



---

## **The Semantic Web: Has the DB Community Missed the Bus (again ?)**

**Vipul Kashyap**

**National Library of Medicine, NIH**

**kashyap@nlm.nih.gov**

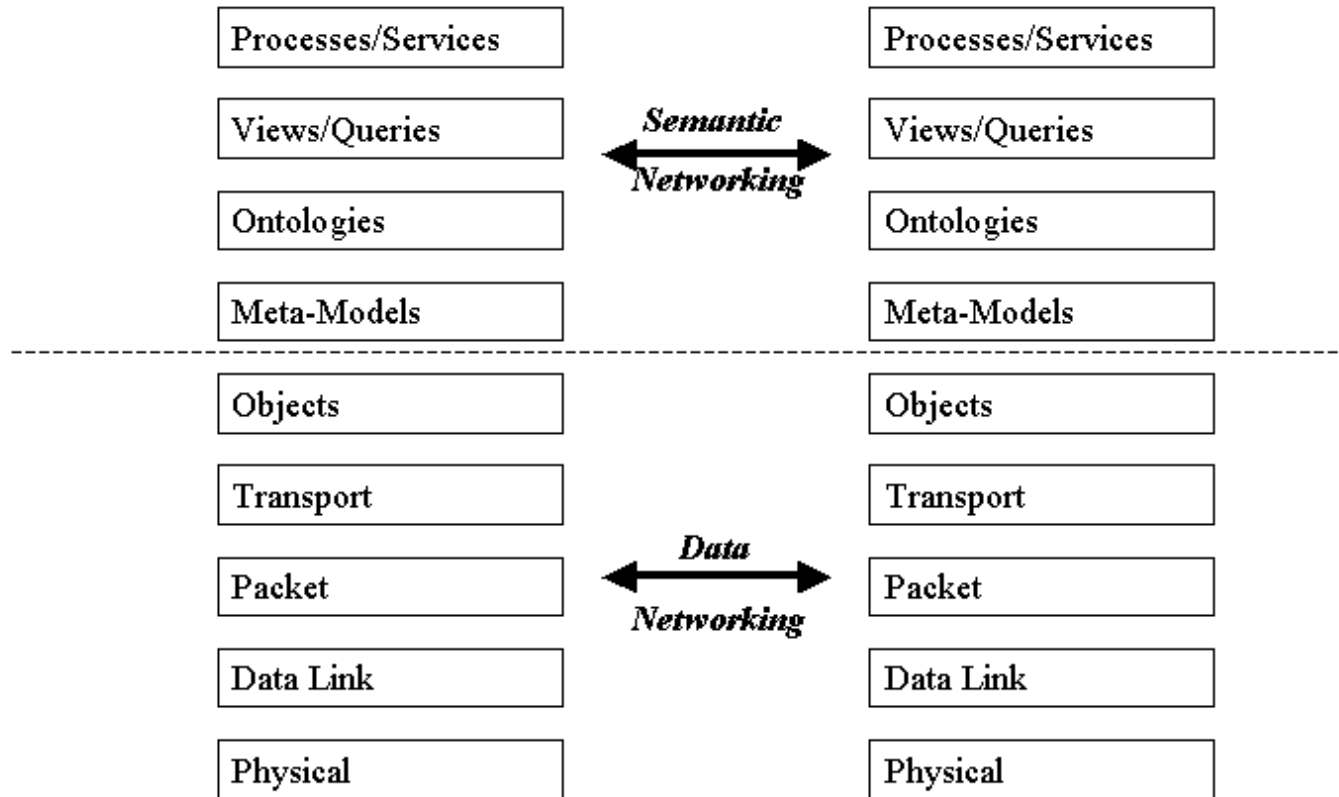
**NSF Workshop on DB & IS Research for Semantic Web and Enterprises**

**April 3, 2002**

# What Makes the “Syntactic” Web click ?

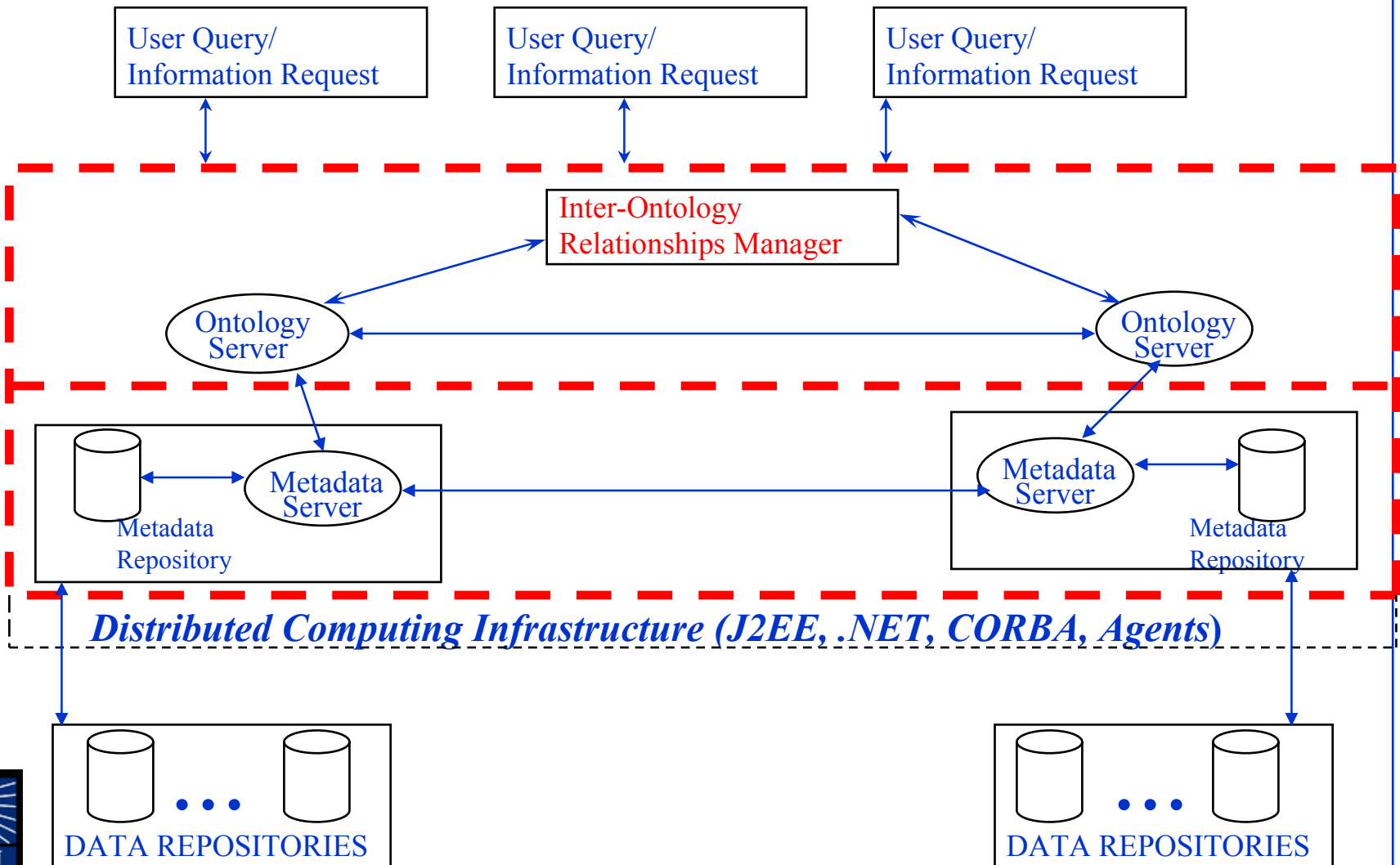
- Technology ?
  - Yes, but ...
  - Why wasn't the internet (telnet, ftp, gopher) as successful ?
  - Why were DBMS servers, CORBA/RMI not as successful ?
- Multimedia ?
  - Probably...
  - Better “cognitive compatibility” as compared to text...
- Ease of use ?
  - We are getting there ... !
  - Just point and click .... Easy to publish information ...
- People ?
  - “For the people/by the people”
  - A “primitive but useful” mechanism for people to “socialize” with each other !
- Questions:
  - What is the semantic web ? How can it make the syntactic web better ?
  - How can DB research help ?

# Semantic “Networking”



*It is crucial for the interoperability layer to migrate from the syntactic to the semantic!*

# The Semantic Web Fabric: A Collection of Metadata Descriptions and Ontologies



# Components of the Semantic Web Fabric

- Bootstrapping, Creation and Maintenance of Semantic Knowledge
  - Collaborative and Sociological Processes, Statistical Techniques
  - Ontology Building, Maintenance and Versioning Tools
- Re-use of Existing Semantic Knowledge (Ontologies)
- Annotation/Association/Extraction of Knowledge with/from Underlying Data
- Information Retrieval and Analysis (Distributed Querying/Search/Inference Middleware)
- Semantic Discovery and Composition of Services
- Distributed Computing/Communication Infrastructures
  - Component based technologies, Agent based systems, Web Services
- Repositories for managing data and semantic knowledge
  - Relational Databases, Content Management Systems, Knowledge Base Systems

# What DB researchers have done ?

- Semantic Data Models
- Multi-database Schema Heterogeneity
- Multi-database/Federated Database Schema Integration
- Schema Evolution
- Object Oriented/XML/Deductive Databases/Rule Based Systems
- Mediators and Wrappers
- Multidatabase/Federated Database Query Processing
- Data Mining
- Probabilistic Databases
- Workflow-based Coordination Systems
- Security in Database Systems
- Multimedia Databases
  - Text and Information Retrieval Systems
  - Image Databases

DB Research is well positioned to contribute to the Semantic Web, but:

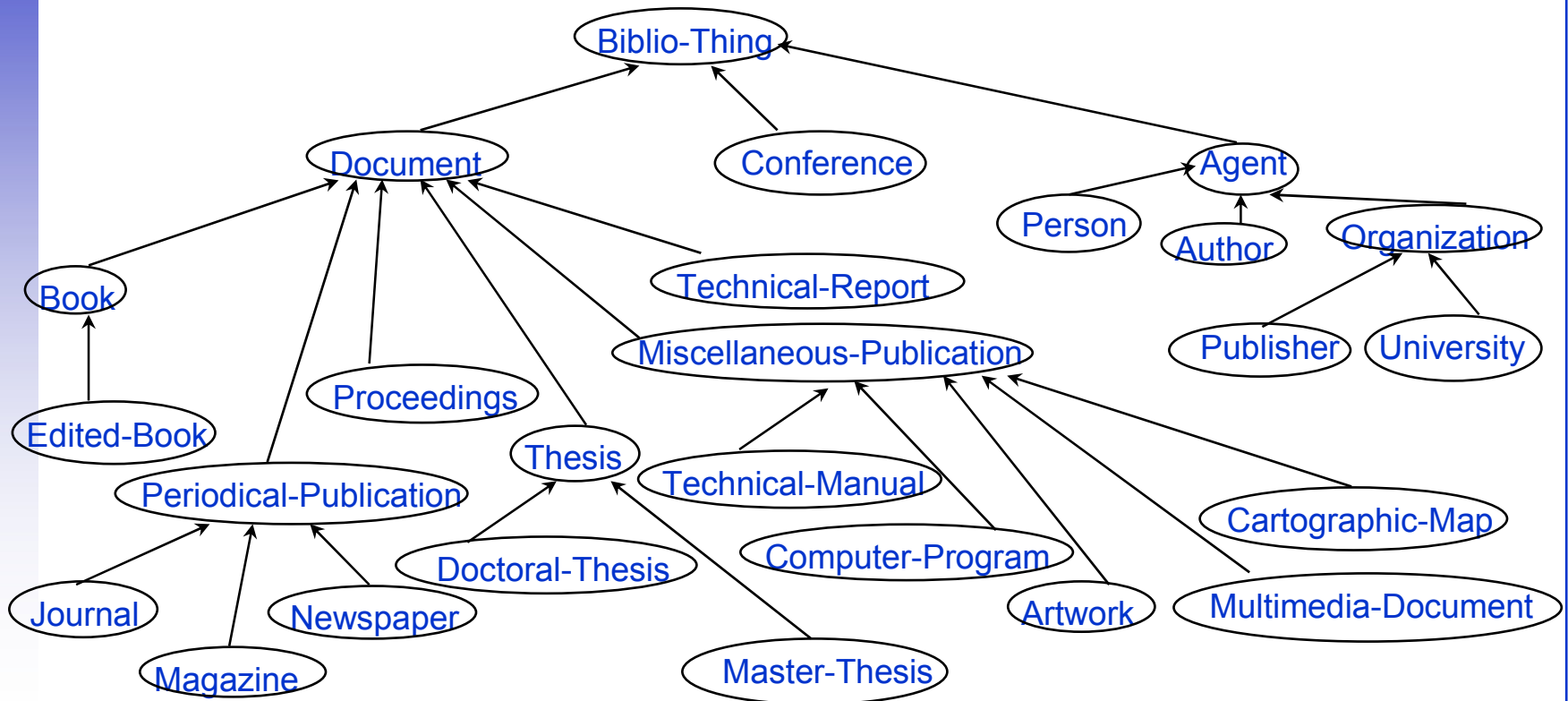
- there has been little interest in issues related to Semantics in the DB community
- the Semantic Web can be the underlying theme that ties in all the disparate pieces of work

# What are the missing gaps ?

- Ontology Integration/Interoperation
  - Problem is different from Schema Integration
  - Need to address “semantics” of relationships such as “synonyms”, “hyponyms”, etc.
- Ontology Impedance/Mismatch
  - Relax the requirements of consistency and completeness
  - Should be able to characterize the “information error/loss” that occurs..
- Dynamic Ontologies
  - Need to relax the assumption of the “staticness” of database schemas
- Inferences based on Semantics of the Data
  - Has been relatively ignored by the DB community
- Semantics of Multimedia Data
  - Need to focus more on non-traditional data such as text, images, etc.
  - Need to focus on “annotation mechanisms” as an addition to wrappers/mediators
- Semantics of Processes/Plans/Workflows
- Performance/Scalability
  - A traditional strong point of DB research

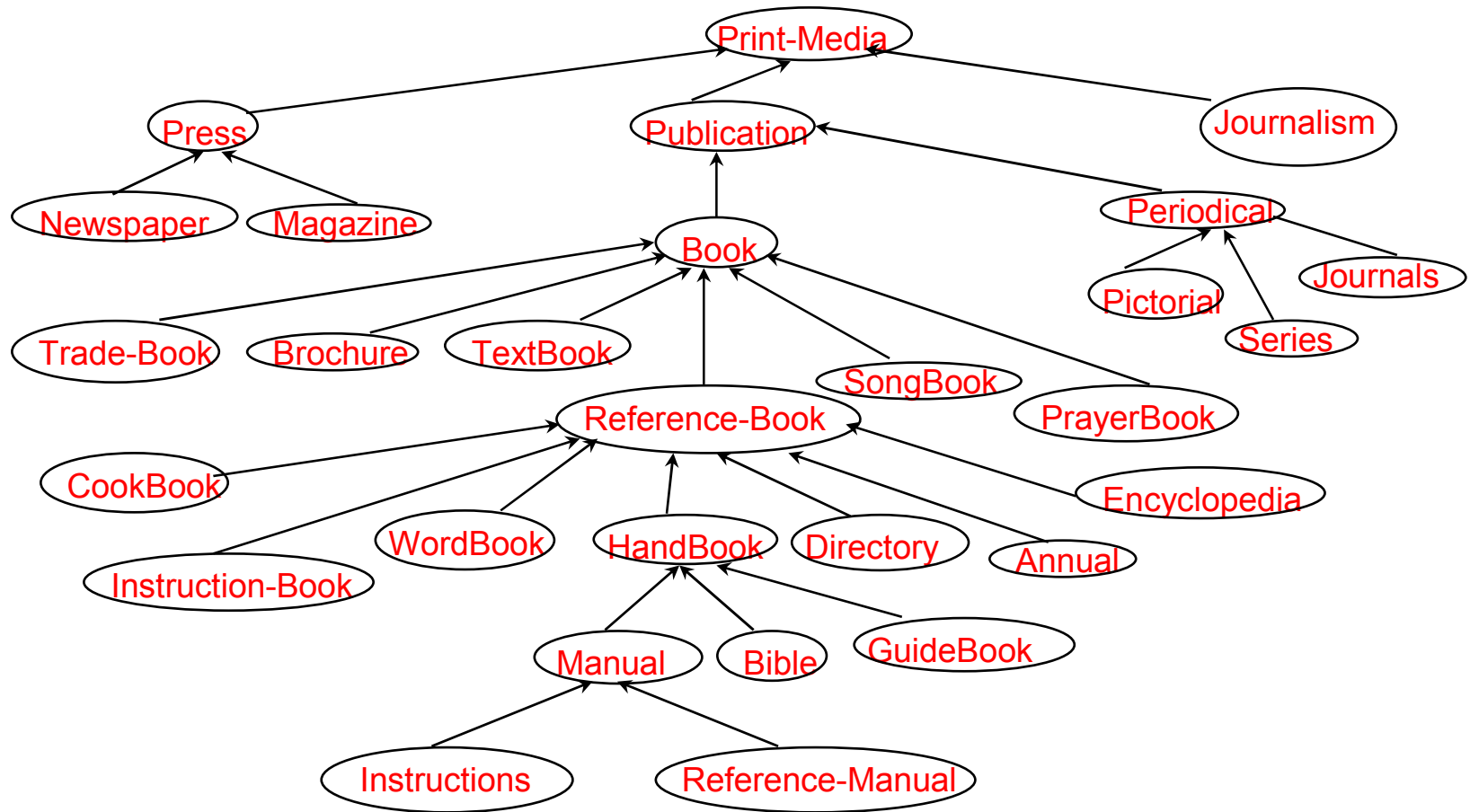
The next wave of research (esp. in the context of the Semantic Web) will focus on re-use of pre-existing data models/schemas/ontologies that describes the content of information sources...

# Bibliography Data Ontology: The **Blue** Ontology



<http://www-ksl.stanford.edu/knowledge-sharing/ontologies/html/bibliographic-data/>

# A subset of WordNet 1.5: The Red Ontology



<http://www.cogsci.princeton.edu/~wn/w3wn.html>

# Inter-ontological relationships

- Synonyms

- leads to semantics preserving translations

- Hyponyms/Hypernyms

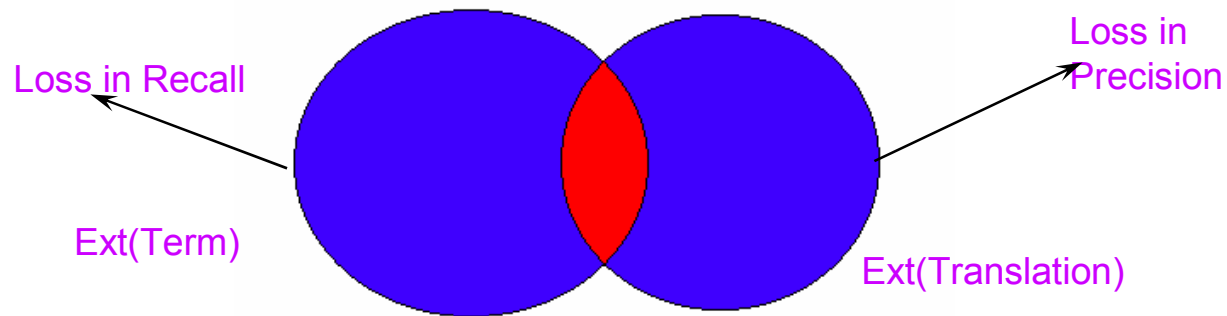
- lead to semantics altering translations ←
- typically results in loss of recall and precision

- List of Hyponyms

- |                     |                |                        |
|---------------------|----------------|------------------------|
| – technical-manual  | <i>hyponym</i> | manual                 |
| – book              | <i>hyponym</i> | book                   |
| – proceedings       | <i>hyponym</i> | book                   |
| – thesis            | <i>hyponym</i> | book                   |
| – misc-publication  | <i>hyponym</i> | book                   |
| – technical-reports | <i>hyponym</i> | book                   |
| – <b>press</b>      | <i>hyponym</i> | periodical-publication |
| – <b>periodical</b> | <i>hyponym</i> | periodical-publication |



# Estimating Loss of Information based on Term Extensions



$$\text{Precision} = \frac{|\text{Ext(Term)} \cap \text{Ext(Translation)}|}{|\text{Ext(Translation)}|}$$

$$\text{Recall} = \frac{|\text{Ext(Term)} \cap \text{Ext(Translation)}|}{|\text{Ext(Term)}|}$$

$$\text{Percentage Loss} = \frac{|\text{Ext(Term)} \Delta \text{Ext(Translation)}|}{|\text{Ext(Term)}| + |\text{Ext(Translation)}|}$$

$$= 1 - \frac{1}{1/2(1/\text{Precision}) + 1/2(1/\text{Recall})}$$

$$\Rightarrow 1 - \frac{1}{(\alpha)(1/\text{Precision}) + (1-\alpha)(1/\text{Recall})} \quad 0 < \alpha < 1$$

# Semantic Adaptation of Precision and Recall

- Term subsumes Translation
  - $\text{Ext}(\text{Translation}) \subseteq \text{Ext}(\text{Term}) \Rightarrow \text{Ext}(\text{Term}) \cap \text{Ext}(\text{Translation}) = \text{Ext}(\text{Translation})$
  - Precision = 1,
  - Recall =  $\frac{|\text{Ext}(\text{Translation})|}{|\text{Ext}(\text{Term})|}$
- However: Term and Translation belong to different ontologies
  - $\text{Ext}(\text{Term}) = \text{Ext}(\text{Term}) \cup \text{Ext}(\text{Translation})$
  - Recall.low =  $\frac{|\text{Ext}(\text{Translation})|.low}{|\text{Ext}(\text{Translation})|.low + |\text{Ext}(\text{Term})|}$
  - Recall.high =  $\frac{|\text{Ext}(\text{Translation})|.high}{\max(|\text{Ext}(\text{Translation})|.high, |\text{Ext}(\text{Term})|)}$
- Need to evolve a common framework for relating subsumption and information loss

# Conclusions

- Data Models/Schemas/Ontologies will form the critical infrastructure for the Semantic Web
- Re-use of pre-existing data models/schemas/ontologies is crucial in describing the semantics of various information sources
- There is a need to relax consistency and completeness requirements and estimate the “error” in the results returned.
- Semantics of information should be used to minimize “error” in the information obtained
- The new environment is likely to be more “dynamic” in nature – schemas, workflows, queries, etc. can no longer be assumed to be static...
- DB research is well positioned to participate in the Semantic Web if it “adapts” to these new requirements....

.... Otherwise it is in danger of missing the “bus” again !!