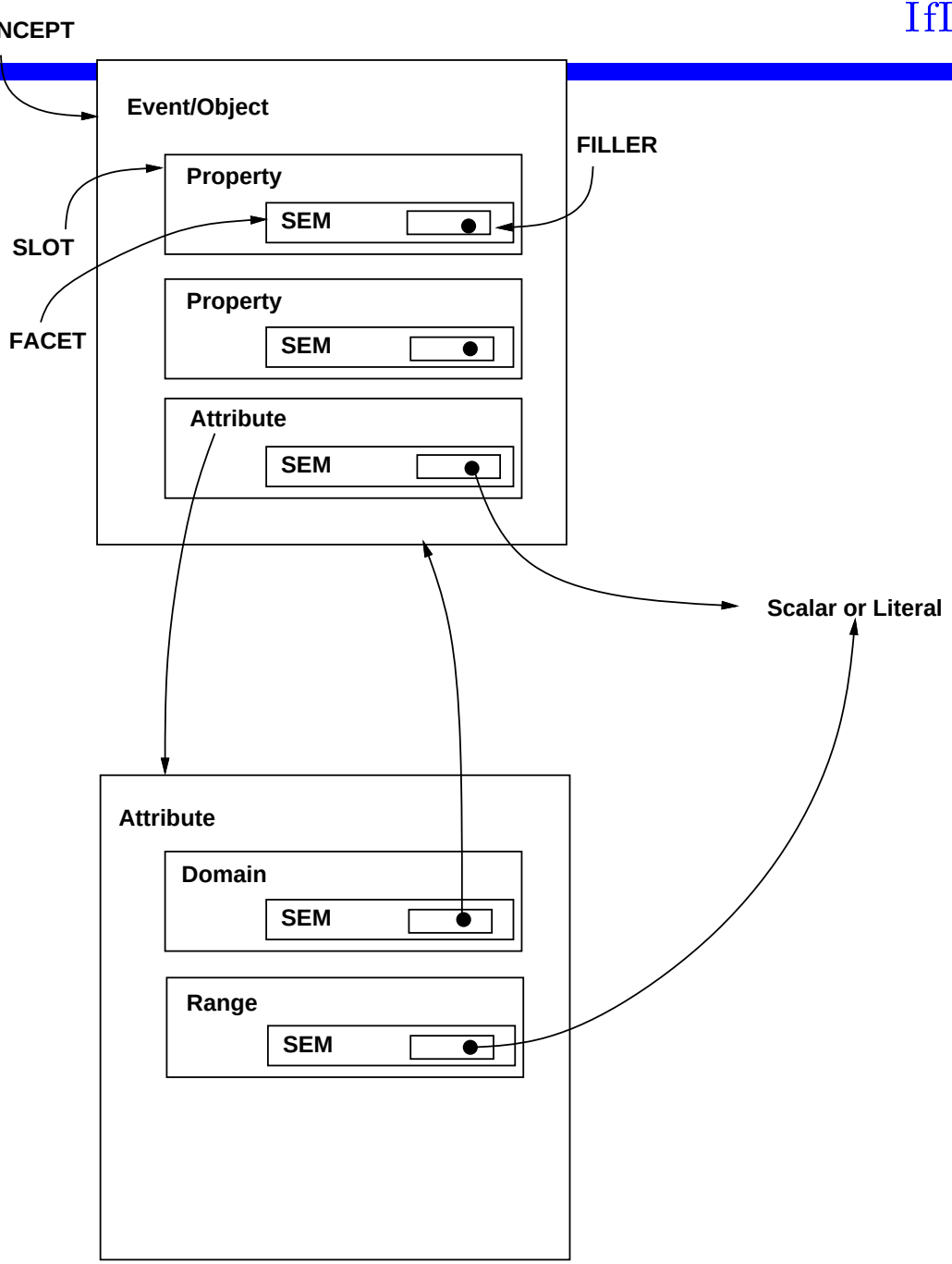

CONCEPT
Event/Object
Property
SEM
SLOT
FACET
FILLER
Property
SEM
Relation
SEM
Event/Object
Property
SEM
Property
SEM
Relation
SEM
Relation
Domain
SEM
Range
SEM
Inverse
VALUE
Relation
Domain
SEM
Range
SEM
Inverse
VALUE


PHYSICAL-OBJECT-RELATION
OBJECT-RELATION
MENTAL-OBJECT-RELATION
SOCIAL-OBJECT-RELATION
SOCIAL-RELATION
ORGANIZATION-RELATION
CORPORATE-OBJECT-RELATION
SOCIAL-ROLE-RELATION
CORPORATE-RELATION
PRODUCTION-RELATION
RELATION
FIELD-OF-STUDY-RELATION
EVENT-RELATION
CASE-ROLE
EVENT-STATE-RELATION
CONCEPTUAL-CASE-ROLE
CASE-ROLE-INVERSE
EVENT-OBJECT-RELATION
SPATIAL-RELATION
INVERSE-EVENT-OBJECT-RELATION
INVERSE-EVENT-STATE-RELATION
INVERSE-EVENT-RELATION

Attribute

Attributes werden in ***attribute-slot*** eingebettet.

Werte werden eingeschränkt


CONCEPT
Event/Object
FILLER
Property
SEM
SLOT
Property
SEM
FACET
Attribute
SEM
Scalar or Literal
Attribute
Domain
SEM
Range
SEM

Literale

Literale sind wichtig, um unendliche Ketten bei der Bedeutungsanalyse zu vermeiden.

- Sind entweder binär oder Werte auf einer beinhaltenden Skala.
- Binäre literale werden Attributspezifische Werte genutzt anstatt von “yes” und “no”. Werden genutzt wenn keine nummerische Skala besteht.

- Auch wenn es eine Skala gibt wird sie nicht unbedingt genutzt. Bsp: Farbe.
- Es wird nicht ***light-red*** eingeführt sondern eine Relation `GREATER-THAN` und `LESS-THEN` eingeführt.

TMR als erweiterter Instanzengraph

TMRs werden als Eingabe für den Sprachgenerator genutzt.

- Linguistische Informationen
- Instanzen von ***concept***s
- Instanzen sind auch ***node***s
- Unterscheiden sich nur in der Belegung der ***slot***s
- ***Value-facet*** sind mit konkreten Skalaren, Literalen oder Instanzen von ***concept***s belegt.

Beispiel

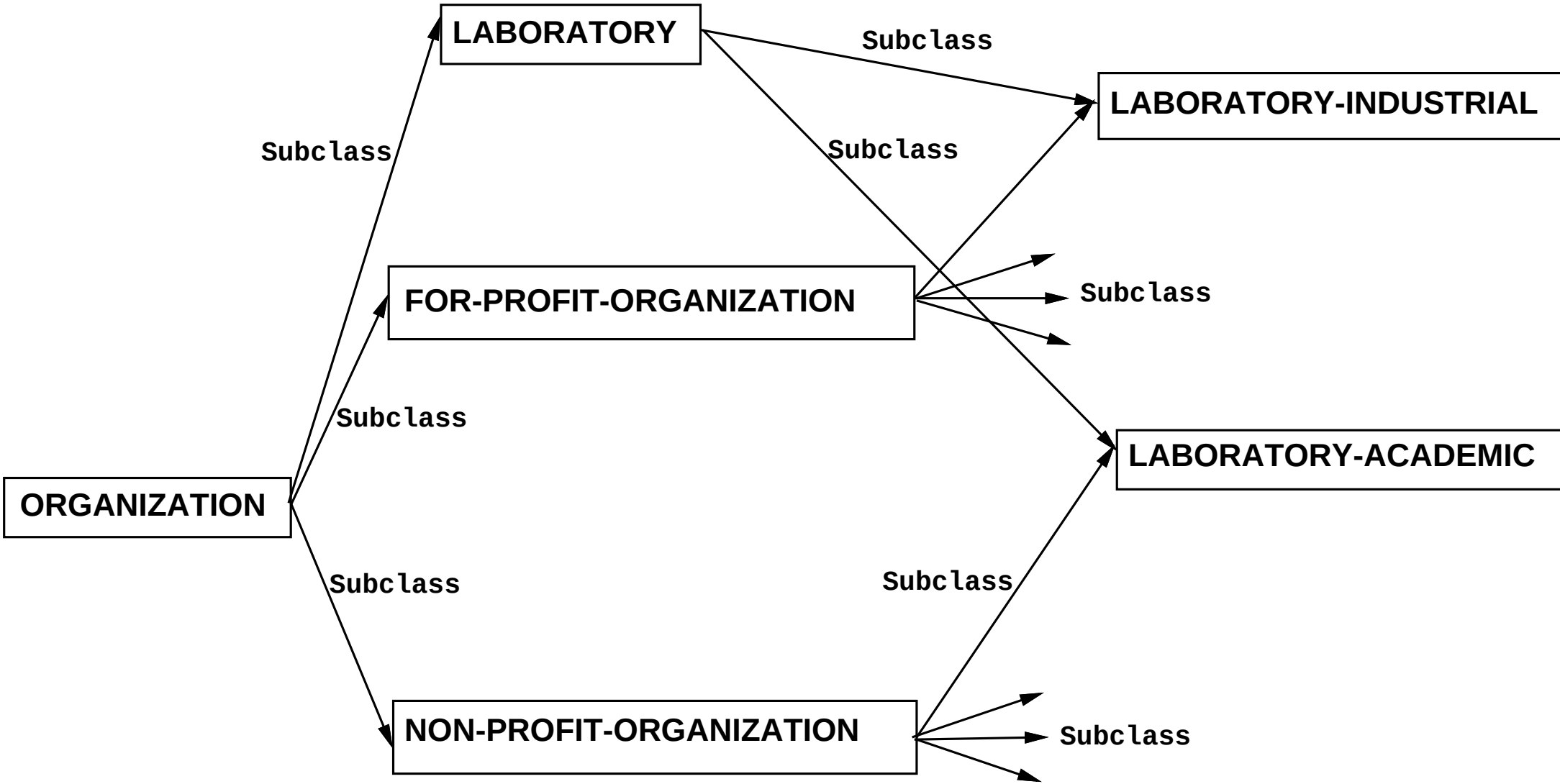
"I eat a green Apple." soll der Ausgangssatz sein.

conzepts:

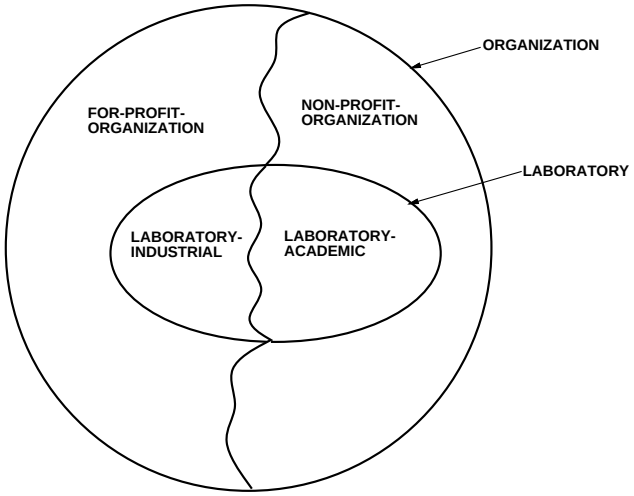
- `HUMAN` (***object***)
- `APPLE` (***object***)
- `EAT` (***event***)

Und-Graph

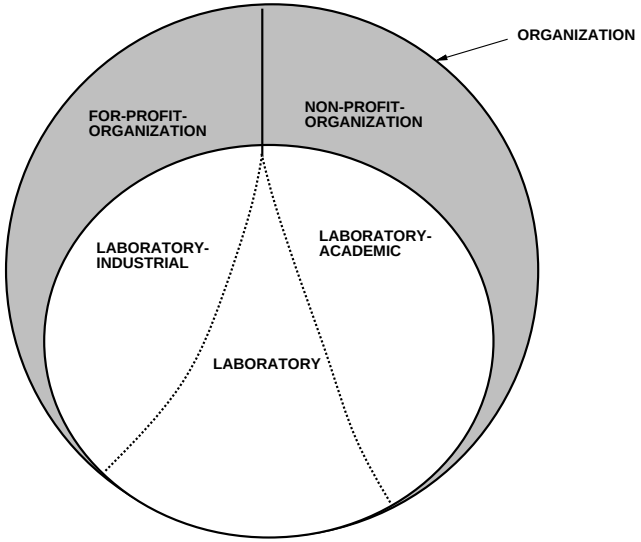
Problem bei Mehrfachvererbung: ist immer Konjunktion
(bezogen auf die Instanzen (Modelle))
Ein Beispiel soll dies verdeutlichen:


LABORATORY
Subclass
LABORATORY-INDUSTRIAL
Subclass
Subclass
FOR-PROFIT-ORGANIZATION
Subclass
Subclass
LABORATORY-ACADEMIC
ORGANIZATION
Subclass
Subclass
NON-PROFIT-ORGANIZATION
Subclass

Problem: Mögliche Fehlinterpretation und keine disjunkte Zerlegung des Elternkonzepts.



Intended Meaning: Organizations are either for-profit or non-profit; Laboratories are either for-profit or non-profit; All laboratories are either for-profit or non-profit organizations.



Possible Meaning: Organizations are either for-profit, non-profit, or laboratories. Some laboratories are for-profit and some are non-profit. There are laboratories that are neither for-profit nor non-profit organizations.

Prinzipien und Probleme der Ontologieentwicklung

Anforderungen an die Ontologie

- Sprachunabhängigkeit

  (keine 1:1-Abbildungen, Idealfall: 2 Lexikas werden neben der Ontologie entwickelt)

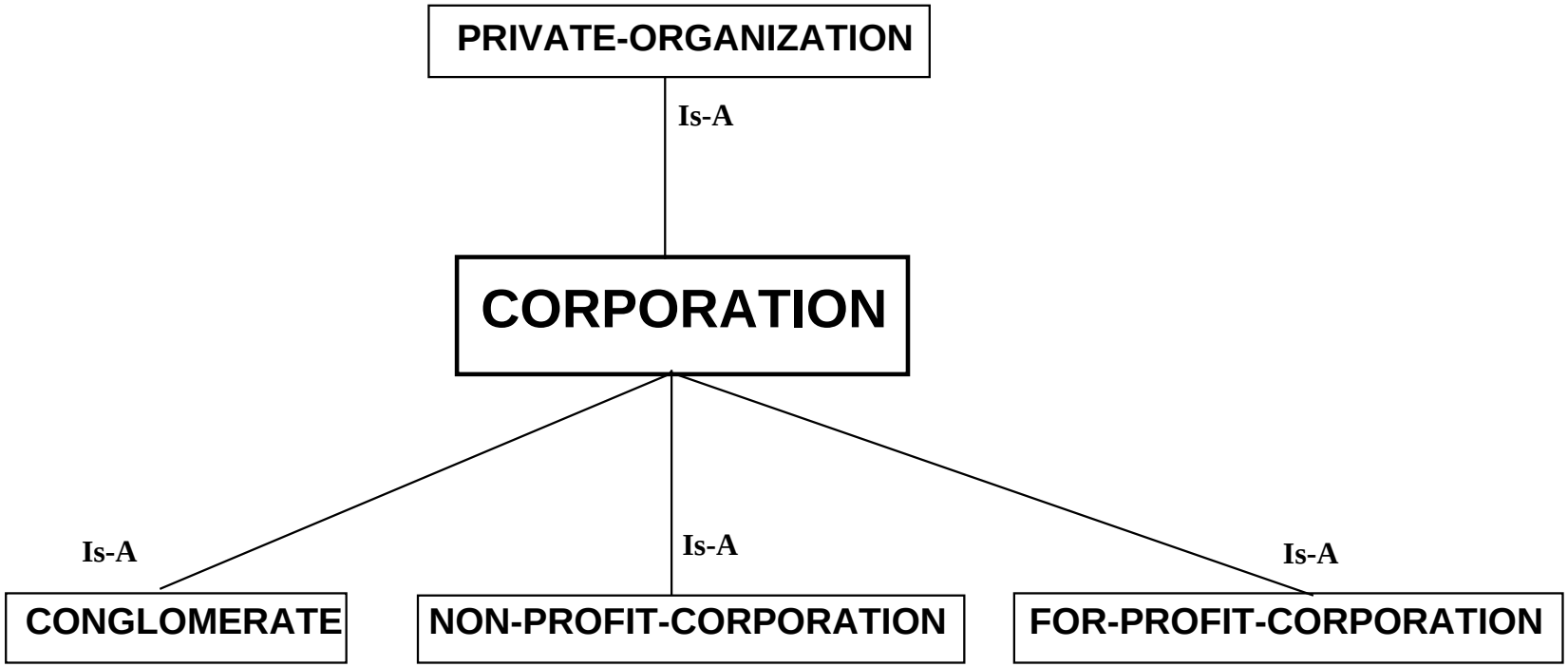
- Entwicklung der Ontologie darf sich nicht der der Lexikas unterordnen

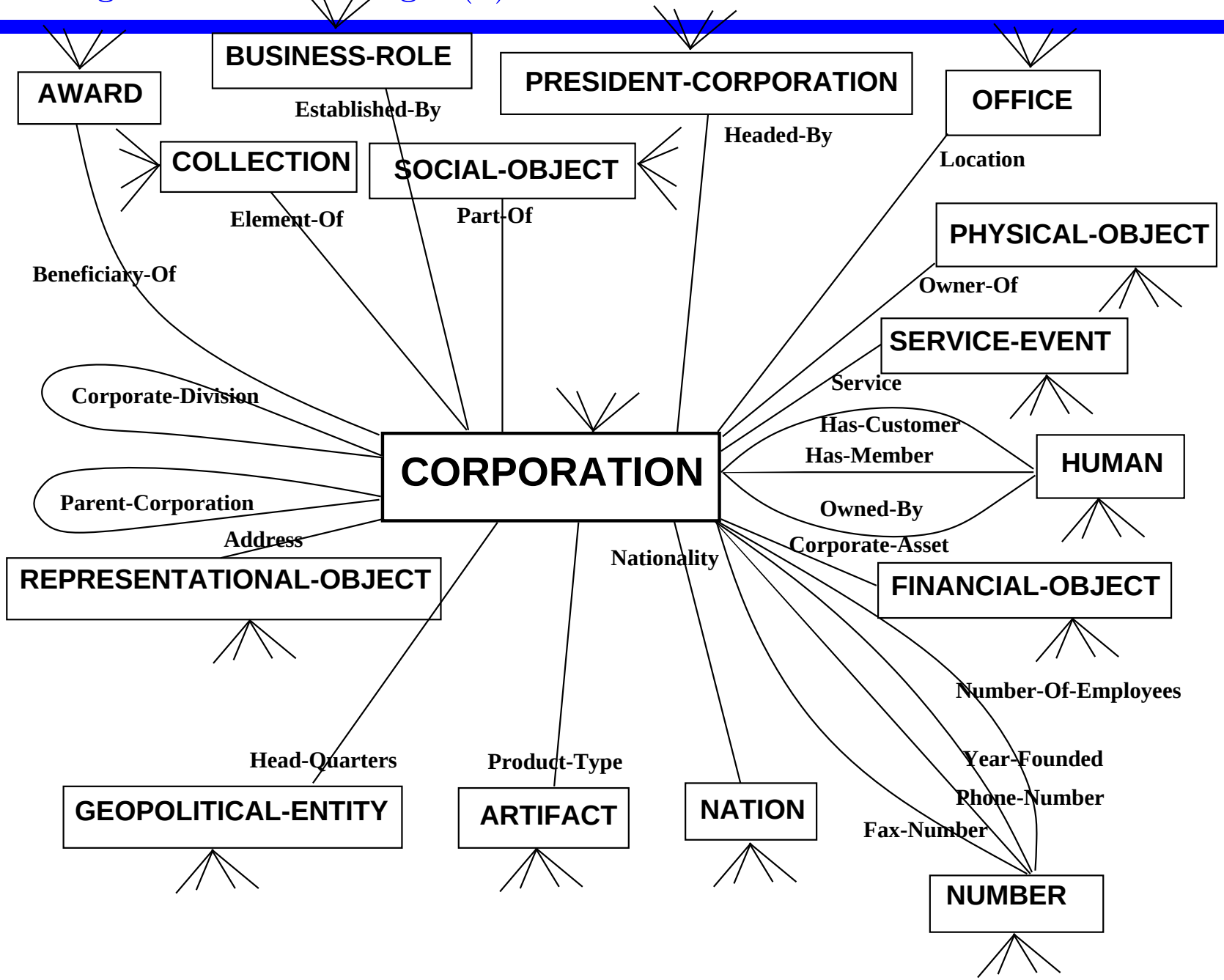
- Wollgeformtheit entsprechend den axiomatischen Spezifikation

- Konsistenz und Kompatibilität mit den anderen Teilen von μK

- reich an Inhalt, hochgradige Vernetzung der Konzepte

  keine bloße Hierarchie von Konzeptnamen




BUSINESS-ROLE
AWARD
PRESIDENT-CORPORATION
OFFICE
Established-By
Headed-By
Location
COLLECTION
SOCIAL-OBJECT
Element-Of
Part-Of
PHYSICAL-OBJECT
Beneficiary-Of
Owner-Of
SERVICE-EVENT
Service
Corporate-Division
Has-Customer
Has-Member
CORPORATION
HUMAN
Parent-Corporation
Owned-By
Address
Corporate-Asset
Nationality
REPRESENTATIONAL-OBJECT
FINANCIAL-OBJECT
Number-Of-Employees
Head-Quarters
Product-Type
Year-Founded
Phone-Number
GEOPOLITICAL-ENTITY
ARTIFACT
NATION
Fax-Number
NUMBER

- Verständlichkeit (statische Darstellung der Konzepte)
- Nutzen für NLP steht im Vordergrund

  Mehrdeutigkeiten auflösen und notwendige Ableitungen in beliebige Tiefe durchführen
- beschränktes Themengebiet
- kein episodisches Wissen (andere Methodologie, s. Onomasticon)
- Automatisierung der Entwicklung: Graphische Browser, Halbautomatisierte Benutzerinterfaces, Programme zum Erkennen von Inkonsistenzen und Richtliniendurchsetzung

Strukturelle Prinzipien

Entwicklung der Ontologie nicht mit vorgefertigter Struktur

→ Modellierung der Ontologie nach der Welt und den Bedürfnissen

Allgemeine Prinzipien von Ontologien

- **Gleichartigkeit**: Ein Kindkonzept muß die Bedeutung des Oberkonzeptes teilen.

  μK : Man kann ***property***s bei der Vererbung mit `*nothing*` blocken $\rightarrow$ Nicht-Monotonie.

- **Spezifizität**: Ein Kind-Konzept muß sich von seinen Eltern unverkennbar unterscheiden.

  μK : Diskriminator wird nicht repräsentiert.

  Beispiel: Unterschied von `ANIMAL` und `INVERTEBRATE` ist nicht dargestellt in der Ontologie.

- **Gegensatzprinzip** Ein Konzept muß sich von seinen Geschwistern unterscheiden und der Unterschied muß repräsentiert werden.

  μK : nicht notwendig, wir brauchen keine Vollständigkeit, Vererbung

  Zwischen `WALK` und `RUN` wird keine Abgrenzung vorgenommen.

- **eindeutige semantische Achse**: Alle Geschwister müssen sich in einer Art und Weise unterscheiden.

  μK : Zu restriktiver Ansatz. Mehrere semantische Achsen mittels ***property***s und ***concept***s möglich.

  Beispiel: `ORGANIZATION` mit dem Kind `NON-PROFIT-ORGANIZATION` und `LABRORATORY` und der ***property*** `CUSTOMER-CONTACT-ATTRIBUTE` (repräsentiert eine Dritte Achse)

Redundanz

- Dualität zwischen ***object***s und ***property***s,

  Wie modelliert man "headquarter of a corporation"?

  a) ***concept*** `ORGANIZATION` noch in `HEADQUARTER` und `BRANCHES` unterteilen oder

  b) als ***attribute***

- Dualität zwischen ***event***s und ***property***s,

  Oft gibt es bei ***event***s mehrere Phasen:

  – Beginphase, `FALL-ASLEEP` oder `ACQUIRE`

  – fortlaufende Phase, ***state***, `FALL-ASLEEP` oder `ACQUIRE`

– Endphase, `WAKE-UP` oder `RELINQUISH`

Möglichkeiten: `OWN` als ***Event*** oder eine `OWNED-BY` und `OWNED-OF` ***property***

`ACQUIRE` ist ein zentrales Konzept von μK wird als ***event*** modelliert.

- Dualität zwischen ***object***s und ***event***s

  Das ***event*** (`OWN`) und ***object*** (`LABRORATORY`) müssen mit einer ***property*** `OWNED-BY` verbunden werden. Außerdem muß `ACQUIRE` mit `OWNED-BY` verbunden werden, damit im TMR eine Verbindung zwischen den beiden existiert.

Erfassung von Concepts

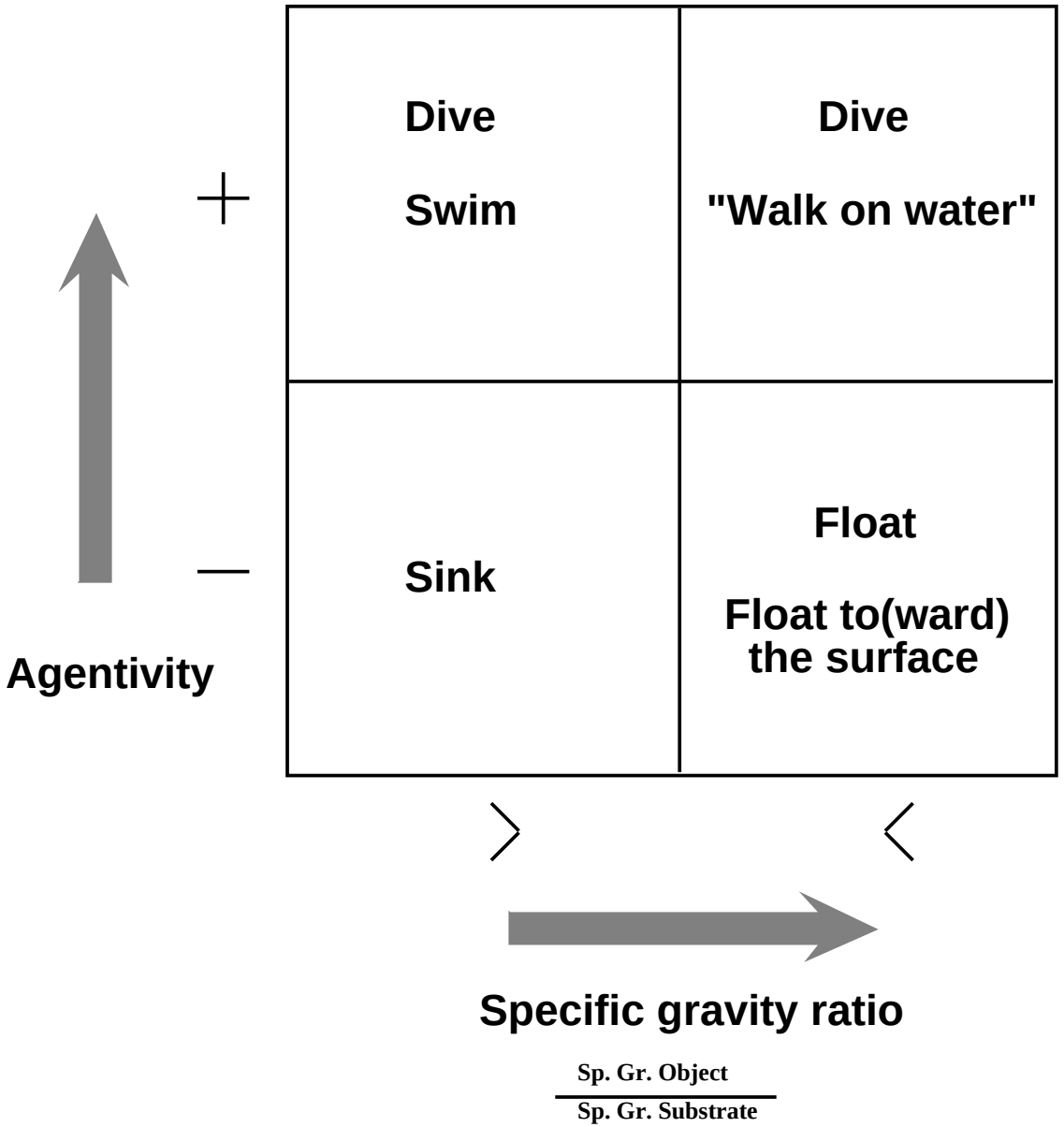
- gegeben: zwei englische Worte `swim` und `float`
- Bedeutung der Wörter verdeutlichen

  English: Swim → `SWIM`, Float → `FLOAT`

  Spanish: Nadar → `SWIM`, Flotar → `FLOAT`

  Russion: Plyt → `SWIM` or `FLOAT`, Plabat → `SWIM` or `FLOAT`

  Bedeutung der Wörter in verschiedenen Sprachen


Dive
Swim
Dive
"Walk on water"
Sink
Float
Float to(ward) the surface
+
—
Agentivity
Specific gravity ratio
Sp. Gr. Object
Sp. Gr. Substrate

- Dimension finden, in der sich die Worte unterscheiden... `AGENTIVITY` (aktive Rolle im Geschehen haben vs. passiv teilhaben)
- ähnliche Konzepte gleich miterfassen (`dive` und `sink`)
- weitere Analysen sind nicht sinnvoll bei dieser Domäne

Integration anderer Ontologien

Diese Punkte müssen bei der Integration sichergestellt werden, um die Qualität zu sichern.

a) breite Themenbasis, d.h. Fachbegriffe herauslassen (keine Terminologie aus der Nervenchirurgie)

b) viele ***properties***, d.h. vor allem ein hoher Verbindungsgrad

c) Leichte Verständlichkeit, einfache Suche und Browsing

d) Wirtschaftlichkeit

Richtlinien

Welche Konzepte in die Ontologie

- ***concept***s nicht hinzufügen nur weil es möglich ist.
- bei ***concept***s mit ähnlicher Bedeutung, daß allgemeine nehmen und ein ***attribute*** für die Abstufungen einführen
- keine ***event***s mit vielen Argumenten, d.h. `walk-to-airport-terminal` und `walk-to-parking-lot` vermeiden.
- englische Worte, keinen Plural, Konzept

Namenskonventionen

Es sind nur alphanummerische Zeichenketten mit Bindestrichen erlaubt.

- normale Regeln: kein Plural, keine Umgangssprache, zusammengesetzte Wörter nicht zweideutig wählen (`unit-of-time` statt `time-unit`), Namensdopplung ausschließen (bei ***properties*** und ***objects***: `employee` als ***object*** und `employed-by` für ***property***)
- ***concept*** Namen: englisch, max. vier Wörter mit Bindestrichen getrennt.

- ***instance*** Namen: Namen der Instanz + ***concept*** Name + Integer (jeweils mit – getrennt)
- Literale: englisch, meist ein Wort.
- Weiter Symbole sind: mathematische Symbole, Lexikonische Symbole, TMR-Symbole

Qualitätssicherung

- Erfassung bleibt lenkbar (man kann sich am Ziel orientieren; keine Automatisierung)
- Waisen entfernen, da jedes Konzept mit jedem auf Ähnlichkeit vergleichbar bleiben muß
- Computerunterstützung: fehlende Erklärungstexte, falsch geschriebene Bezeichner, Waisenerkennung, Relationen ohne inverse Relationen
- Nutzer finden die tieferliegenden Redundanzen und Inkonsistenzen

Resourcen

Paper:

Ontology Development for MT: Ideology and Methodology,
Kavi Mahesh, 1995

Internet:

`http://crl.nmsu.edu/users/mahesh/onto-intro-page.html`