
Enabling the Semantic Web:

The role of metadata, semantics and domain ontologies

Vipul Kashyap

National Library of Medicine


kashyap@nlm.nih.gov

<http://cgsb2.nlm.nih.gov/~kashyap>

Colloquium Talk, CSEE Department, UMBC

October 3, 2003

Outline

- What is the Semantic Web ? 
- Metadata and Ontologies
 - A Three Level Approach for the Semantic Web
- The Semantic Web Fabric: A Collection of Metadata and Ontologies
 - Components of the Semantic Web Fabric
 - Metadata-based approach for Heterogeneous Digital Data
- Ontologies: A critical Semantic Web “bottleneck”
 - Bootstrapping
 - Enhancement of Existing Resources
 - Re-use: Multiple Ontology-based Query Processing
- Conclusions and Future Work

What is the Semantic Web?

■ Semantics:

- “meaning or relationship of meanings, or relating to meaning ...” (Webster),
- meaning and use of data (Information System perspective)


■ Semantic Web:

- An extension of the current web, in which **information is given well-defined meaning**, better enabling computers and people to work in cooperation [Berners-Lee, Hendler, Lassila, 2001]

■ “Emergent” Semantics:

- **Creation, validation** and use of **dynamic knowledge**, where semantics “**emerges**” from the interactions between people and applications on the web.

Outline

- What is the Semantic Web ?
- Metadata and Ontologies 
 - A Three Level Approach for the Semantic Web
- The Semantic Web Fabric: A Collection of Metadata and Ontologies
 - Components of the Semantic Web Fabric
 - Metadata-based approach for Heterogeneous Digital Data
- Ontologies: A critical Semantic Web “bottleneck”
 - Bootstrapping
 - Enhancement of Existing Resources
 - Re-use: Multiple Ontology-based Query Processing
- Conclusions and Future Work

Metadata and Ontologies

Get the **titles, authors, documents, maps** published by the United States Geological Service (USGS) about **regions** having a **population** greater than 5000, **area** greater than 1000 acres having a low density urban area **land cover**

Domain specific metadata: **terms** chosen from **domain specific ontologies**

What is Metadata ?

- data/information about data
- useful/derived properties of media
- properties/relationships between objects
- may or may not capture information content of underlying data

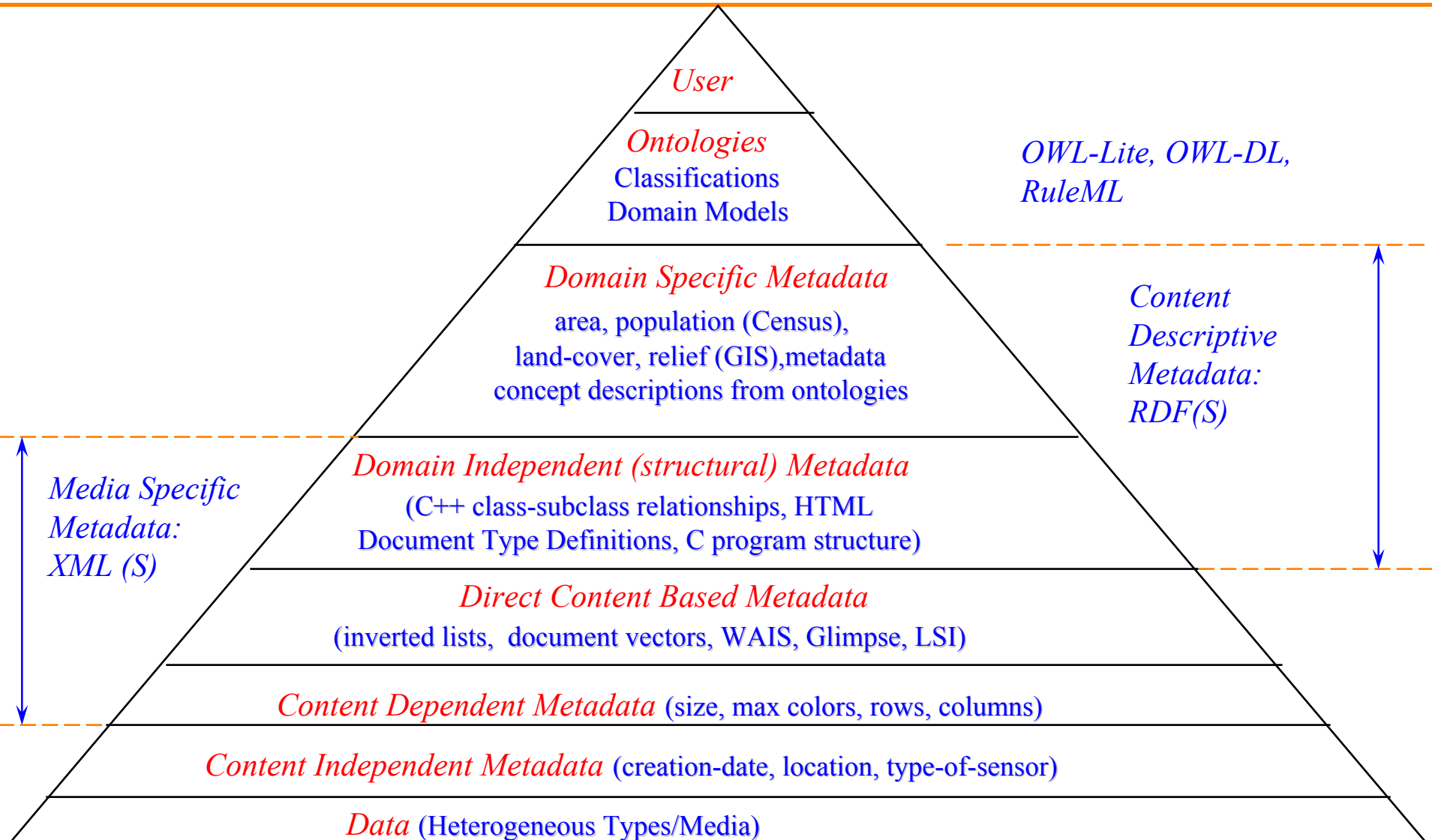
What are Ontologies ?

- collection of terms, definitions and interrelationships
- specification of a representational vocabulary for a shared domain of discourse
- Semantically rich metadata capturing the information content of underlying data repositories
- Lattice of OWL-DL expressions

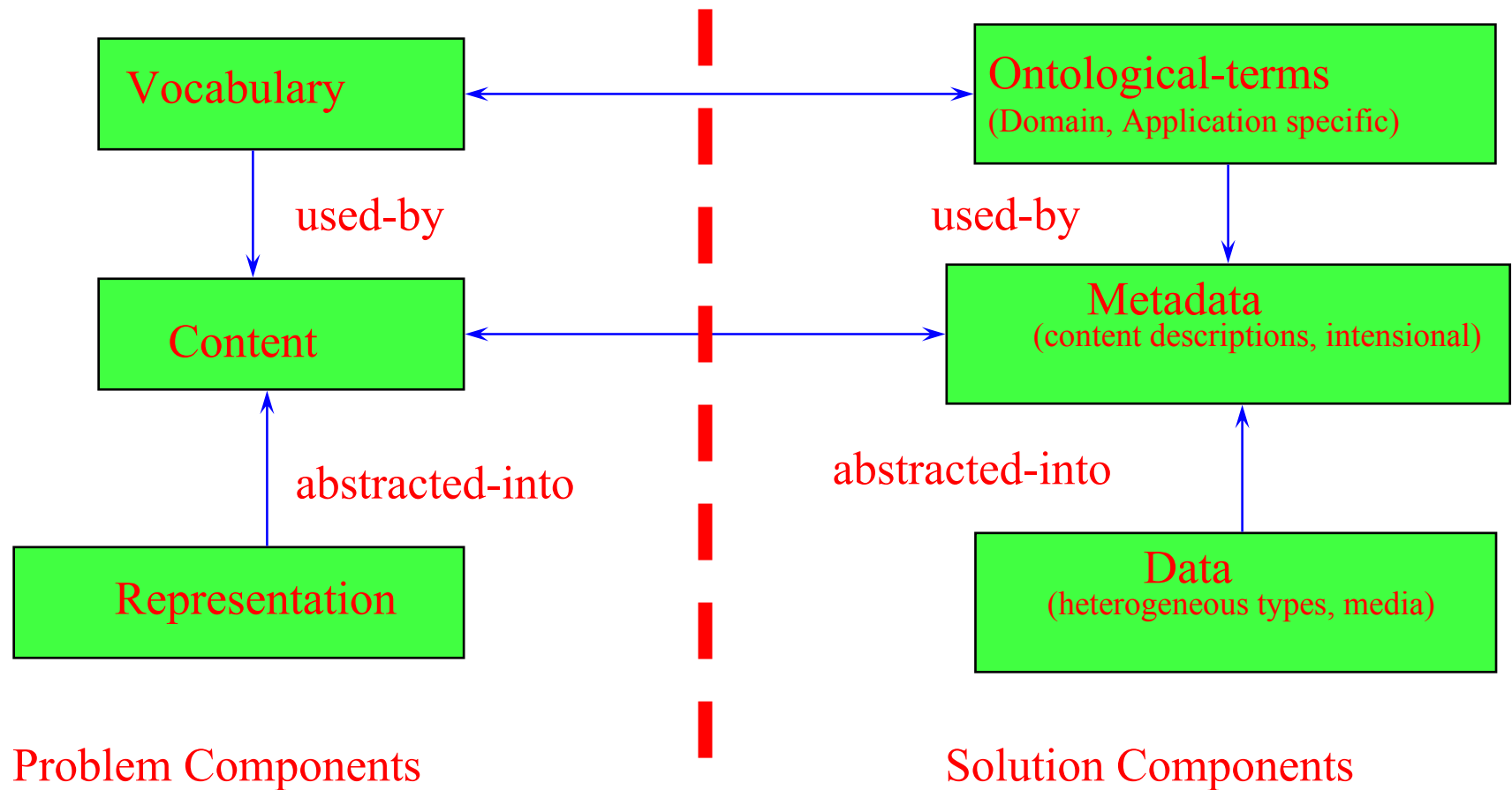
Metadata for Digital Data: Examples

<i>Metadata</i>	<i>Data Type</i>	<i>Metadata Type</i>
Q-Features [Jain and Hampapur]	Image, Video	Domain Specific
R-Features [Jain and Hampapur]	Image, Video	Domain Independent
Meta-Features [Jain and Hampapur]	Image, Video	Content Independent
Impression Vector [Kiyoki et al.]	Image	Content Descriptive
NDVI, Spatial Registration [Anderson and Stonebraker]	Image	Domain Specific
Speech Feature Index [Glavitsch et al.]	Audio	Direct Content Based
Topic Change Indices [Chen et al.]	Audio	Direct Content Based
Document Vectors [Deerwester et al.]	Text	Direct Content Based
Inverted Indices [Kahle and Medlar]	Text	Direct Content Based
Content Classification Metadata [Bohm and Rakow]	MultiMedia	Domain Specific
Document Composition Metadata [Bohm and Rakow]	MultiMedia	Domain Independent
Metadata Templates [Ordille and Miller]	Media Independent	Domain Specific
Land Cover, Relief [Sheth and Kashyap]	Media Independent	Domain Specific
Parent Child Relationships [Shklar et al.]	Text	Domain Independent
Contexts [Sciore et al., Kashyap and Sheth]	Structured	Domain Specific
Concepts from Cyc [Collet et al.]	Structured	Domain Specific
User's Data Attributes [Shoens et al.]	Text, Structured	Domain Specific
Domain Specific Ontologies [Mena et al.]	Media Independent	Domain Specific


A Metadata Classification: The Information Pyramid



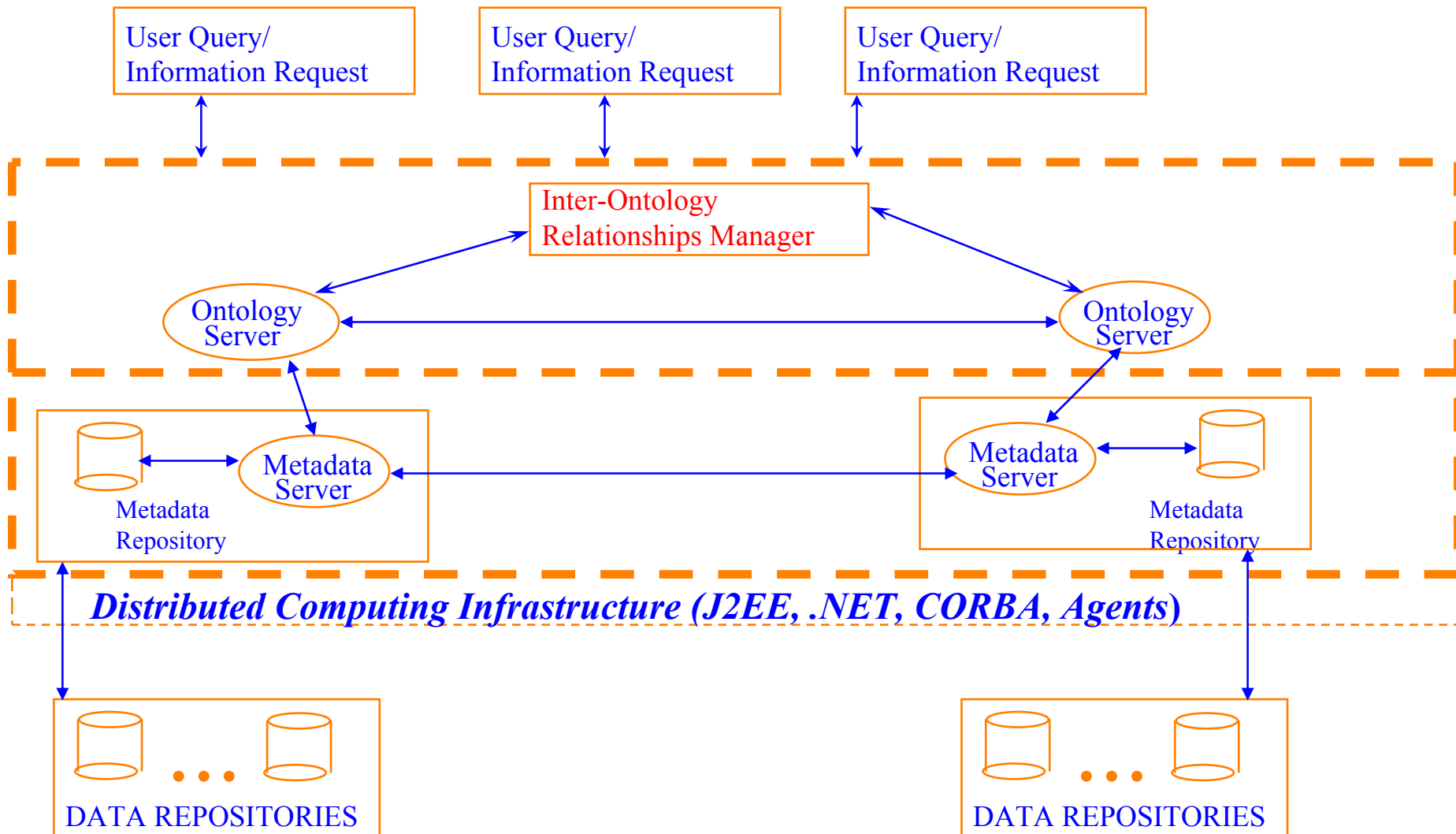
The Semantic Web: A Three Layer Approach



Outline

- What is the Semantic Web ?
- Metadata and Ontologies
 - A Three Level Approach for the Semantic Web
- The Semantic Web Fabric: A Collection of Metadata and Ontologies
 - Components of the Semantic Web Fabric 
 - Metadata-based approach for Heterogeneous Digital Data
- Ontologies: A critical Semantic Web “bottleneck”
 - Bootstrapping
 - Enhancement of Existing Resources
 - Re-use: Multiple Ontology-based Query Processing
- Conclusions and Future Work

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

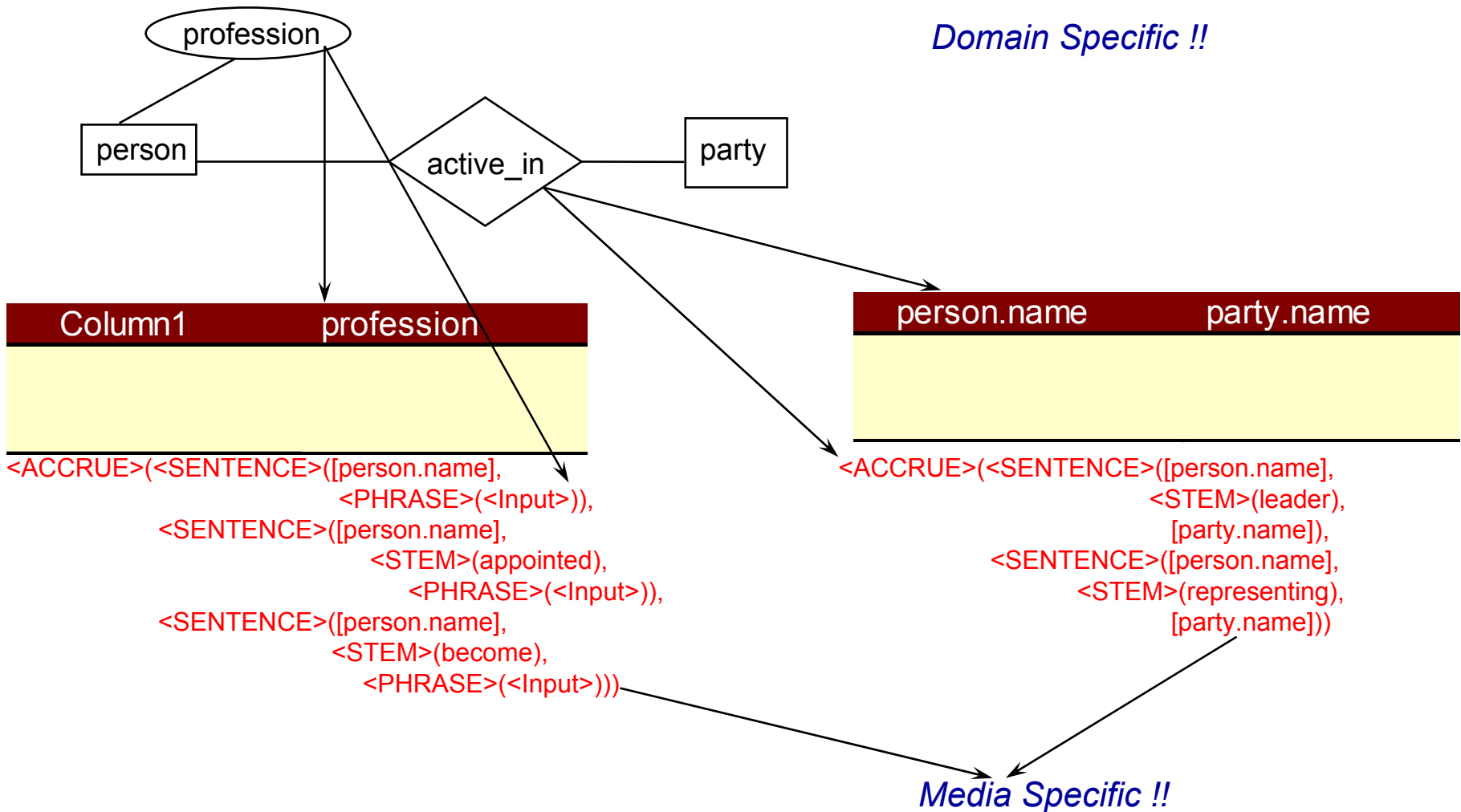
*Significant
Human Involvement*

Associating Knowledge with Data:

From “media specific” to “domain specific” metadata

- Annotation/Association/Extraction of Knowledge with/from Underlying Data
 - Structured Databases
 - 📄 Mapping concepts in domain ontologies to schema metadata elements
 - Text Databases
 - 📄 Mapping of concepts in domain ontologies to text patterns, e.g., sentence, phrase, etc.
 - Image Databases
 - 📄 Mapping of concepts in domain ontologies to image patterns, e.g., color, texture, shape, etc.
- Information Retrieval and Analysis
 - Structured Databases
 - 📄 Distributed Query Processing across Multiple Information Sources
 - Text Databases
 - 📄 Mapping SQL/Description Logic based queries into text retrieval expressions
 - Image Databases
 - 📄 Mapping “Ontological Exemplars” into image processing routines

Metadata-based approach: Mapping ontological elements to textual data



Metadata-based approach:

Mapping OWL-DL expressions to Topic Expressions

[has_document] from (AND person (FILLS name "Alexandr Shokhin")
(FILLS profession 'Prime Minister'))

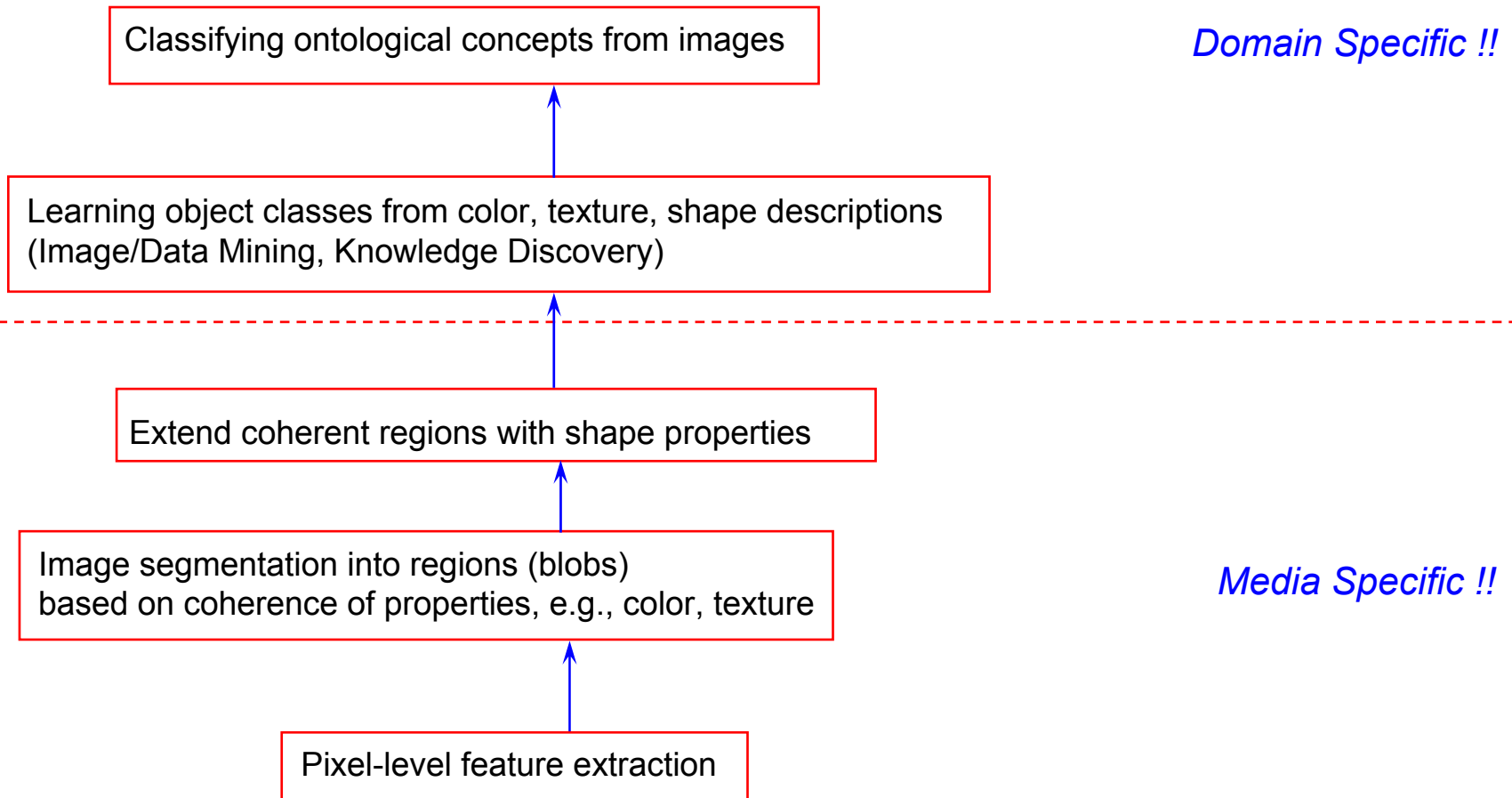
<ACCRUE>(<TOPIC>(person),

<PHRASE>(<WORD>(Aleksandr), <WORD>(Shokhin)),

<ACCRUE>(<SENTENCE>(<PHRASE>(<WORD>(Aleksandr),
<WORD>(Shokhin)),
<STEM>(appointed),
<PHRASE>(<WORD>(Prime), <WORD>(Minister))),
<SENTENCE>(<PHRASE>(<WORD>(Aleksandr),
<WORD>(Shokhin)),
<STEM>(becomes),
<PHRASE>(<WORD>(Prime), <WORD>(Minister))))))

Metadata-based approach:

Selecting and using appropriate metadata for image retrieval



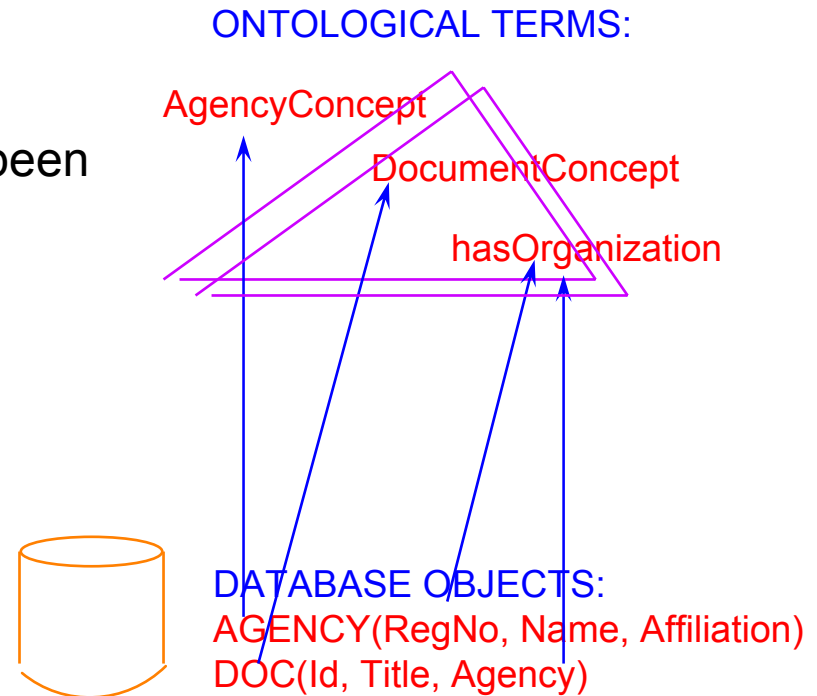
Note: Future Work, Current Status “Thoughtware”

Metadata-based approach:

Describing database objects using OWL/DL expressions

“All documents stored in the database have been published by some agency”

Database Documents
≡ (AND DocumentConcept
(hasOrganization AgencyConcept))

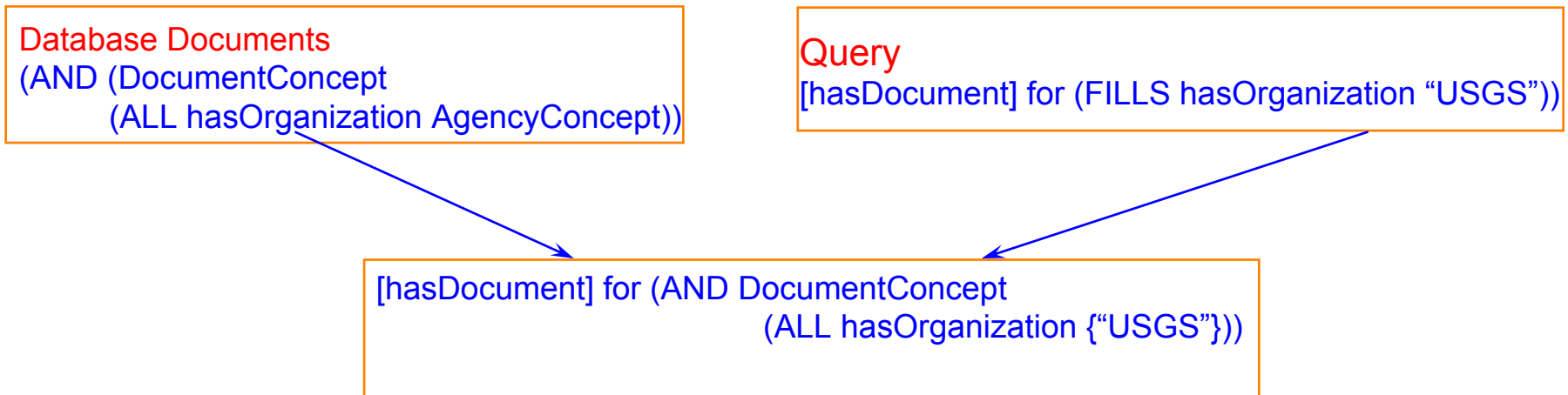


Advantages:

- Use of **ontologies** for an **intensional** domain specific description of data
- Representation of extra information
 - Relationships between objects not represented in the database schema
 - Using terminological relationships in the ontology


Metadata-based approach:

Using OWL/DL expressions to reason about underlying data



- Reasoning with OWL-DL Expressions
- Ontological Inferences:
 - DocumentConcept
 - (hasOrganization, { "USGS" })
- Types of Reasoning:
 - Subsumption
 - Most specific subsumer/Most general subsumee

Outline

- What is the Semantic Web ?
- Metadata and Ontologies
 - A Three Level Approach for the Semantic Web
- The Semantic Web Fabric: A Collection of Metadata and Ontologies
 - Components of the Semantic Web Fabric
 - Metadata-based approach for Heterogeneous Digital Data
- Ontologies: A critical Semantic Web “bottleneck” 
 - Bootstrapping
 - Enhancement of Existing Resources
 - Re-use: Multiple Ontology-based Query Processing
- Conclusions and Future Work

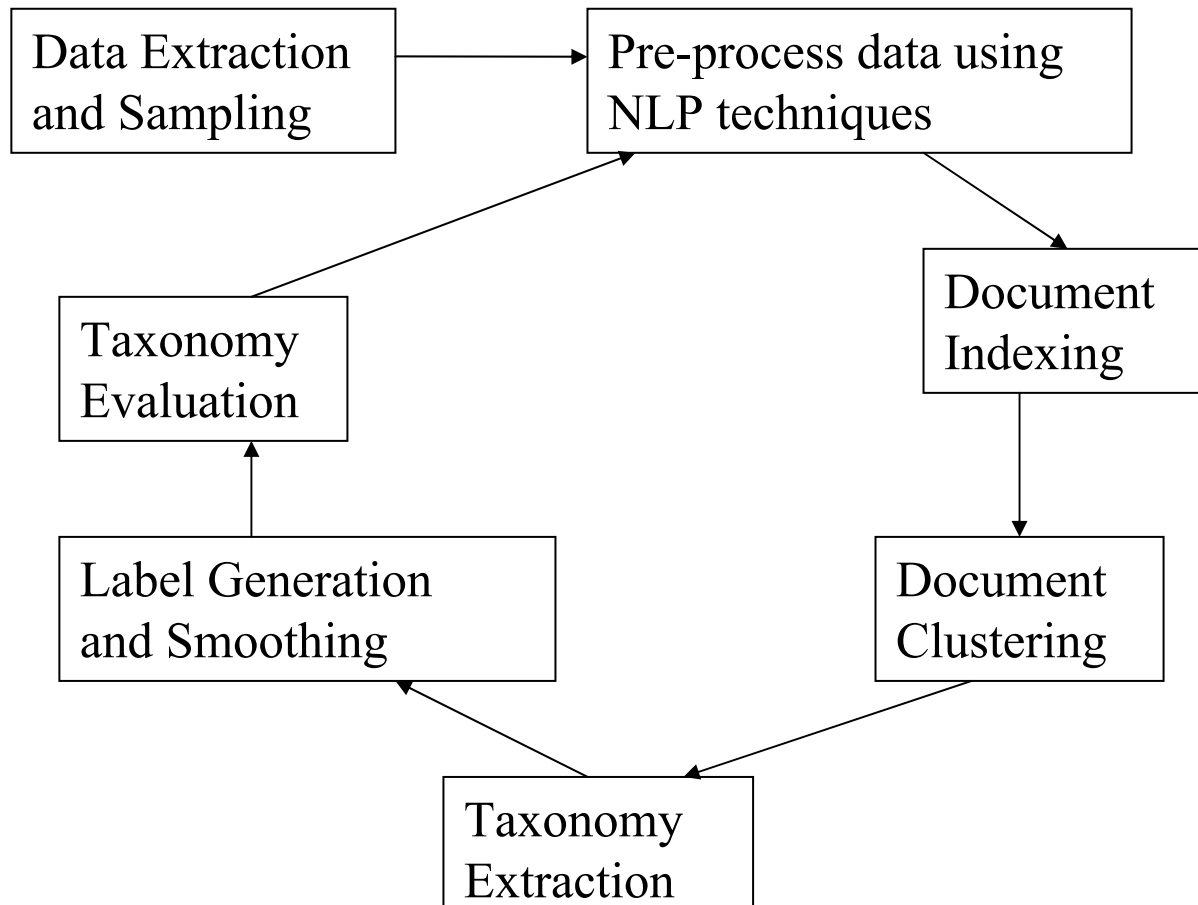
Ontologies:

A critical Semantic Web “bottleneck”

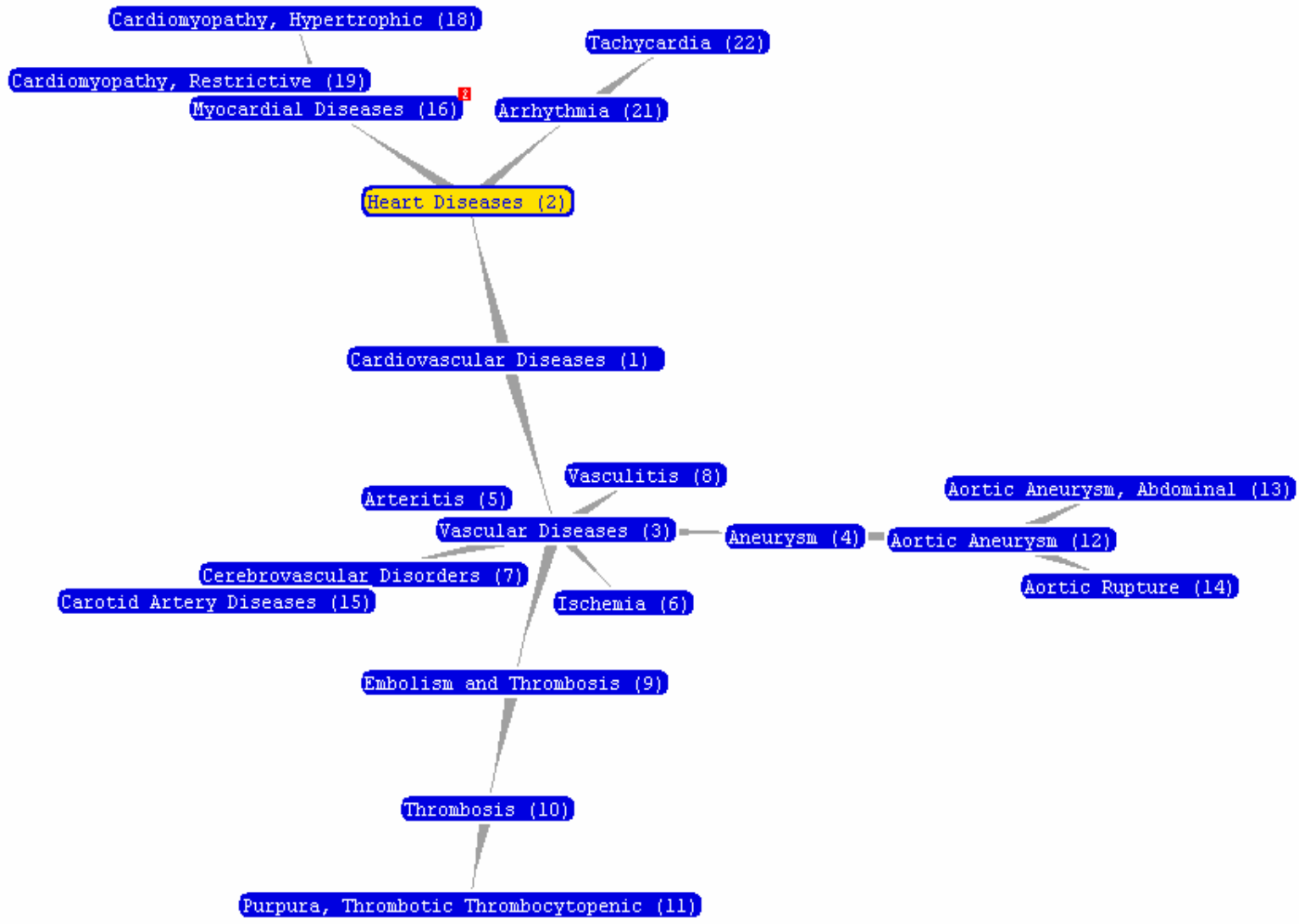
- Where do we get the ontologies from? How do we minimize human effort in creating them?
 - Bootstrapping approaches
- Can we re-use existing resources to create new ontologies?
 - E.g., database schemas, thesauri
- Can we re-use pre-existing independently developed ontologies?
 - Multi-Ontology Query Processing

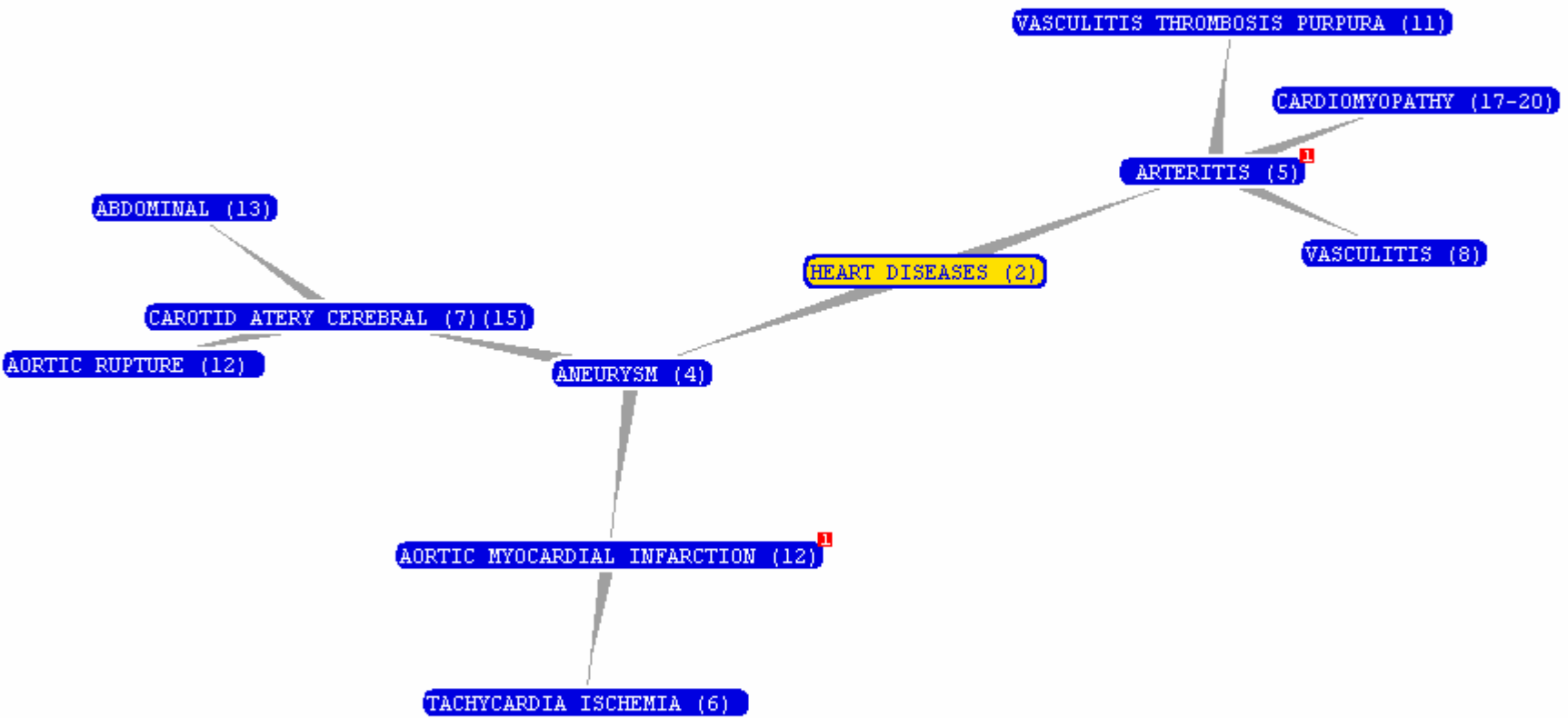
Bootstrapping:

An approach involving Statistical and NLP techniques



Component of “Emergent” Semantics
Ongoing work – Initial Promising results





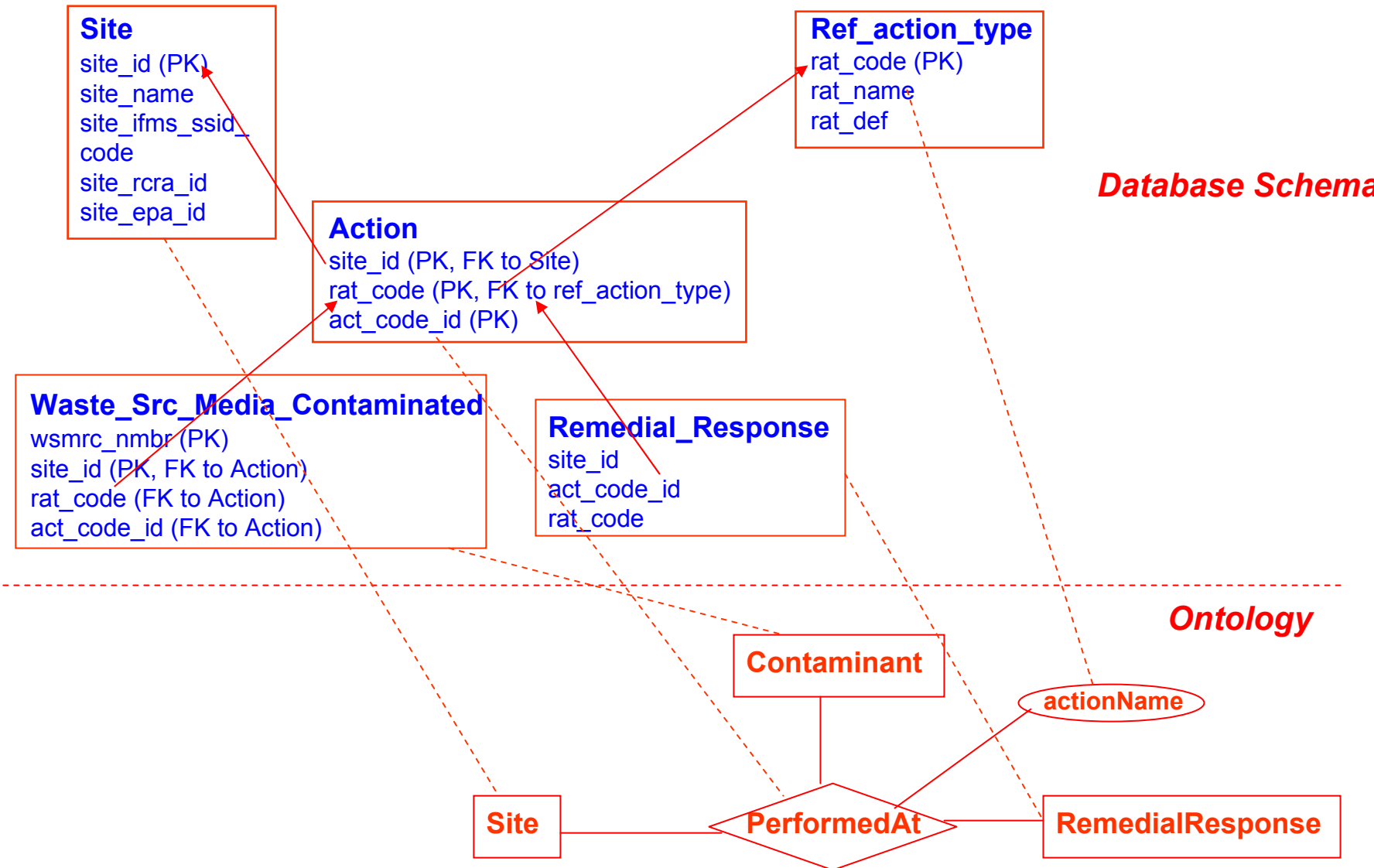
Enhancing Existing Resources: Thesauri

- Thesauri:
 - Characterized by broader-than/narrower than hierarchical relationships
 - Provide an excellent source of knowledge for creating ontologies
- Analysis of major syntactic strategies for encoding hypernymy
 - Verbs (about 20%)
 - ☞ Nimodipine is an isopropyl calcium channel blocker
 - Appostives (about 40%)
 - ☞ Arginine, a semi-essential amino acid, has been shown to increase...
 - Nominal modification
 - ☞ The anticonvulsant gabapentin has proven effective for neuropathic pain
 - Lexico syntactic patterns identified by Marti Hearst
 - Check for hierarchical relationships in a thesauri

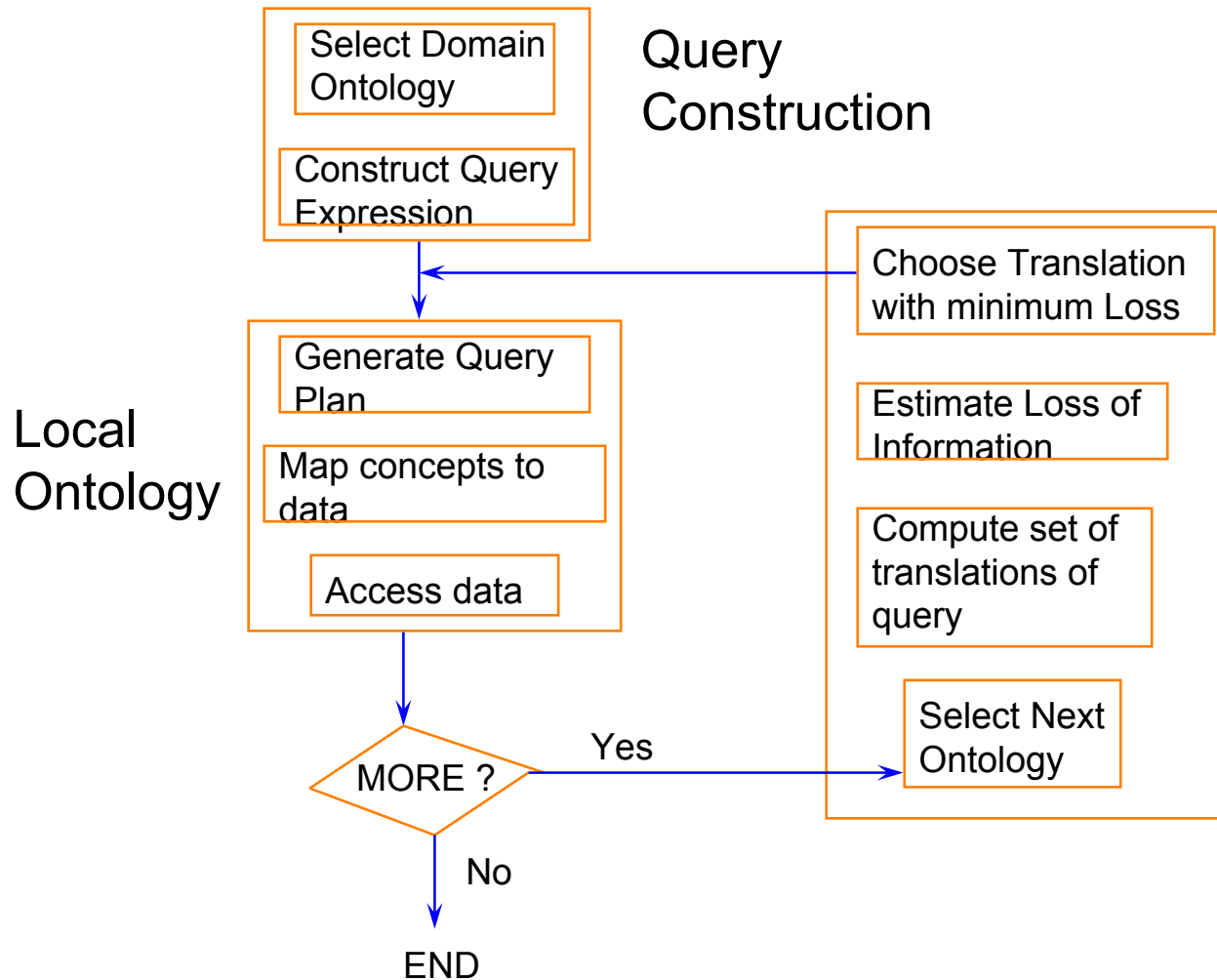
Part of Semantic Knowledge Representation Project at the NLM
Re-use and adapt these techniques for Automatic Taxonomy Generation

Enhancing Existing Resources: DB Schemas

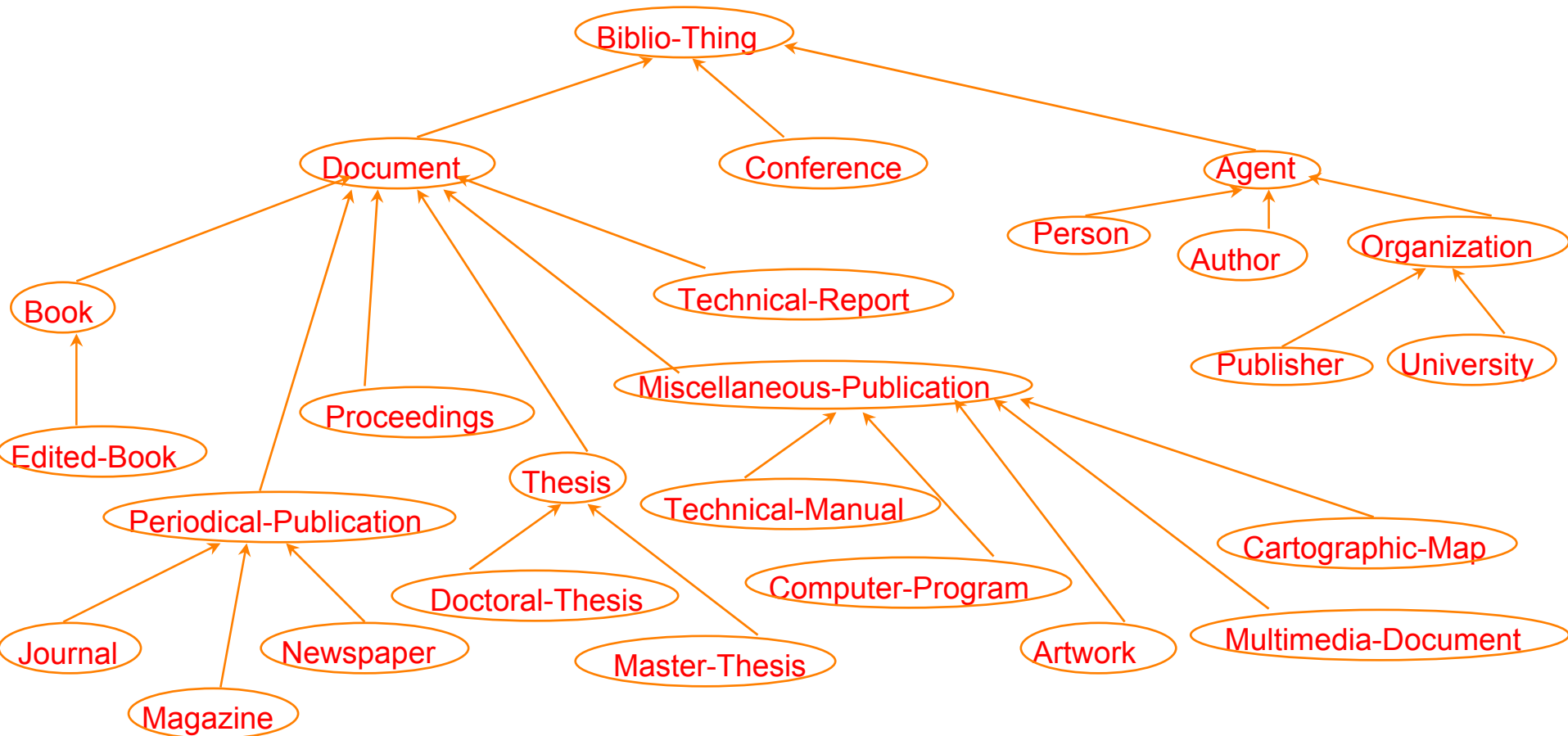
EDEN Project at MCC



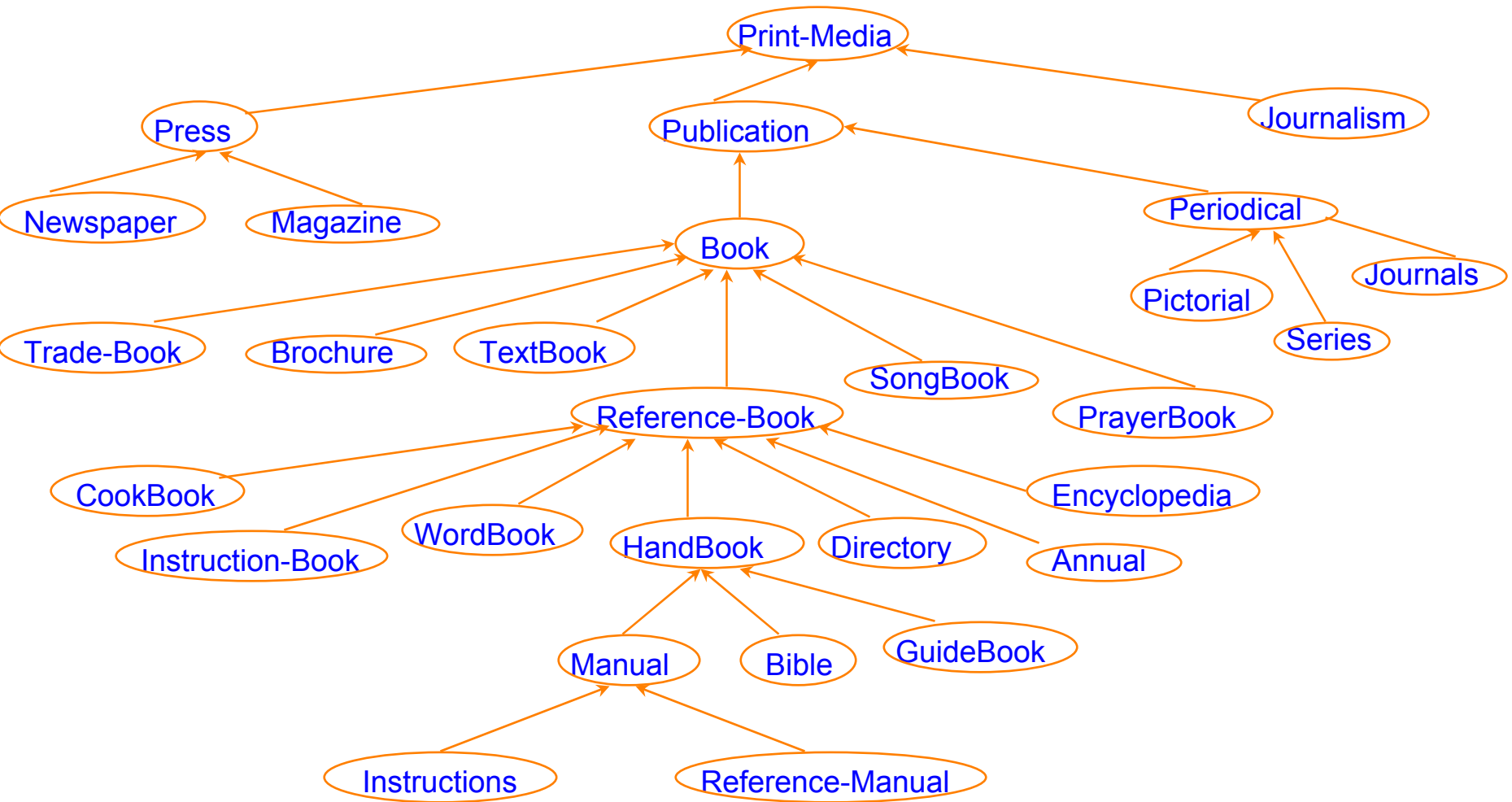
Re-use: Multi-Ontology Query Processing



The Bibliography Data (Red) Ontology



The WordNet (subset, Blue) Ontology



Inter-Ontological Relationships

■ Synonyms

- leads to semantics preserving translations

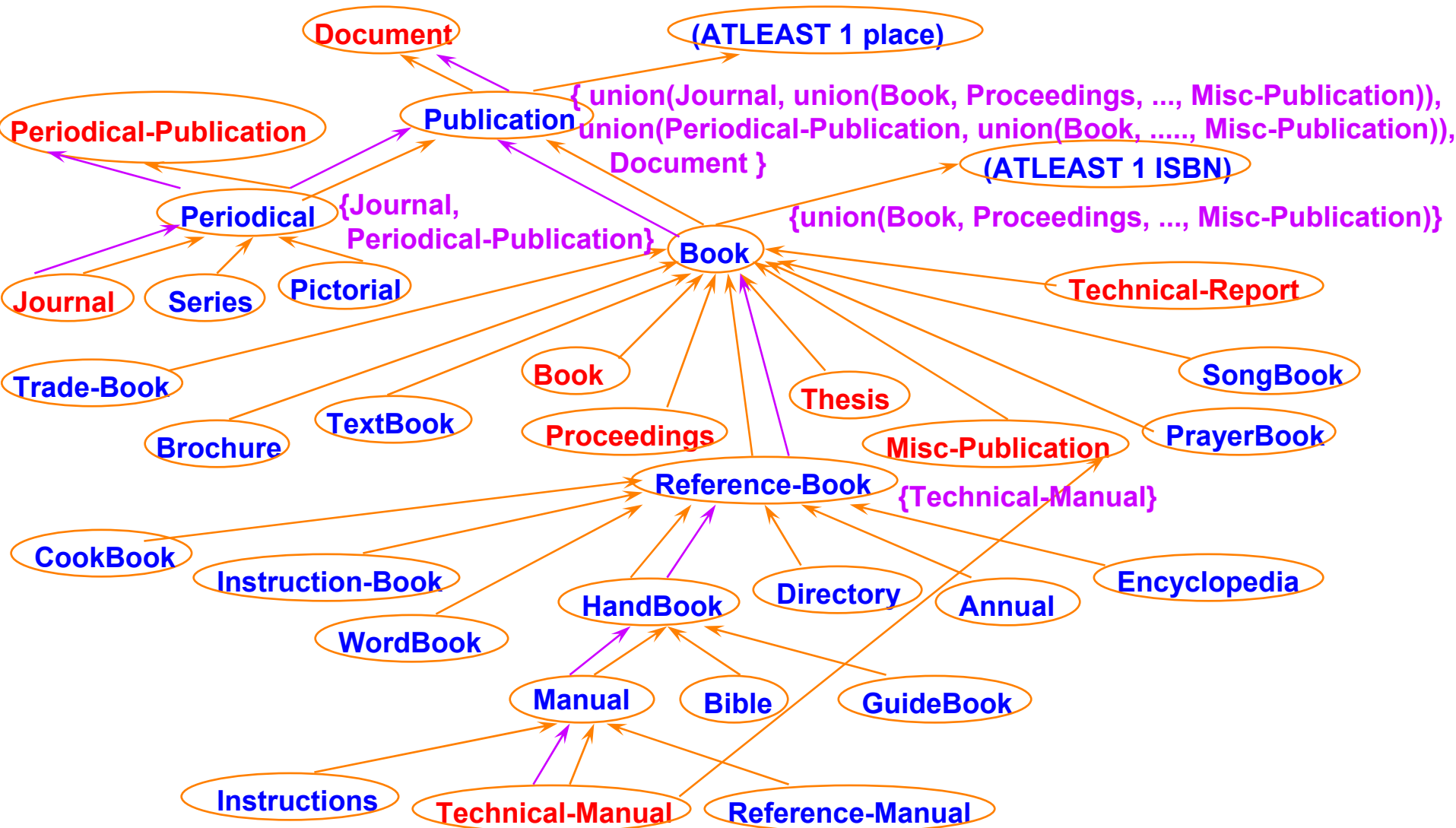
■ Hyponyms/Hypernyms ←

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

■ List of Hyponyms

- technical-manual *hyponym* manual
- book *hyponym* book
- proceedings *hyponym* book
- thesis *hyponym* book
- misc-publication *hyponym* book
- technical-reports *hyponym* book
- press *hyponym* periodical-publication
- periodical *hyponym* periodical-publication

Ontology Integration and Query Re-writing



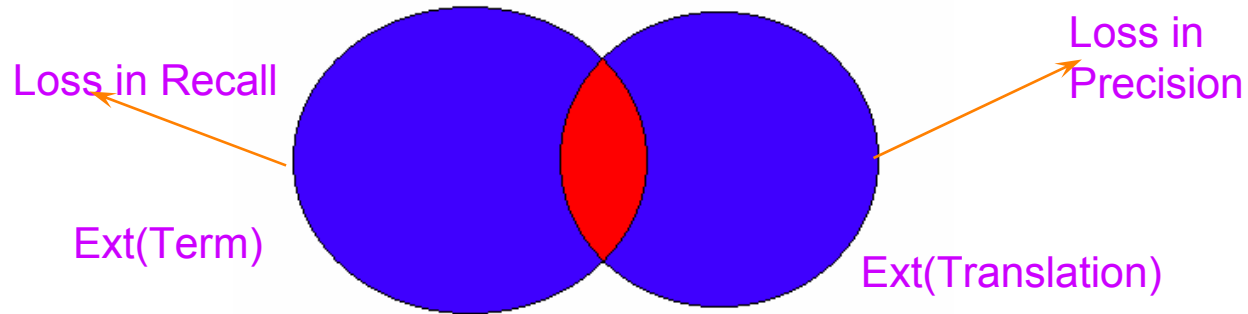
Loss of Information (Intensional)

- Original Query:
 - [NAME PAGES] for (AND BOOK (FILLS CREATOR “Carl Sagan”))
- Modified Query:
 - [NAME PAGES] for (AND document (FILLS doc-author-name “Carl Sagan”))
- Terminological Relationships:
 - BOOK \Leftrightarrow (AND PUBLICATION (ATLEAST 1 ISBN))
 - PUBLICATION \Leftrightarrow (AND document (ATLEAST 1 PLACE-OF-PUBLICATION))
- Terminological Difference:
 - (AND (ATLEAST 1 ISBN) (ATLEAST 1 PLACE-OF-PUBLICATION))
- Loss of Information:
 - Instead of books authored by Carl Sagan, OBSERVER returns those documents by Carl Sagan that may not have an ISBN or may not have been published

Intensional Loss of Information: Advantages and Disadvantages

- May not make sense as it mixes two vocabularies,
 - e.g., does **Book** - **Book** make any sense ?
- The problem becomes worse if the two ontologies are in different languages,
 - e.g., English and Italian
- Makes it hard for the system to differentiate between the various alternatives
- **On the other hand:**
 - An information loss interval doesn't make much sense to the user.

Loss of Information (Extensional)



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

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

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

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

Loss of Information: Semantic Adaptation

- 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 = $\frac{|\text{Ext}(\text{Translation})|}{|\text{Ext}(\text{Translation})| + |\text{Ext}(\text{Term})|}$

- Need to evolve a common framework for relating subsumption and information loss

Loss of Information: Semantic Adaptation

- Translation subsumes Term
 - Dual of the previous case
 - Recall = 1
 - Precision = $\frac{|\text{Ext}(\text{Term})|}{|\text{Ext}(\text{Translation})|}$
- Cases of no Information Loss
 - Translation of a term by the intersection of its immediate parents which is also its definition
 - Translation of a term by the union of its immediate children if there exists a “covering” relationship between the two
- Need for “extensional” inter-ontological relationships
 - e.g., 20% of publications are 50% of books
 - characterizing degree of overlap

Challenges: Biomedical Informatics

- Scale:
 - Huge number of concepts: in the 1000s
 - May only want to merge relevant portions of the vocabularies
- Semantic Poverty
 - UMLS lacks “semantics”
 - ☞ BT/NT
 - ☞ Parent/Child
 - Need to convert hierarchical relationships to “is-a” or “part-of”
 - How does one compute “Information Loss” ?
- Inconsistency
 - Circular relationships in the UMLS Metathesaurus
 - ☞ A ParentOf B ParentOf C ParentOf A
 - ☞ How does one break these cycles?

Conclusions

- Analysis of the Semantic Web Technology Space
 - Proposed a Three Layered Approach
 - Identified components of the Semantic Web Fabric
- Building out the Semantic Web Infrastructure
 - Semantic Knowledge needs to be associated with heterogeneous digital data
 - ☞ E.g., structured, text and image data
 - Metadata plays a crucial role in the above endeavor
 - Ontologies are both a crucial component and a critical bottleneck for the Semantic Web
- Ontologies: A critical bottleneck for the Semantic Web
 - Bootstrapping approaches to create “seed” ontologies
 - Enrichment of existing resources: e.g., DB Schemas, Thesauri
 - Techniques for re-use of pre-existing ontologies (“off the shelf”)
 - Issues related to loss of information and semantic distance

Ongoing and Future Work

- Automatic Taxonomy Extraction
 - TaxaMiner Project
 - <http://cgsb2.nlm.nih.gov/~kashyap/projects/TaxaMiner>
- Challenges from Biomedical Informatics
 - Semantic Vocabulary Interoperation Project
 - <http://cgsb2.nlm.nih.gov/~kashyap/projects/SVIP>
- Semantics, Loss of Information and Semantic Distance
 - Experimentation and Validation
 - Common Framework to deal with subsumption, meronymy and Loss of Information
- Web Services and Bio-Informatics
- Flexible Infrastructures for Bio-Informatics Information Integration
- Trust, Information Quality and Security
- Emergent Semantics
 - Investigate Socio-cultural and Anthropological approaches