

Wintersemester 2006 • Institut für Germanistik I

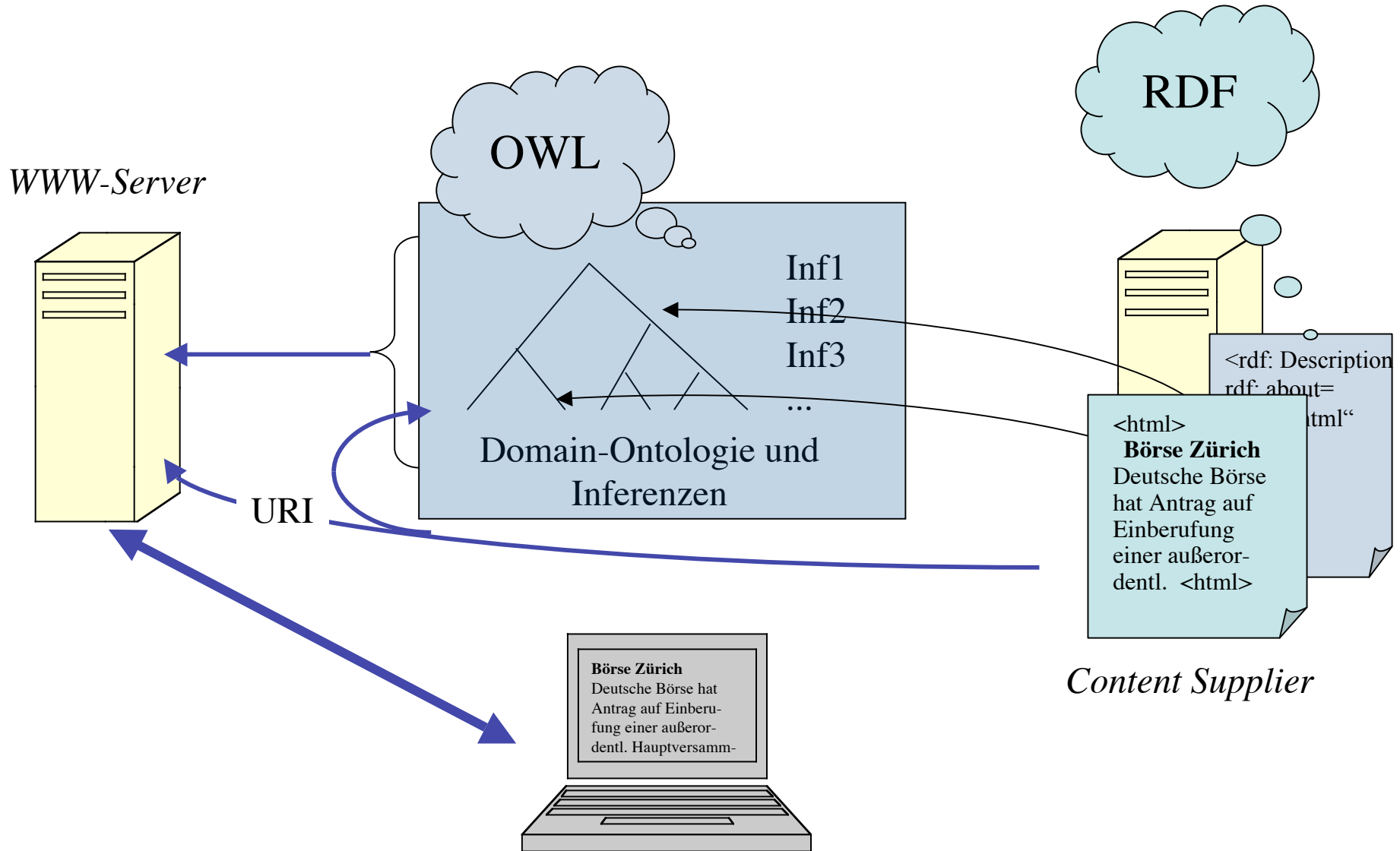
Vorlesung  
Computerphilologie

# Semantic Web

Wozu benutzt man Ontologien?

Warum sind Ontologie heute als Technologie so interessant?

# Was ist das Semantic Web?



# Ziel des Semantic Web

- Das Semantic Web (nach Tim Berners-Lee 1998) dient dem verbesserten Informationszugang im WWW durch Referenz aus der Site auf standardisierte gemeinsame semantische Meta-Daten:
  - Ontologien als Beziehungen zwischen Domain-Objekten
    - Z.B. “Trakehner” ist Unterbegriff von “Pferd”, und
  - Inferenzen über Ontologie-Objekten
    - Z.B. “Wenn  $A \subset B$  und  $B \subset D$ , dann  $A \subset D$  (Transitivität):  
“Ein Trakehner ist ein Tier, weil alle Trakehner Pferde und alle Pferde Tiere sind”

# Der Stand der Technik und die Verbreitung

- Es gibt inzwischen eine Reihe von fachbezogenen Ontologien, die von vereinzelt Web-Anbietern benutzt werden.
- Mit RDF(S) und OWL gibt es inzwischen einen einigermaßen festen Standard, der die zahlreichen vorherigen Formalismen (nur XML, OIL, DAML-OIL, RDF) ablöst bzw. zusammenfasst.
- Es gibt noch wenig Verbindung zwischen Semantic Web Forschung, Fachsprachenforschung und Terminologien bzw. Dokumentationswissenschaft.
- Es gibt sehr viele Verbindungen zwischen Semantic Web Forschung und Knowledge Management, bzw Betriebswirtschaft.

# Warum müssen sich (Fachsprachen-)Linguisten damit befassen?

- Fachsprachen-Linguisten haben sich immer mit der Beschreibung von Fachtexten für praktische Bedarfe befasst. Die praktischen Bedarfe wechseln inzwischen immer mehr vom Papier zu Computeranwendungen.
- Fachsprachler haben immer mit Terminologen zusammengearbeitet, die die (problematische) Ähnlichkeit, - wenn nicht wo möglich Identität -, einer Gegenstands-Systematik, einer Begriffssystematik und eines terminologischen Systems anstreben.
- Besonders ist die internationale Terminologienormung und deren Einhaltung in der Praxis wegen der unterschiedlichen Sachsysteme und historisch gewachsenen Überlappungsphänomenen immer problematisch.
  - Die korrekte Korrespondenz von engl. *attorney* im Deutschen oder Finnischen zu finden, scheitert zunächst an unterschiedlichen Rechtssystemen und dann an der Unkenntnis über das Fach und über die Textsorte.

# Ähnlich im WWW

Dieselben Probleme hat das Semantic Web, aber

- die Texte sind meist sehr heterogen oder in sich unklare Mischungen mit Alltagstexten, z.B. in Werbe-Sites,
- Texte in verschiedenen Sprachen oder mehrsprachige Texte sollen gefunden werden,
- es gibt Texte, die auf ein anderssprachiges Sachsystem referieren, und
- die Gemeinsamkeiten sind oft vage.

Umso mehr sind Fachsprachen-Spezialisten gefragt

# (Grubers) Ontologien

- Gruber'sche Ontologien (oft auch Taxonomien genannt) sind Objektnamen einer Domäne (Fach, Sachgebiet, z.B. Produkte, Hobbies, Orte, etc) mit einer homogenen Klasse von begrifflichen Beziehungen unter ihnen, wie Ober-/Unterbegriff oder Begriff/Instanz oder Teil/Ganzes (Meronymien).

*Vgl. Tom Gruber, A Translation Approach to Portable Ontology Specifications. Knowledge Systems Laboratory. Stanford Univ. Tech. Report KSL 92-71*

- Die Inferenzen sind eine zusätzliche Forderung des Semantik Web, abgeleitet aus der älteren Forschung über Künstliche Intelligenz und dem Wissensmanagement.

# Berners-Lee's Layer Cake

Trust	
Proof	
Logic	
XML Familie	Ontologies & OWL
	RDF Schemas
	RDF, Topic Maps, ähnliche Technologien
	XML-Schemas, DTDs, Namespaces
	XML-Dokumente
URI	Unicode



# Die nötigen Darstellungsformalismen

- Es muss für die WWW-Texte eine deklarative **Annotation** (außer HTML) für die Stellen und Inhalte geben, von denen aus man Ontologien befragt:  
⇒ Formalismus RDF, geschrieben in der Sprache XML
- Es muss für die Ontologien eine deklarative **Repräsentation** geben, um sie aus WWW-Texten heraus zu benutzen zu können:  
⇒ Formalismus RDF(S) und OWL, geschrieben in der Sprache XML

Deklarativ = Propositional und positiv  
aufgeschriebenes Wissen (z.B. Lexikoneintrag)

Prozedural = an einen Vorgang gebundenes  
Wissen (z.B. Programm)

# Das RDF Model (Resource Description Framework)

- Ist eine W3C Empfehlung, die Metadaten für Web-basierte **Ressourcen** standardisiert.
- Jeder Block in RDF ist ein Tripel <Objekt-Attribut-Wert> („Subjekt“, „Prädikat, „Objekt“):
- authorOf(‘http://www.w3.org/employee/id1321’,  
‘http://www.books.org/ISBN0621’)

*Attribut*

*Wert*

*Objekt*

XML  
Serialisierung

```
<rdf:Description rdf:about:"http://www.w3.org/employee/id1321">  
  <authorOf rdf:ressource="http://www.books.org/ISBN0621"/>  
</rdf:Description>
```

# RDFS (RDF Schema)

- Beschreibt die grundlegenden Elemente, die für die Modellierung der Ontologien benutzt werden können.
- Man kann **Klassen, Unterklassen, Merkmale und Untermerkmale** definieren.
- RDFS gibt keine syntaktischen Bedingungen an.
- Es gibt Schlüsselwörter, wie: Class, subPropertyOf, subClassOf etc.

```
<rdfs: Class rdf: about="HardCover">
```

```
<rdfs: subClassOf rdf:resource=#Book"/>
```

```
</rdfs: Class>
```

# OWL -1-

- Veröffentlichung & Austausch von Ontologien
- Dedizierte Beschreibungs-Sprache
- Austausch-Syntax ist RDF/XML
- Elemente
  - Taxonomische Beziehungen zwischen Klassen
  - Eigenschaften und Datentypen
  - Objekteigenschaften (Eigenschaften von Individuen)
  - Instanzen von Klassen und Eigenschaften

# OWL -2-

- Menge von OWL-Aussagen & Schlussfolgerungen = Wissensbasis (Knowledge base (KB) )
- Typen von OWL:
  - Drei zunehmend komplexe Sprachen
    - OWL Lite
    - OWL DL
    - OWL Full

# OWL-Header

- Header
  - Namensräume (rdf, rdfs, owl, dc, ...)
  - Kommentare
  - Versionskontrolle (lite, ...)
  - Verbindungen anderer Ontologien

## Ein Beispiel für die Annotierung von (Fach-)Texten

- Verbindung eines historischen Dokuments mit einer Ontologie
- Beispiel: CIDOC-CRM Ontology
- Entwickelt am ICS-FORTH- Heraklion-Kreta ([http://cidoc.ics.forth.gr/comprehensive\\_intro.html](http://cidoc.ics.forth.gr/comprehensive_intro.html))
- Ziel: Beschreibung von künstlerischen und historischen Objekten für Museen.

## Protocol of Proceedings of Crimea Conference

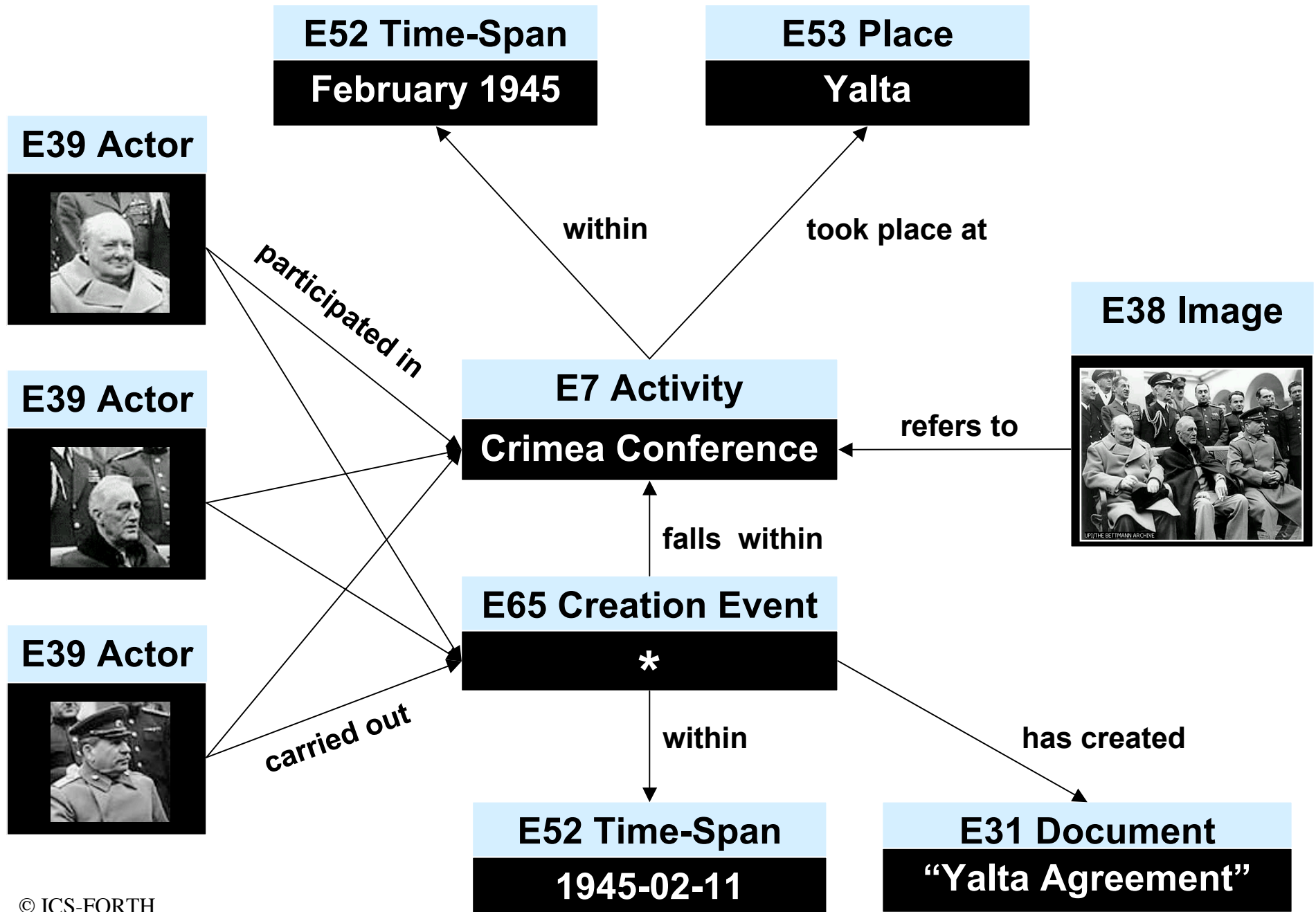


### Allied Leaders at Yalta

*“The following declaration has been approved:*

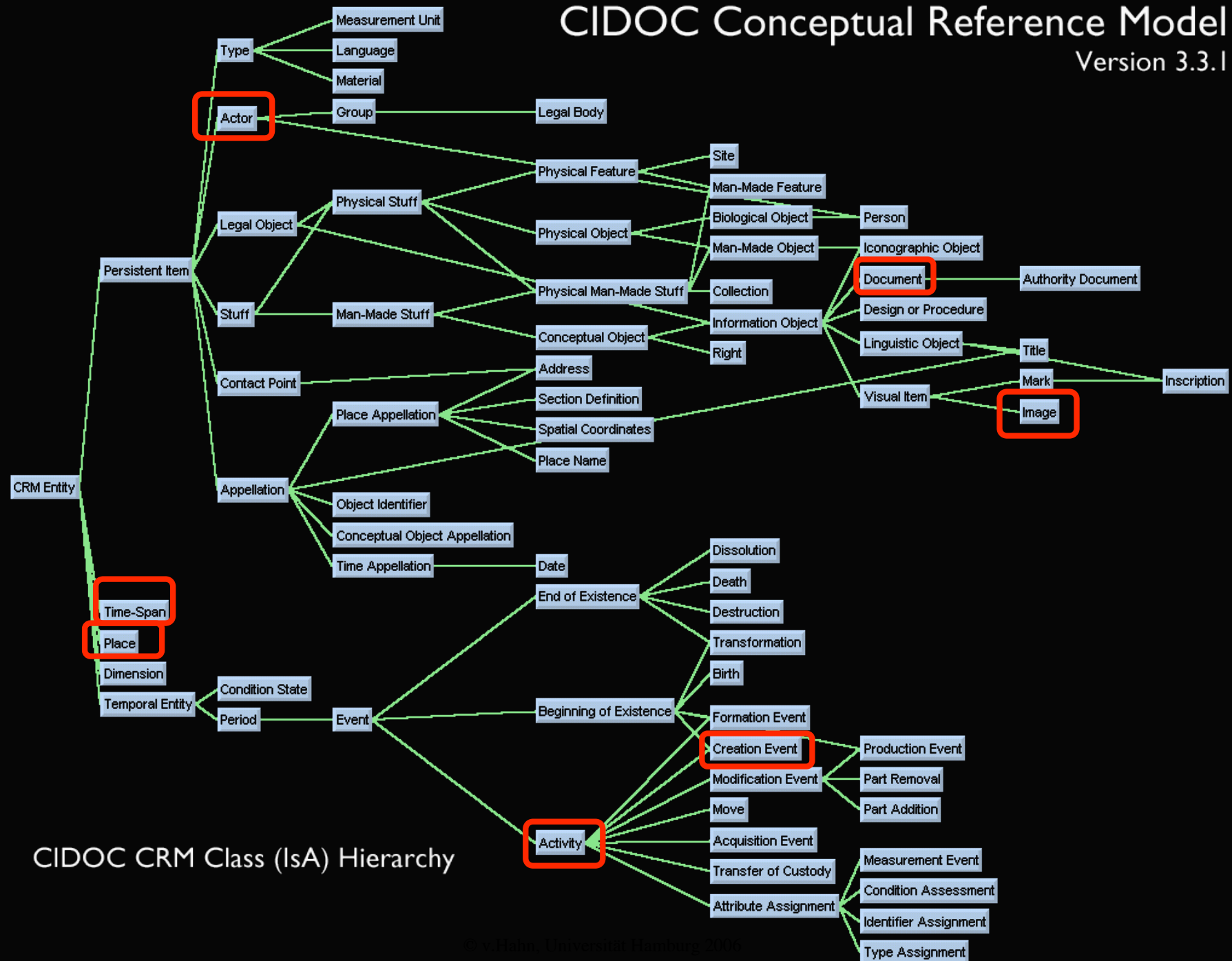
The Premier of the Union of Soviet Socialist Republics, the Prime Minister of the United Kingdom and the President of the United States of America have consulted with each other in the common interests of the people of their countries and those of liberated Europe. They jointly declare their mutual agreement to concert... and to ensure ...”





# CIDOC Conceptual Reference Model

Version 3.3.1



CIDOC CRM Class (IsA) Hierarchy

# Klassenkodierung in einer OWL-Fassung -1-

```
<?xml version=„1.0“?>
```

```
<!--
```

```
CIDOC CRM v3.4.9 Encoded in OWL>
```

```
<rdf:RDF xml:lang=„en“
```

```
  xmlns:rdf=„http://www.w3.org/1999/02/22-rdf-syntax-ns#“
```

```
  xmlns:rdfs=„http://www.w3.org/2000/01/rdf-schema#
```

```
  xmlns:owl =http://www.w3.org/2002/07/owl#>
```

```
<owl:class rdf:ID=„E7.Activity“>
```

*(„Wert / Subject“)*

```
  <rdfs:subClassOf rdf:resource=#E5.Event“/>
```

*(„Prädicat/ Attribut“)(„Object“)*

```
</owl:Class>
```

```
<owl:class rdf:ID=„E31.Document“>
```

```
  <rdfs:subClassOf rdf:resource=#E73.InformationObject“/>
```

```
</owl:Class>
```

```
<owl:class rdf:ID=„E38.Image“>
```

```
  <rdfs:subClassOf rdf:resource=#E36.Visualltem“/>
```

```
</owl:Class>
```

## Klassenkodierung in OWL-Fassung -2-

```
<owl:class rdf:ID=„E39.Actor“>  
  <rdfs:subClassOf rdf:resource=#E77.Persistent_Item“/>  
</owl:Class>  
<owl:class rdf:ID=„E52.Time-Span“>  
  <rdfs:subClassOf rdf:resource=#E1.CRM_Entity“/>  
</owl:Class>  
<owl:class rdf:ID=„E53.Place“>  
  <rdfs:subClassOf rdf:resource=#E1.CRM_Entity“/>  
</owl:Class>  
<owl:class rdf:ID=„E65.Creation_Event“>  
  <rdfs:subClassOf rdf:resource=#E7.Activity“/>  
  <rdfs:subClassOf rdf:resource=#E63.Beginning_of_Existence“/>  
</owl:Class>
```

Mehrfachvererbung

# Merkmalskodierung in OWL -1-

```
<owl:ObjectProperty rdf:ID=„P7F.took_place_at“>  
  <rdfs:domain rdf:resource=#E4.Period“/>  
  <rdfs:range rdf:resource=„E53.Place“/>  
</owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:ID=„P10.falls_within“>  
  <rdfs:domain rdf:resource=#E4.Period“/>  
  <rdfs:range rdf:resource=„E4.Period“/>  
</owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:ID=„P11B.participated_in“>  
  <rdfs:domain rdf:resource=#E39.Actor“/>  
  <rdfs:range rdf:resource=„E5.Event“/>  
  <rdfs:subPropertyOf rdf:resource=„#P12B.was_present_at“/>  
</owl:ObjectProperty>
```

## Merkmalskodierung in OWL -2-

```
<owl:ObjectProperty rdf:ID=„P67F.refers_to“>  
  <rdfs:domain rdf:resource=#E73.Information_Object“/>  
  <rdfs:range rdf:resource=„E1.CRM_Entity“/>  
</owl:ObjectProperty>
```

```
<owl:ObjectProperty rdf:ID=„P86F.within“>  
  <rdfs:domain rdf:resource=#E52.Time-Span“/>  
  <rdfs:range rdf:resource=„E52.Time-Span“/>  
</owl:ObjectProperty>
```

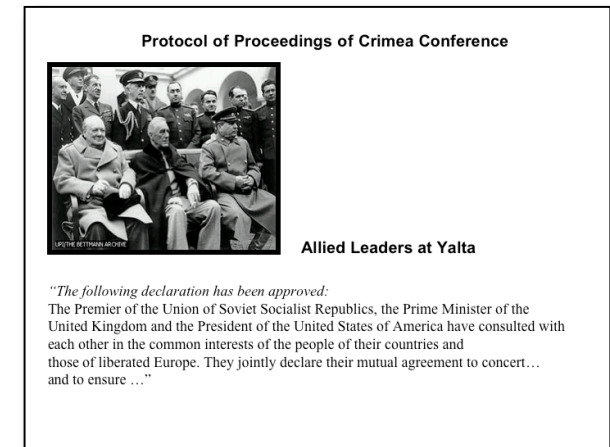
```
<owl:ObjectProperty rdf:ID=„P49F.has_created“>  
  <rdfs:domain rdf:resource=#E65.Creation_event“/>  
  <rdfs:range rdf:resource=„E38.Conceptual_Object“/>  
  <rdfs:subPropertyOf rdf:resource=„P92F.brought_into_existence“/>  
</owl:ObjectProperty>
```

# Dokumentbeschreibung in RDF - Beispiel -

```

<?xml version=„1.0“?>
<rdf:RDF xml:lang=„en“
  xmlns:rdf=„http://www.w3.org/1999/02/22-rdf-syntax-ns#“
  xmlns:cidoc=„http://cidoc.ics.forth.gr/rdfs/cidoc_crm_v3.4.9.rdfs“
  <rdf:Description rdf:ID=„Winston Churchill“>
    <rdf:type rdf:resource=„#cidoc:E39.Actor“>
    <cidoc:P11B.participated_in>
      <cidoc:E7.Activity rdf:ID=„Crimea Conference“>
    </cidoc: P11B.participated_in>
    <cidoc:P14F.carried_out>
      <cidoc:E65.Creation_event rdf:ID=„*“>
    </cidoc: P14F.carried_out>
  </rdf:Description>

```

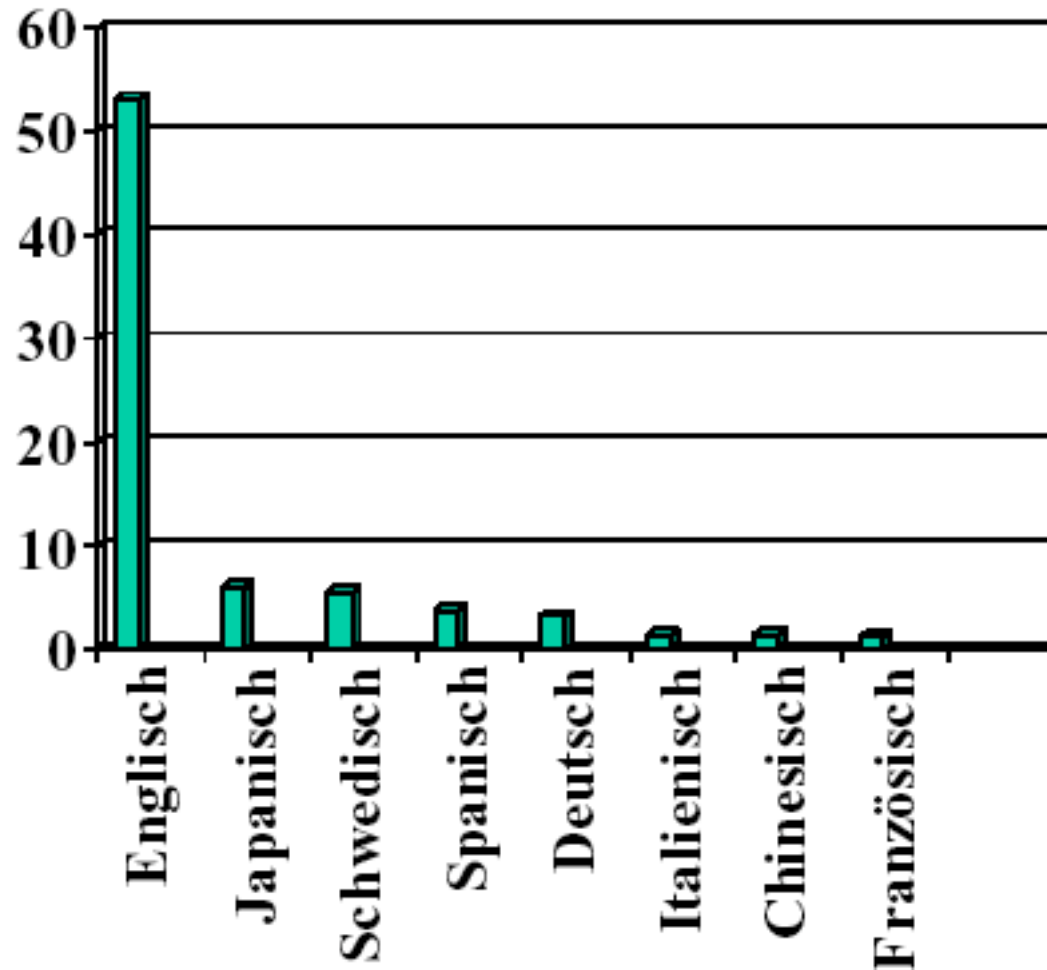


# Was kann man sonst mit einer Web-Ontologie machen?

- Übersetzung und Übersetzungsunterstützung  
Beispiel: DBR-MAT als wissensbasiertes Übersetzungs-Unterstützungssystem, oder wissensbasierte Terminologiedaten.
- Wissensmanagement  
Beispiel: Aufbau einer Ressource für gruppen-, projekt- oder firmenbezogenes Wissen, besonders in der mehrsprachigen Form für internationale Institutionen
- Internationale Kommunikationsbasis für Industrie und Handel
  - Beispiel: Produktlisten, Warennamen.



# Mehrsprachigkeit im Web



McLaughlin and  
Schwall

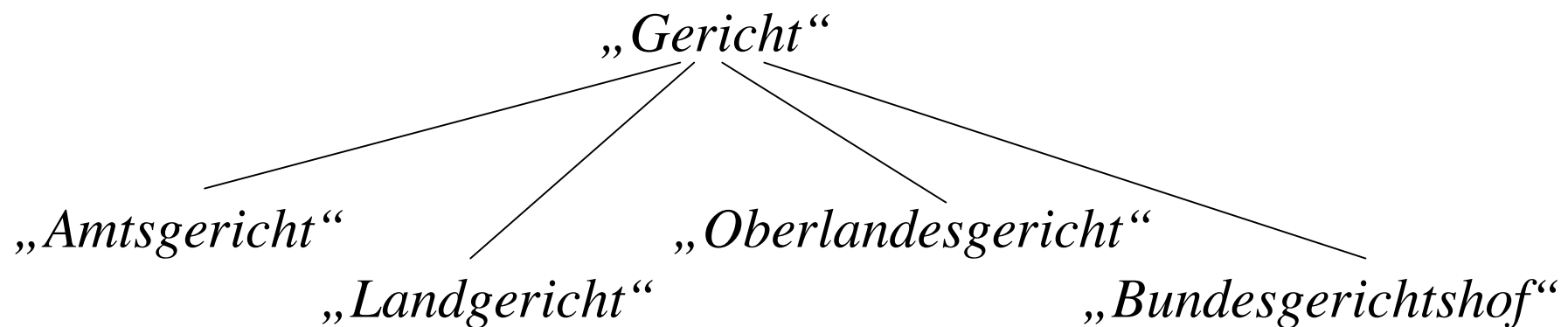
# Warum ein mehrsprachiges Semantic Web?

Wenn man ontologische Meta-Daten aufbaut, sollte man die Internationalität des WWW berücksichtigen. Bisher wird entweder

1. Auf eine englischsprachige Ontologie referenziert
  - *Vorteil:* Eindeutigkeit,
  - *Nachteil:* Jede nichtenglische Site muss die Abbildung aus der Einzelsprache liefern. Sie ist oft unterschiedlich, ambig oder unmöglich.
2. Je eine einzelsprachliche Ontologie aufgebaut
  - *Vorteil:* keine Abbildung zwischen Sprachen nötig,
  - *Nachteil:* Keine internationale Recherche oberhalb von lexikalischer Suche.

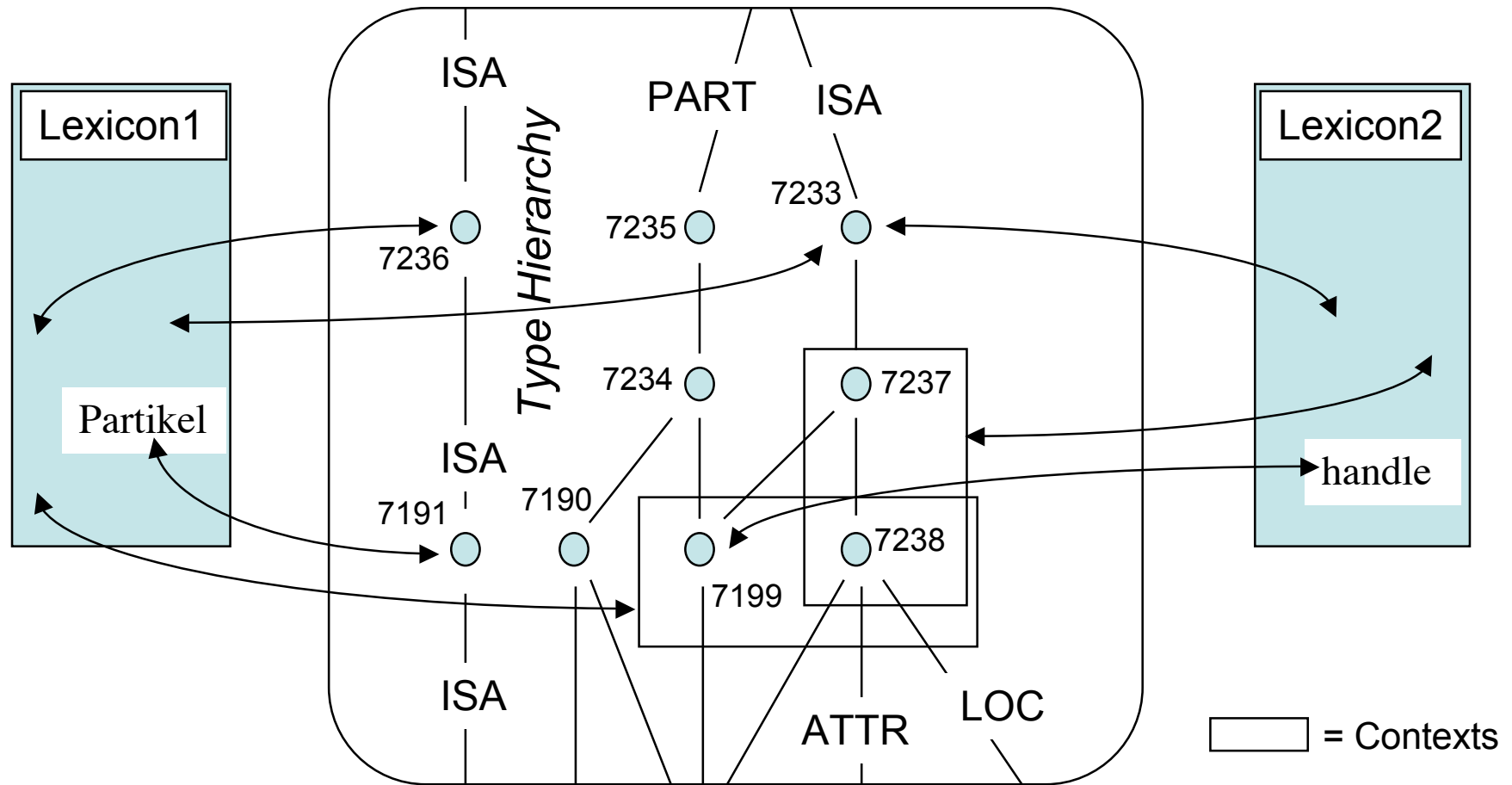
# Schwierigkeiten

- Übliche Gruber'sche Ontologien unterscheiden nicht zwischen Sachsystematik, Begriffshierarchie und Bezeichnern (Namen)



Hätte man eine Sach- oder Begriffshierarchie, könnte man durch zusätzliche einzelsprachliche Bezeichnungsrelationen die sprach- und kulturspezifischen Bezeichner eintragen.

# Eine mehrsprachige Ontologie (DBR-MAT)



# Linguistische Aufgaben

- Ontologien aufbauen und pflegen,
- Inferenzen definieren und pflegen,
- Mehrsprachige fach- und gemeinsprachige Äquivalenzen beschreiben, einschließlich
  - lexikalischer Lücken und
  - (partieller) Überschneidungen,
- Texte entsprechend den Ontologien annotieren,
- An Suchmaschinen mitarbeiten.