

---

# Can I read a publication like I read a document ?

- Semantic Web, Inter-Ontology Interoperation and Loss of Information

Vipul Kashyap

Applied Research, Telcordia Technologies

Colloquium Talk, Department of Computer Science,  
University of Maryland, College Park  
1st November, 2001

# Outline

---

- Ontologies and the Semantic Web
- The OBSERVER System
- Controlled and Incremental Query Expansion across Multiple Ontologies
  - Ontology Integration and Query Rewriting
  - Intensional Loss of Information
  - Extensional Loss of Information
- Conclusions and Future Work

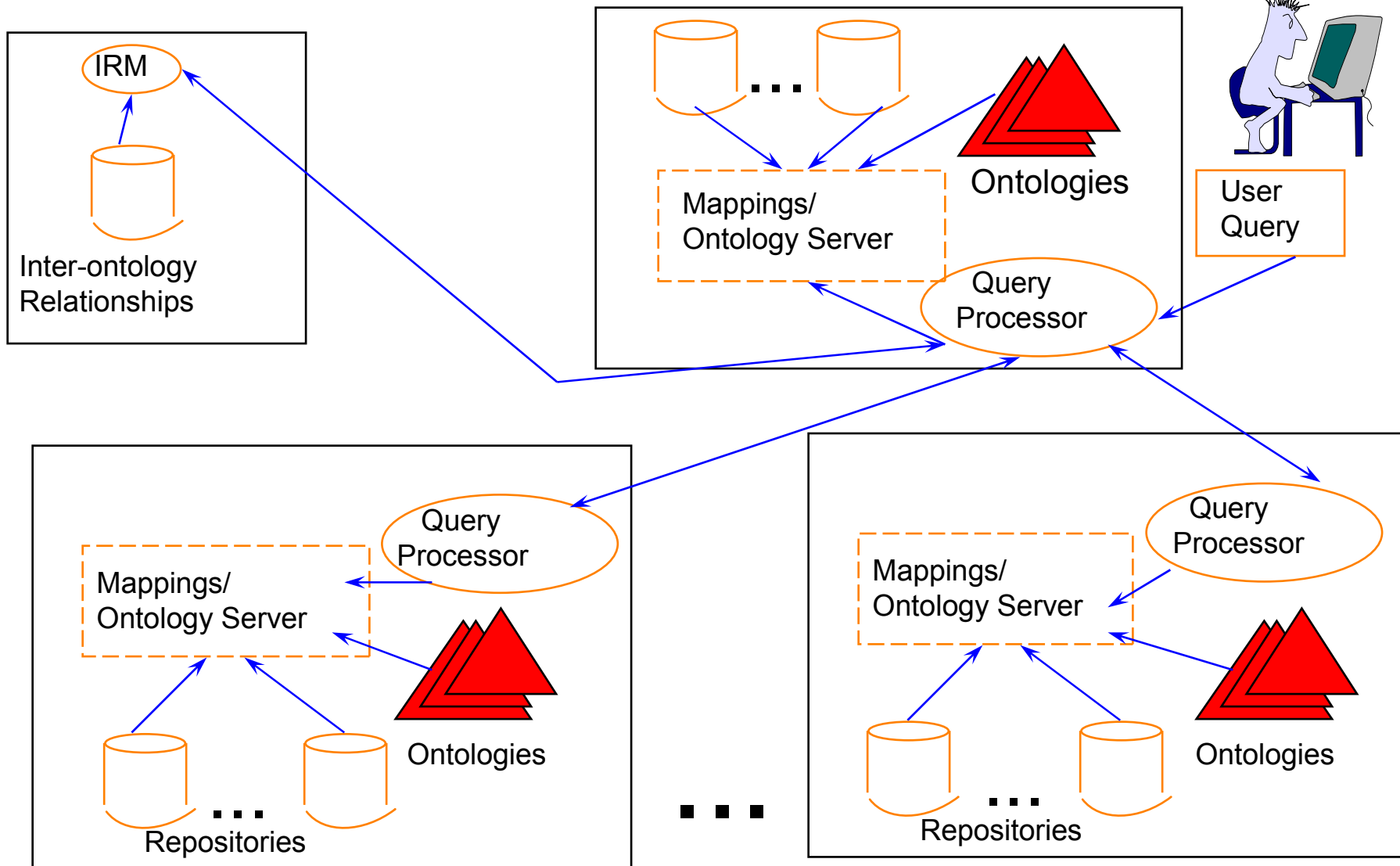
# Ontologies and the Semantic Web

---

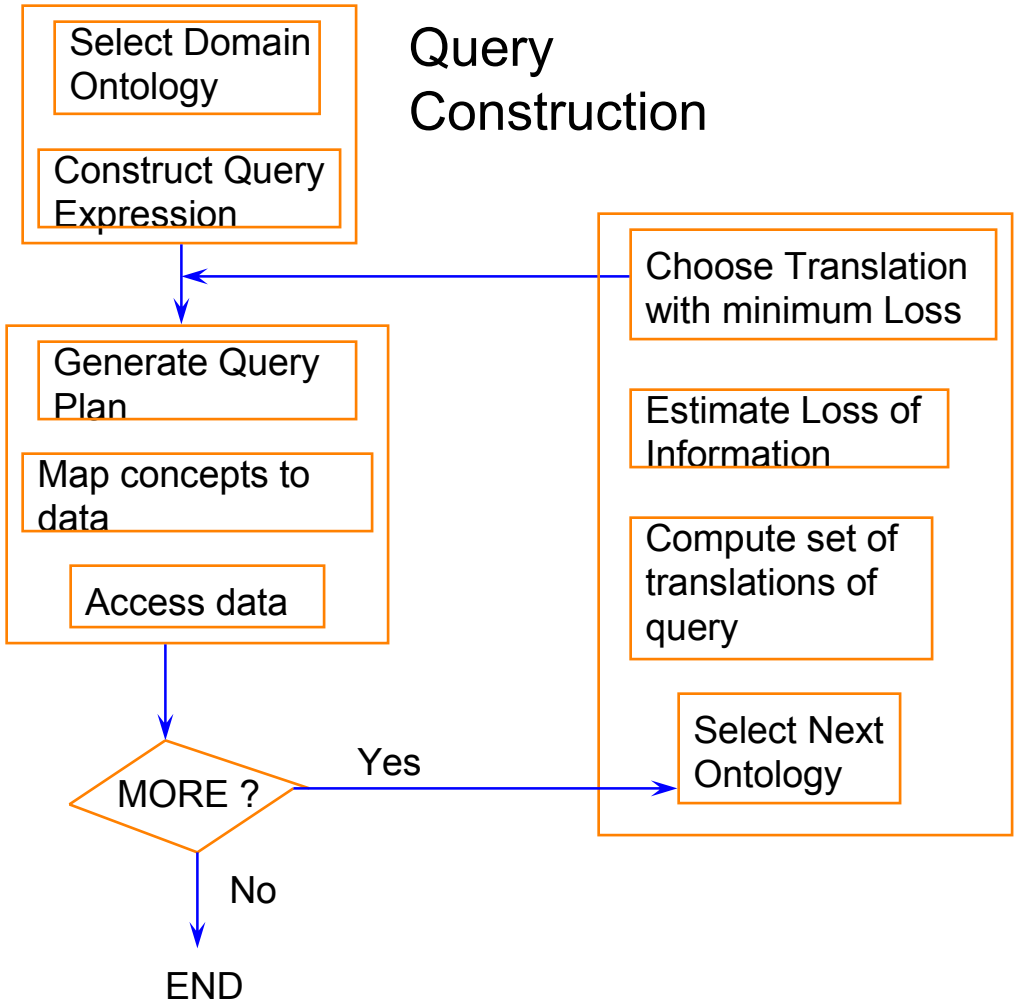
- Semantic Web:
  - An extension of the current web, in which information is given well-defined meaning, better enabling computers and people to work in cooperation [Berners-Lee, Hendler, Lassila, 2001]
- Ontologies: A critical component of the Semantic Web
  - A description (similar to a formal specification of a program) of concepts and relationships that can exist for a community of agent
  - Semantically rich metadata capturing the information content of underlying data repositories
  - DL descriptions organized as a lattice
- Ontology-based Information Access
  - Ability to express information needs at higher semantic level of abstraction
  - Ability to “navigate” multiple pre-existing ontologies enabling wider and focused access to information

# OBSERVER:

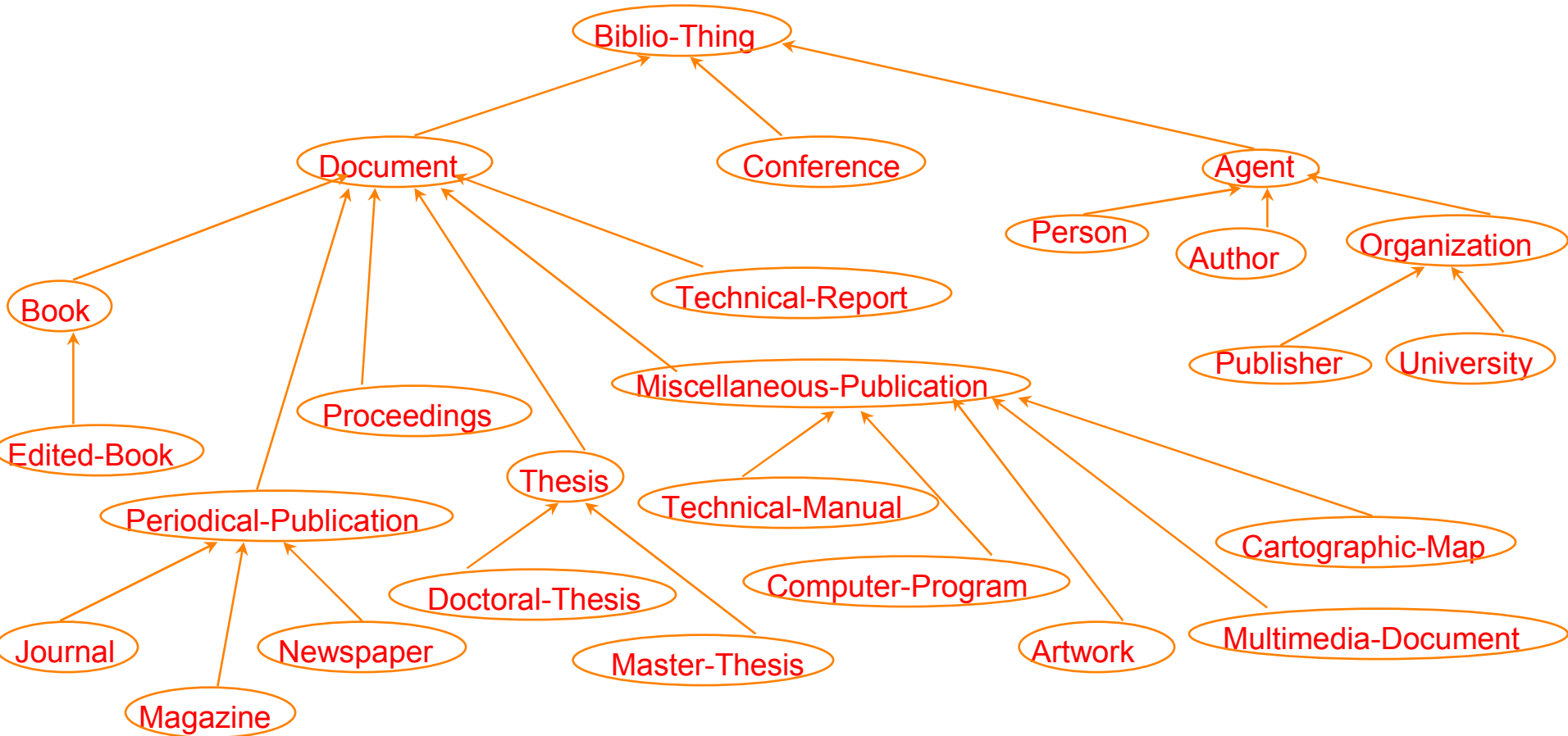
## Ontology-based System Enhanced with (terminological) Relationships for Vocabulary hETerogeneity Resolution



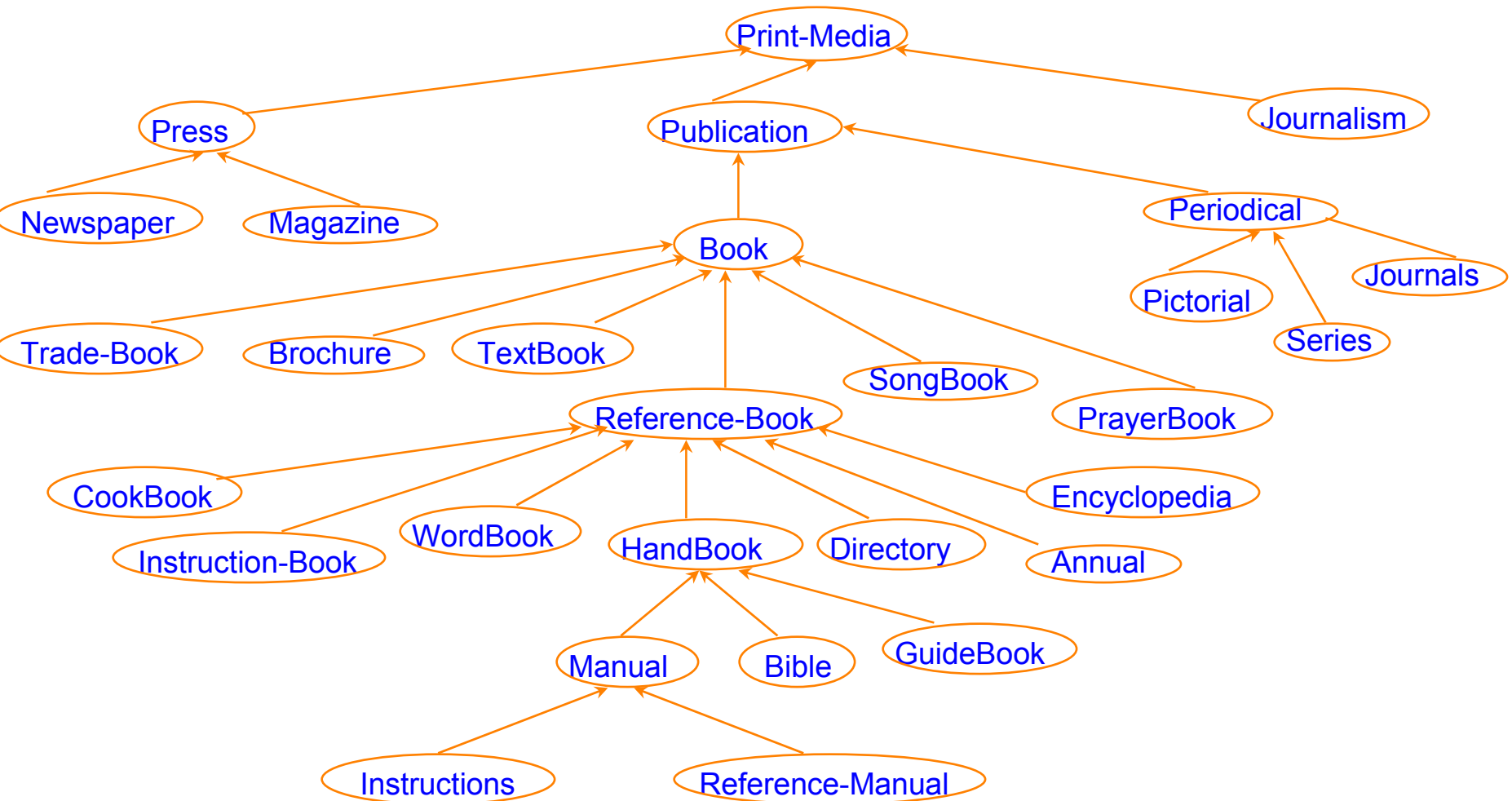
# Controlled and Incremental Query Expansion to a new Ontology



# Bibliography Data Ontology: The Red Ontology



# A subset of WordNet 1.5: The Blue Ontology



# Inter-ontological relationships

---

## ■ Synonyms

- leads to semantics preserving translations

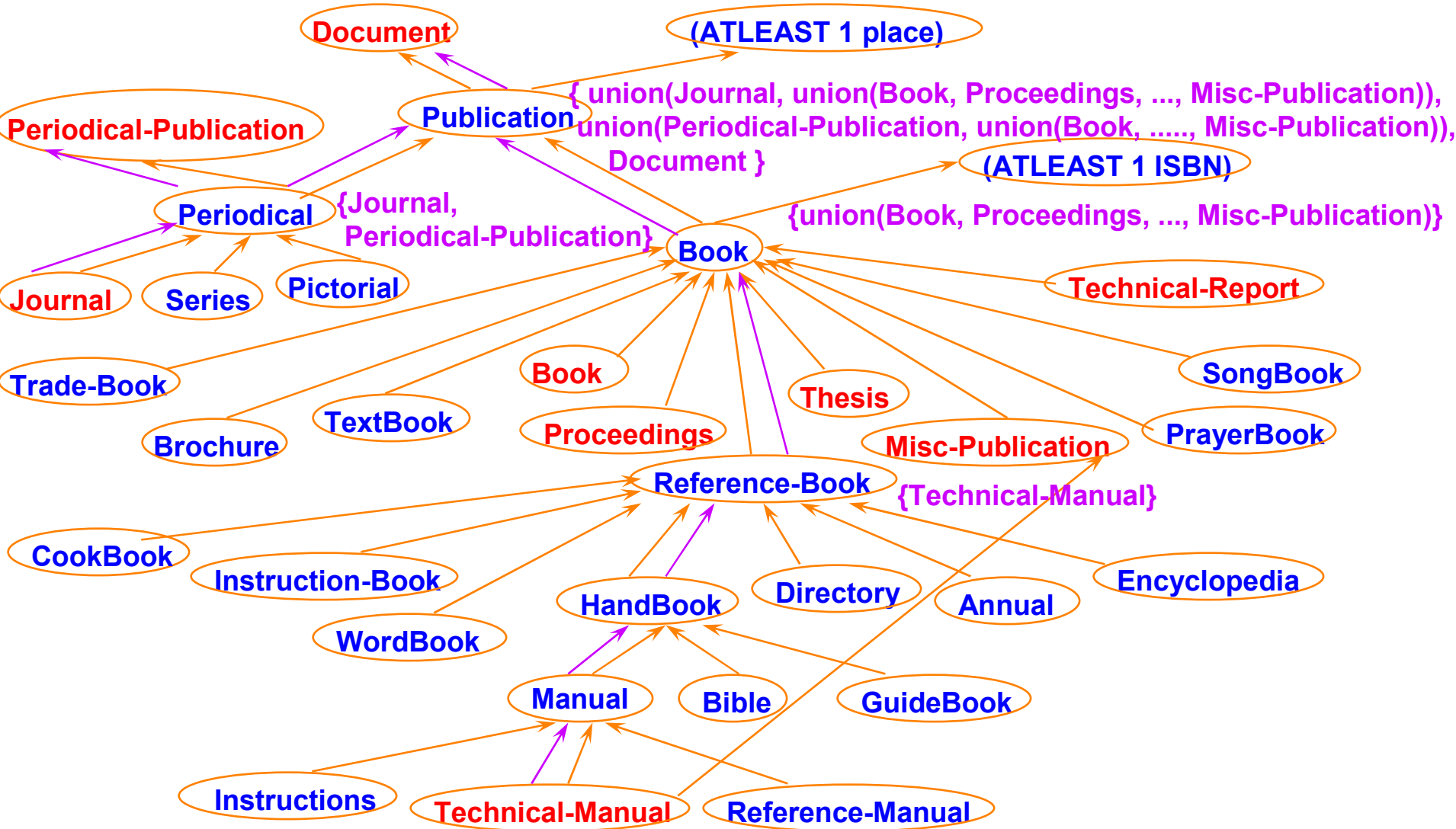
## ■ Hyponyms/Hypernyms ←

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

## ■ List of Hyponyms

- |                     |                |                        |
|---------------------|----------------|------------------------|
| - technical-manual  | <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                   |
| - press             | <i>hyponym</i> | periodical-publication |
| - periodical        | <i>hyponym</i> | periodical-publication |

# Ontology Integration and Query Rewriting



# Intensional Loss of Information

---

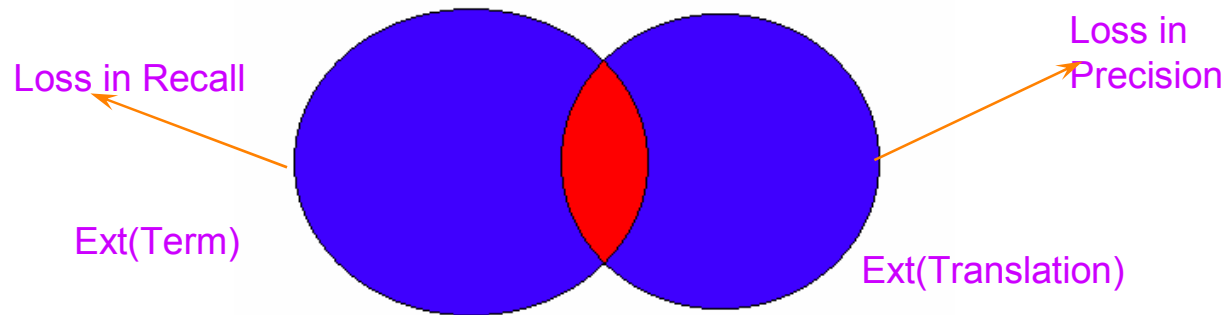
- Original Query:
  - [NAME PAGES] for (AND BOOK (FILLS CREATOR “Carl Sagan”))
- Modified Query:
  - [NAME PAGES] for (AND document (FILLS doc-author-name “Carl Sagan”))
- Terminological Relationships:
  - BOOK  $\Leftrightarrow$  (AND PUBLICATION (ATLEAST 1 ISBN))
  - PUBLICATION  $\Leftrightarrow$  (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: Disadvantages and Advantages

---

- 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.

# Estimating Loss of Information based on Term Extensions



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

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

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

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

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

# Estimating Term Extension Intervals

---

## ■ Intersections

- $|\text{Ext}(\text{Expr}_1) \cap \text{Ext}(\text{Expr}_2)|.\text{low} = 0$
- $|\text{Ext}(\text{Expr}_1) \cap \text{Ext}(\text{Expr}_2)|.\text{high} = \min (|\text{Ext}(\text{Expr}_1)|.\text{high}, |\text{Ext}(\text{Expr}_2)|.\text{high})$

## ■ Unions

- $|\text{Ext}(\text{Expr}_1) \cup \text{Ext}(\text{Expr}_2)|.\text{low} = \max (|\text{Ext}(\text{Expr}_1)|.\text{low}, |\text{Ext}(\text{Expr}_2)|.\text{low})$
- $|\text{Ext}(\text{Expr}_1) \cup \text{Ext}(\text{Expr}_2)|.\text{high} = |\text{Ext}(\text{Expr}_1)|.\text{high} + |\text{Ext}(\text{Expr}_2)|.\text{high}$

## ■ Term

- $|\text{Ext}(\text{Term})|.\text{high} = |\text{Ext}(\text{Term})|.\text{low} = |\text{Ext}(\text{Term})|$

# Estimating Intervals of Information Loss

---

- Intervals of Precision and Recall
  - Precision.high, Precision.low
  - Recall.high, Recall.low
- Leads to Intervals of Information Loss

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

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

# Comparison of two translations

---

- Consider two translations:
  - Trans<sub>1</sub> with bounds low<sub>1</sub> and high<sub>1</sub>
  - Trans<sub>2</sub> with bounds low<sub>2</sub> and high<sub>2</sub>
- Choosing the appropriate translation.
  - Compute  $mLoss_i = (low_i + high_i)/2$ 
    - ☞ if  $mLoss_1 < mLoss_2$ , choose Trans<sub>1</sub>
    - ☞ if  $mLoss_2 < mLoss_1$ , choose Trans<sub>2</sub>
    - ☞ if  $mLoss_1 = mLoss_2$ , choose translation with lesser interval ( $high_i - low_i$ )
- Need for probabilistic models
  - Let  $(low_1, high_1) = (10\%, 80\%)$  and  $(low_2, high_2) = (20\%, 60\%)$
  - $mLoss_2 (40\%) < mLoss_1 (45\%) \Rightarrow$  Trans<sub>2</sub> is chosen
  - However there are cases for which Trans<sub>1</sub> returns a lower (10% - 20%) loss !

# Semantic Adaptation of Precision and Recall

- Term subsumes Translation
  - $\text{Ext}(\text{Term}) \subseteq \text{Ext}(\text{Translation}) \Rightarrow \text{Ext}(\text{Term}) \cap \text{Ext}(\text{Translation}) = \text{Ext}(\text{Translation})$
  - 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

# Semantic Adaptation of Precision and Recall

---

- Translation subsumes Term
  - Analogous (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

# Computation of Precision and Recall in the absence of Semantic Relationships

## ■ Precision

– Precision.low = 0

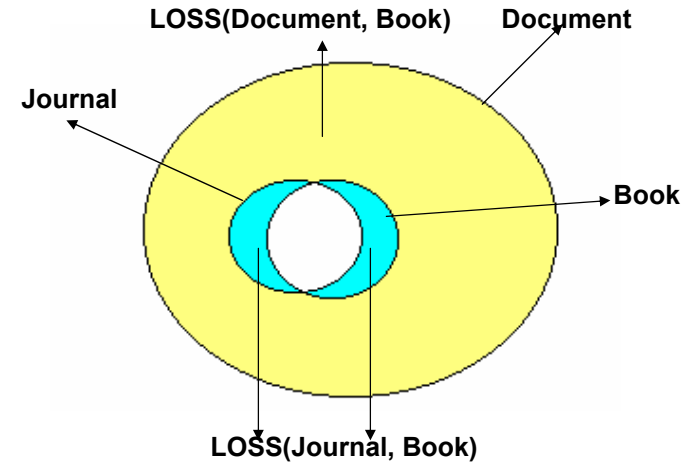
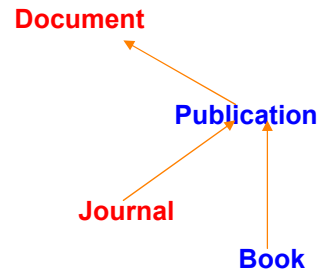
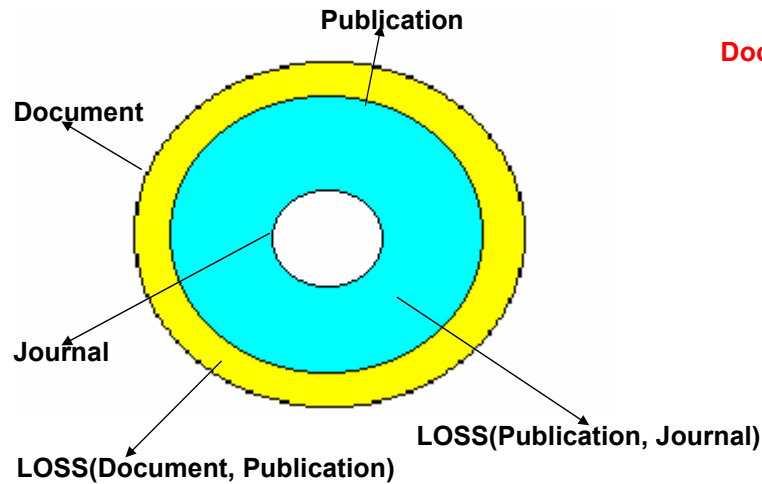
– Precision.high =  $\max\left[ \frac{\min(|\text{Ext}(\text{Term})|, |\text{Ext}(\text{Translation})|.high)}{|\text{Ext}(\text{Translation})|.high}, \frac{\min(|\text{Ext}(\text{Term})|, |\text{Ext}(\text{Translation})|.low)}{|\text{Ext}(\text{Translation})|.low} \right]$

## ■ Recall

– Recall.low = 0

– Recall.high =  $\frac{\min(|\text{Ext}(\text{Term})|, |\text{Ext}(\text{Translation})|.high)}{|\text{Ext}(\text{Term})|}$

# Choosing an optimal translation: Local v/s Global Decision Making



## ■ Local Decision Making

- $\text{LOSS}(\text{Publication}, \text{Journal}) > \text{LOSS}(\text{Document}, \text{Publication})$
- **Document** is chosen as the translation
- But  $\text{LOSS}(\text{Book}, \text{Document}) > \text{LOSS}(\text{Book}, \text{Journal})$  !!

## ■ Global Decision Making

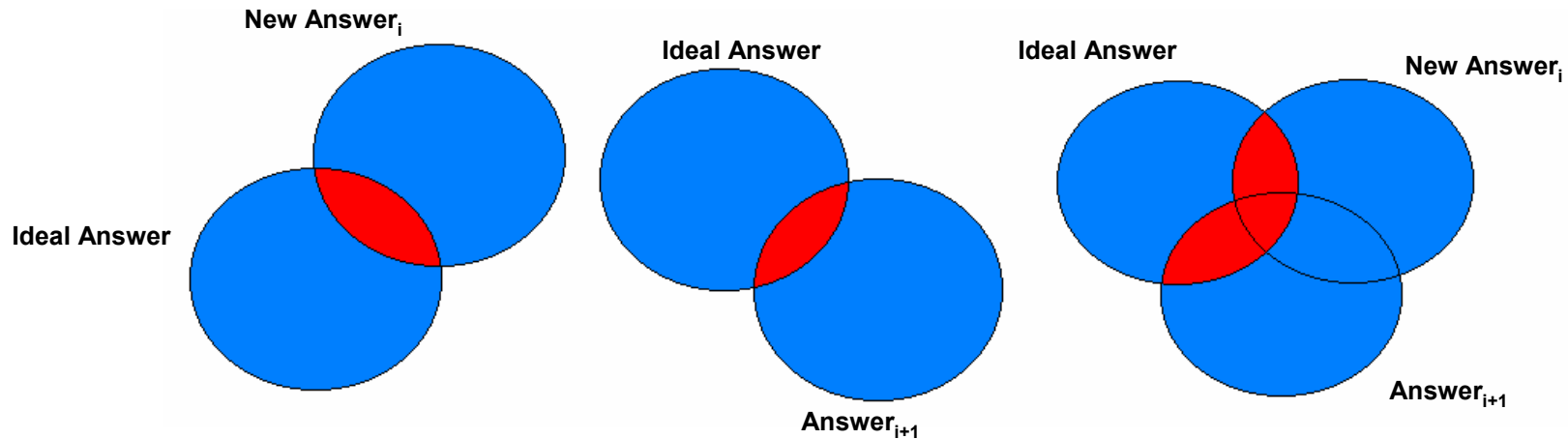
- Both translations {**Document**, **Journal**} are passed on to the next level
- **Journal** is chosen as the appropriate translation

# Loss of Information for Correlated Answers across Ontologies

---

- Let:
  - Answer<sub>i</sub> be obtained from Query<sub>i</sub> executed over ontology O<sub>i</sub>
  - i = 1 corresponds to the user ontology and i = 2,3 correspond to target ontologies
  - NewQuery<sub>2</sub> = Query<sub>1</sub> ∪ Query<sub>2</sub>
  - NewAnswer<sub>2</sub> = Answer<sub>1</sub> ∪ Answer<sub>2</sub>
- Suppose Answer<sub>2</sub> and Answer<sub>3</sub> are obtained from the same ontology.
  - If either has Recall = 1
    - ☐ then system cannot improve answer
  - If both have Recall = 1
    - ☐ then NewAnswer<sub>3</sub> = Answer<sub>1</sub> ∪ (Answer<sub>2</sub> ∩ Answer<sub>3</sub>)
  - If both have Recall < 1
    - ☐ then NewAnswer<sub>3</sub> = NewAnswer<sub>2</sub> ∪ Answer<sub>3</sub>

# Loss of Information for Correlated Answers across Ontologies



- $\text{NewAnswer}_i = \text{Correlated answer from previous ontologies } (O_1, \dots, O_i)$
- $\text{Answer}_{i+1} = \text{Answer obtained from new target ontology } O_{i+1}$
- The following case arise:
  - $\text{NewAnswer}_{i+1} = \text{NewAnswer}_i \cup \text{Answer}_{i+1}$
  - $\text{Loss}(\text{NewAnswer}_{i+1}) > \text{Max loss defined by user}$
  - $\text{NewAnswer}_i$  and  $\text{Answer}_{i+1}$  are displayed separately to the user with an appropriate warning

# Conclusions and Future Work

---

- Pre-existing real world ontologies (“off the shelf”) can be re-used to access information
- Mechanisms for adapting queries across different ontologies.
- Loss of information measures to determine the semantic appropriateness of a particular ontology and translation
- Approach for adaptation of information loss based on semantic relationships
  
- Extensions to current work
  - Evolve a common framework to relate subsumption with loss of information
    - ▣ role subsumption and loss of information
    - ▣ relational algebra-like operators and loss of information
    - ▣ “extensional” and other inter-ontological relationships such as “part-of”
  - Explore relationships with standards such as SQL, XML/RDF based QLs, DAML+OIL
  - Complex probabilistic modeling for ranking translations
- Experimentation and Validation of measures for Loss of Information
- Re-use of existing vocabularies, ontologies and metadata standards
  - Cyc, InterMed, UMLS, GEMET, EDR, SDTS, FGDC, OGIS
- Semi-Automatic Generation of Domain Ontologies and Inter-Ontological Relationships
  - Data Mining, Clustering Techniques, Machine Learning