Knowledge Mining for Intelligent Geospatial Data Discovery

Wenwen Li
GeoDa Center for Geospatial Analysis and Computation
School of Geographical Sciences and Urban Planning
Arizona State University
05-07-2013

Outline

- Background
- Intelligent geospatial data discovery
- A knowledge mining approach
- Results
- Conclusion and discussion

From GIS to CyberGIS

- Analysis
  - Single desktop to cluster-based, cloud-based remote computing
- Data
  - Centralized database to distributed web-accessible database/catalog on the CyberSpace
- Resource discovery
  - Distributed data resources
  - Distributed analytical resources

Current Efforts

DATA.GOV
EMPOWERING PEOPLE

European Commission
INSPIRE Geoportal

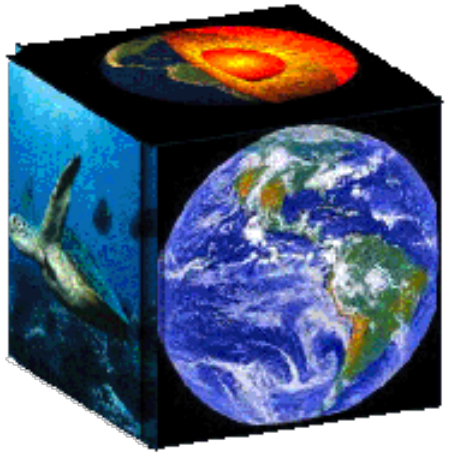
http://inspire-geoportal.ec.europa.eu

GEON

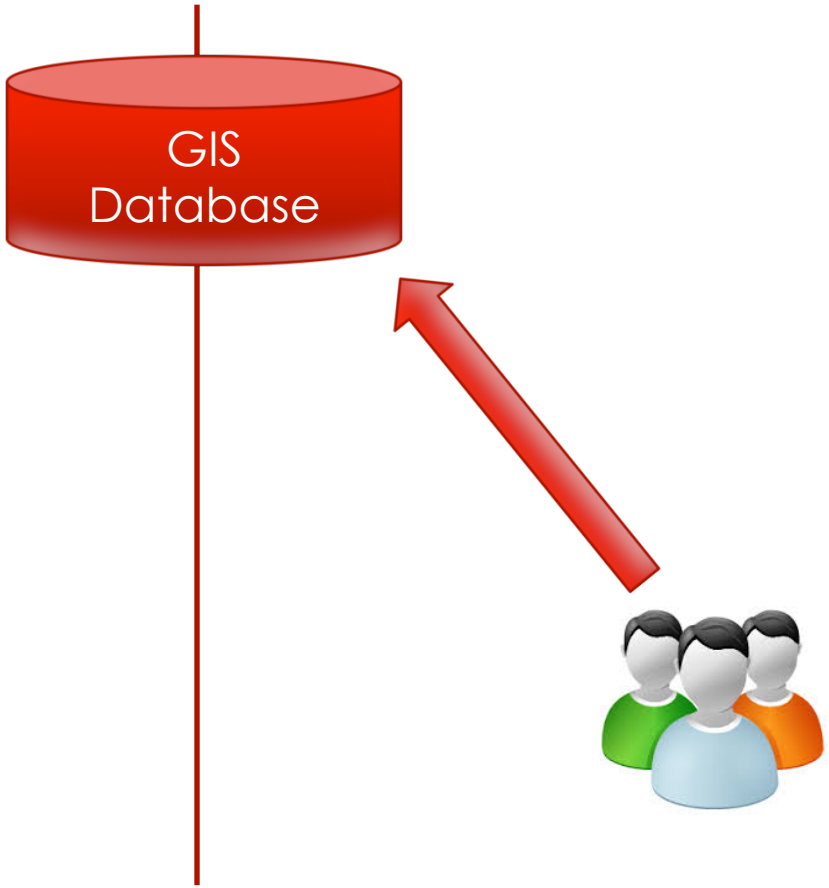
http://www.geongrid.org

GeoNetwork
Opensource

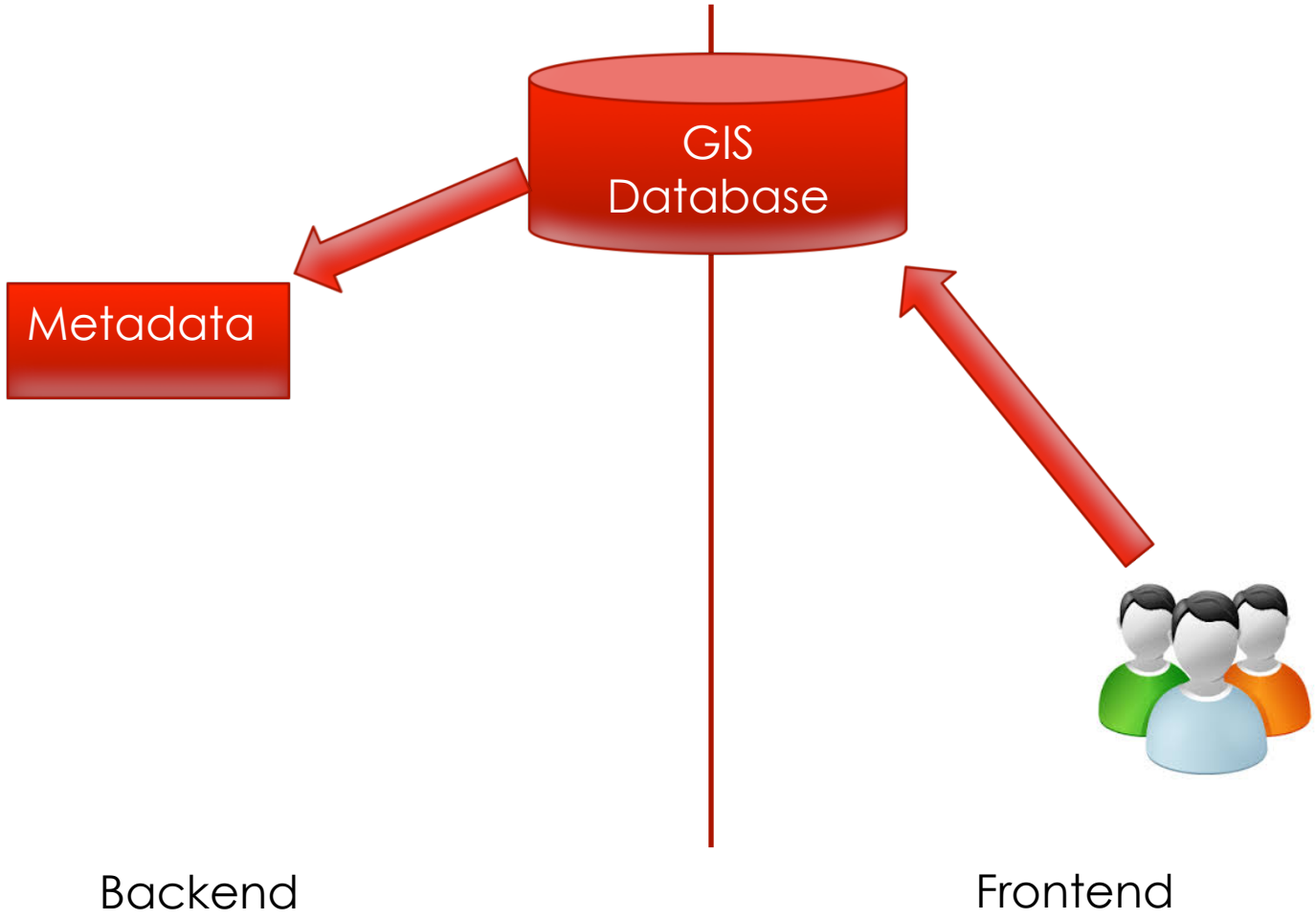
http://geonetwork-opensource.org/



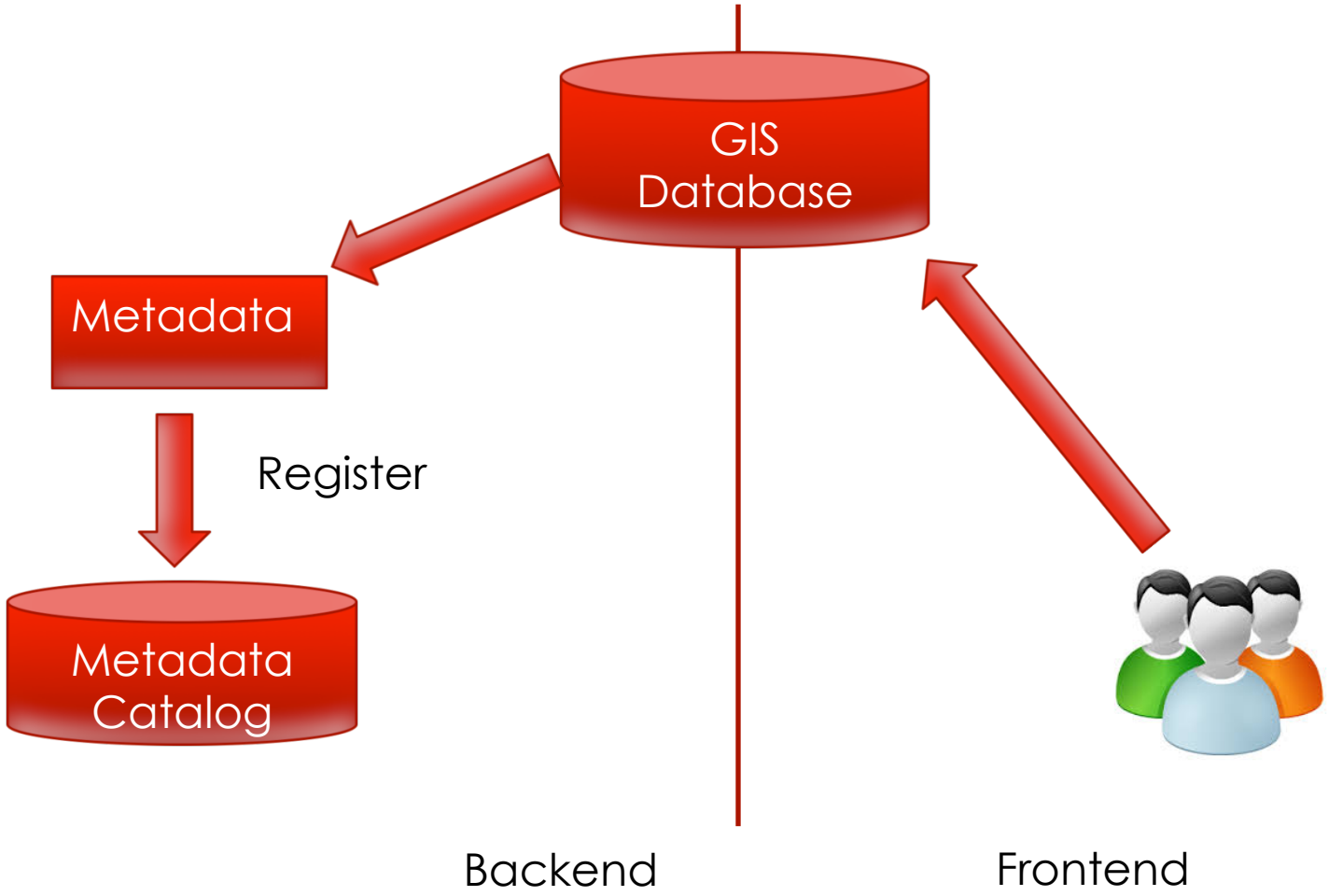
Problem Statement



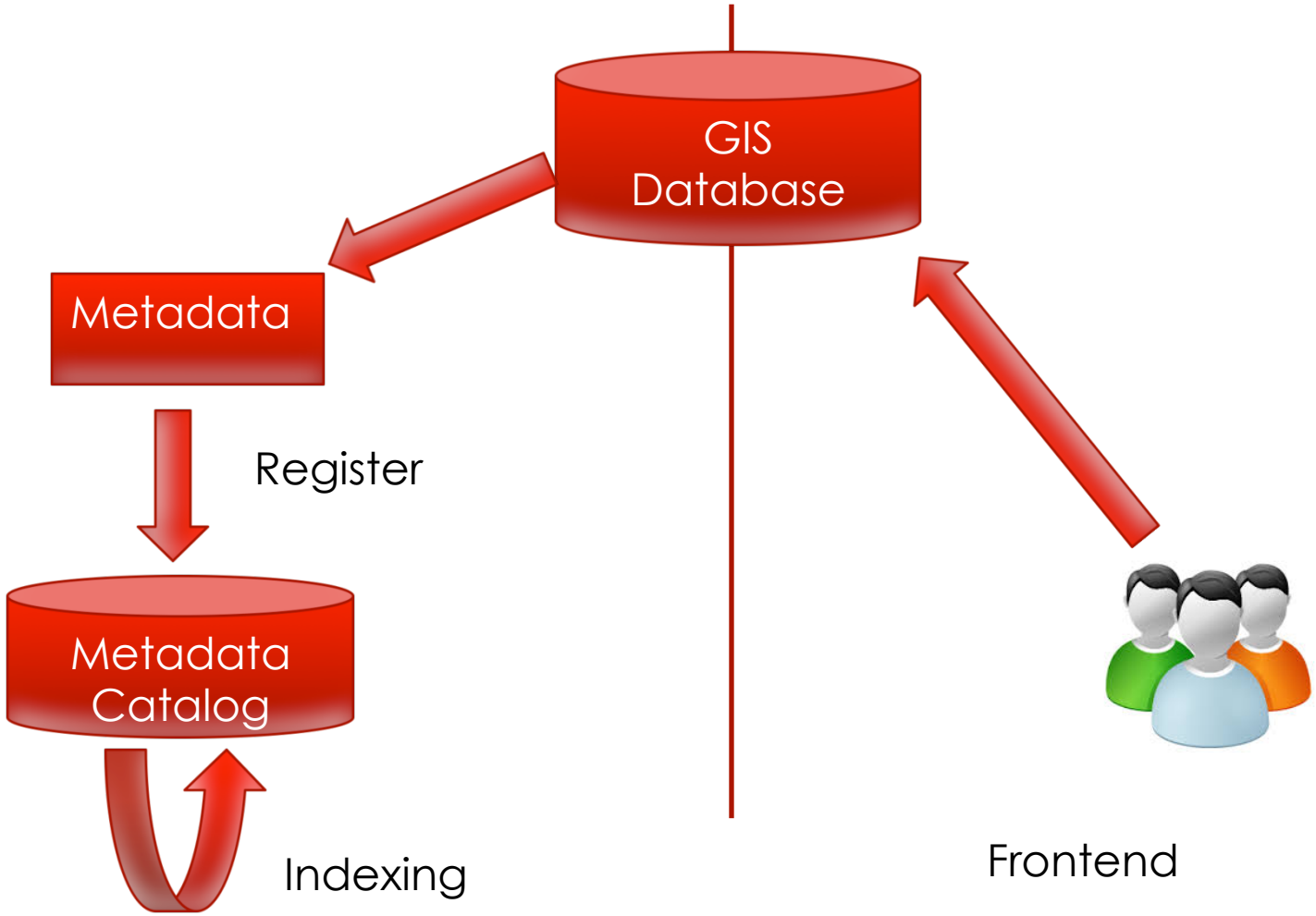
Problem Statement



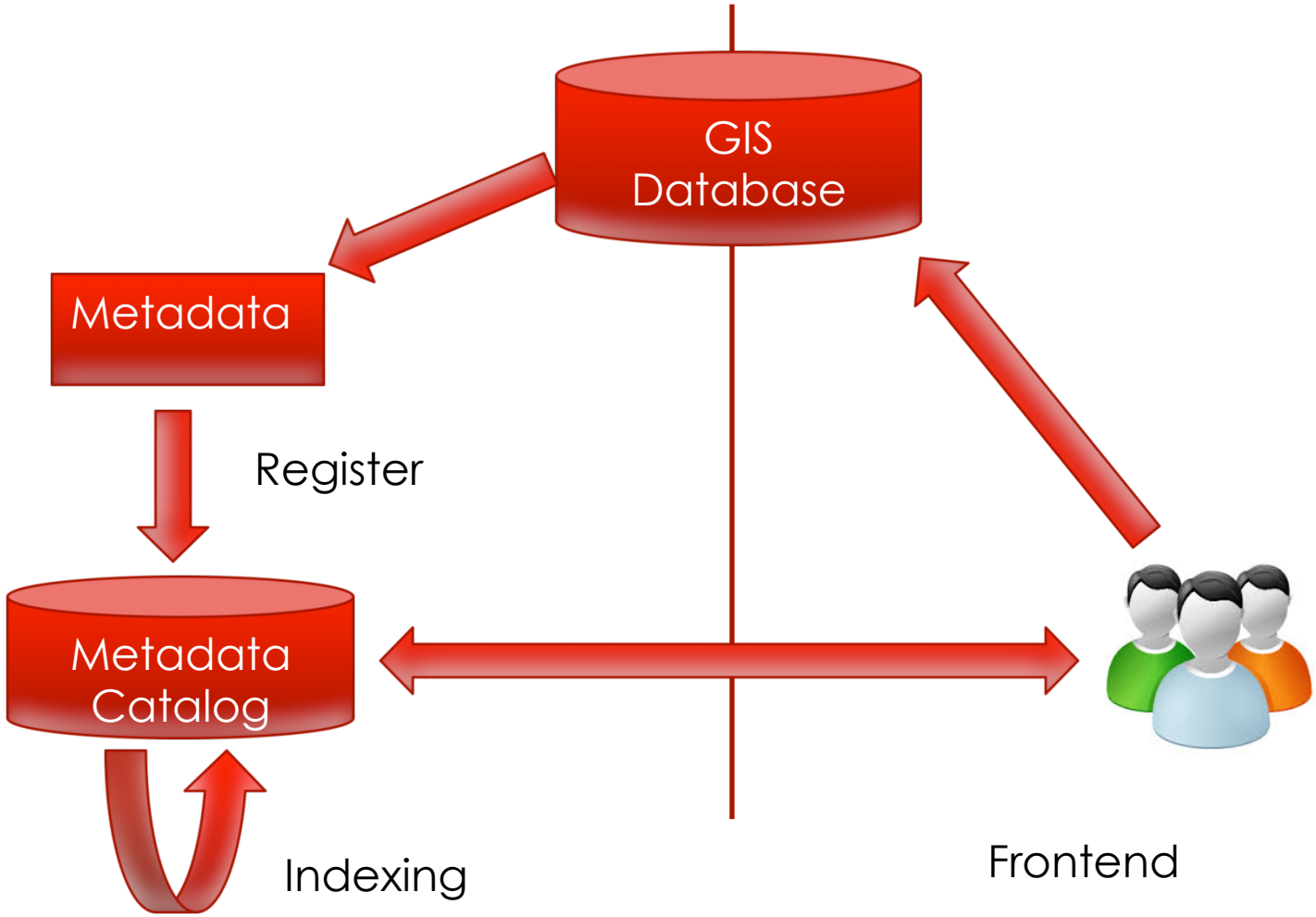
Problem Statement



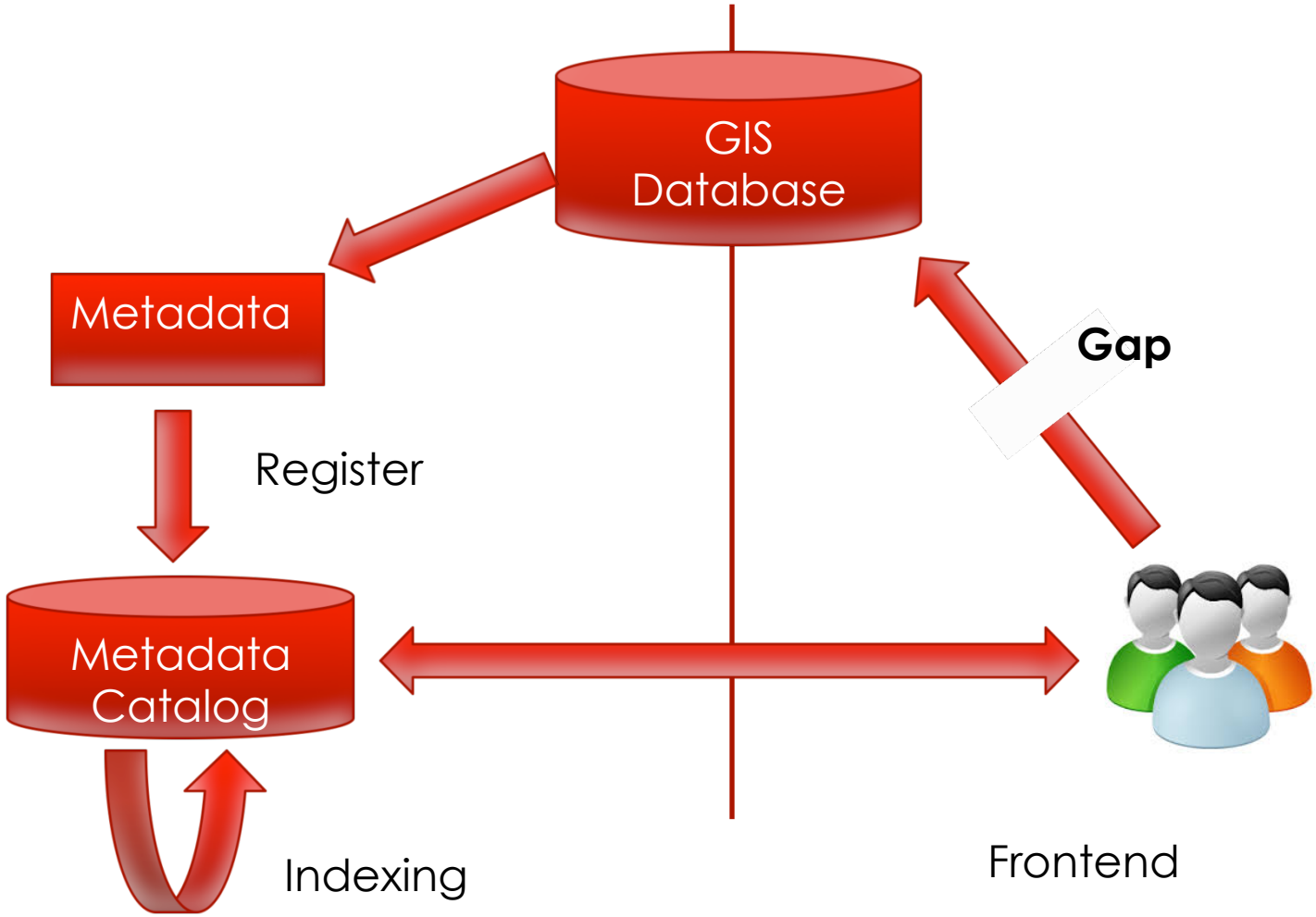
Problem Statement



Problem Statement



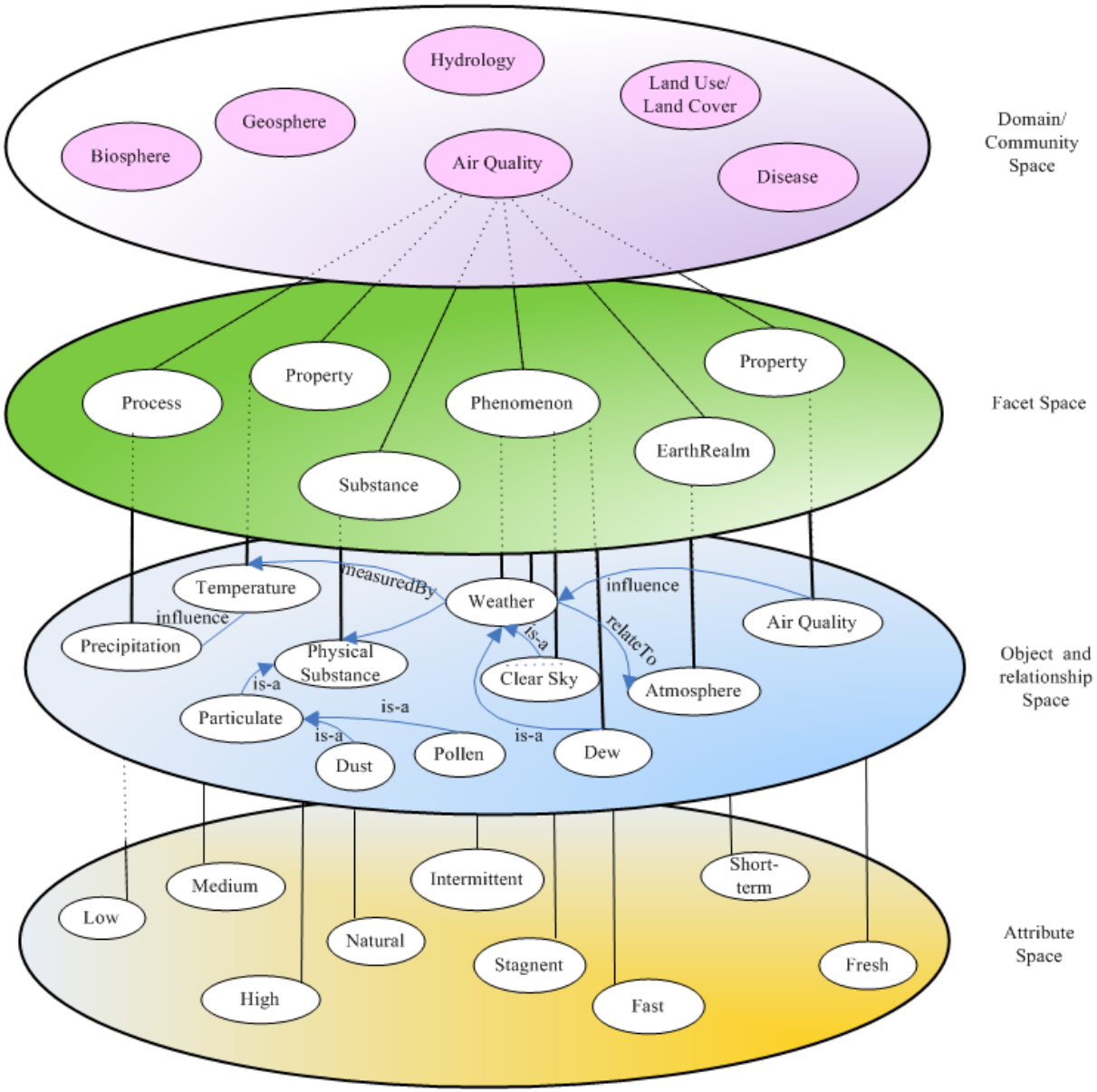
Problem Statement



Intelligent Geospatial Data Discovery

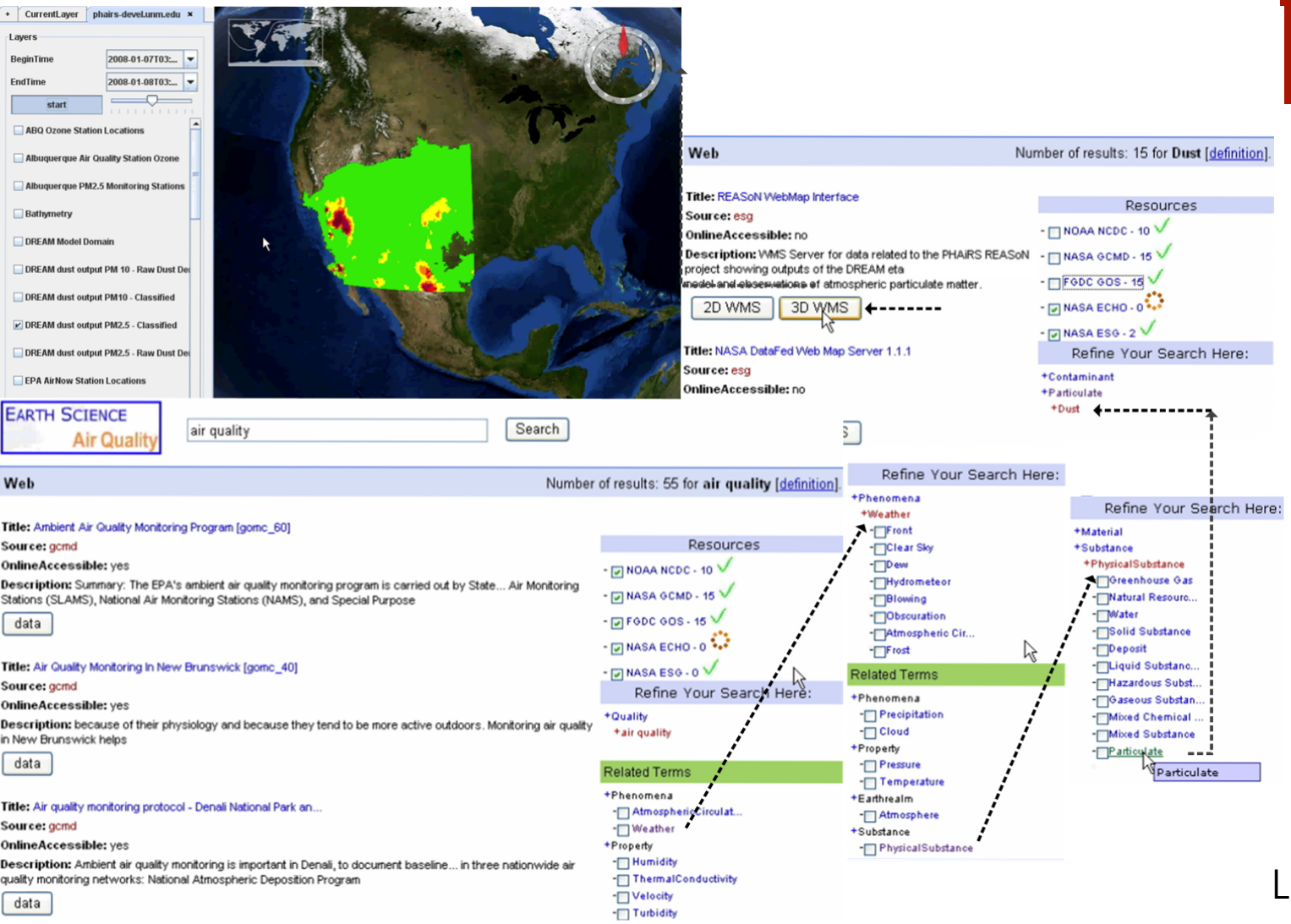
- Ontology-based approach
  - Formalize the definition of domain knowledge
  - Classes, individuals, properties, relationships
- Current Work:
  - GEON (Bowers et al. 2004)
  - LEAD (Droegemeler et al. 2005)
  - VSTO (Fox et al. 2008)
  - SWEET based semantic search (Li et al. 2011, Li et al. 2012a)

A Conceptual Framework



Li et al. 2010

Semantic Search



Li et al. 2010

Limitation

- Hard to model spatial relationship (Shi, 2011)
  - Equal, within, touch, disjoin, intersect..
- Hard to build a consensus domain ontology (ontology mapping)

Ontology	Full Name	Creator
SWEET	Semantic Web for Earth and Environmental Terminology	NASA JPL
CUAHSI	Consortium of Universities for the Advancement of Hydrologic Science	CUAHSI
MMI	Marine Metadata Interoperability	NSF
INSPIRE	Infrastructure for Spatial Information in Europe	European Commission
GEMET	General Multilingual Environmental Thesaurus	EEA, ETC/CDS

- Limited spatial reasoning capability
  - Similarity reasoning
    - Rodriguez and Egenhofer, 2004; Janowicz et al. 2008; Li et al. 2012a
  - Natural language processing

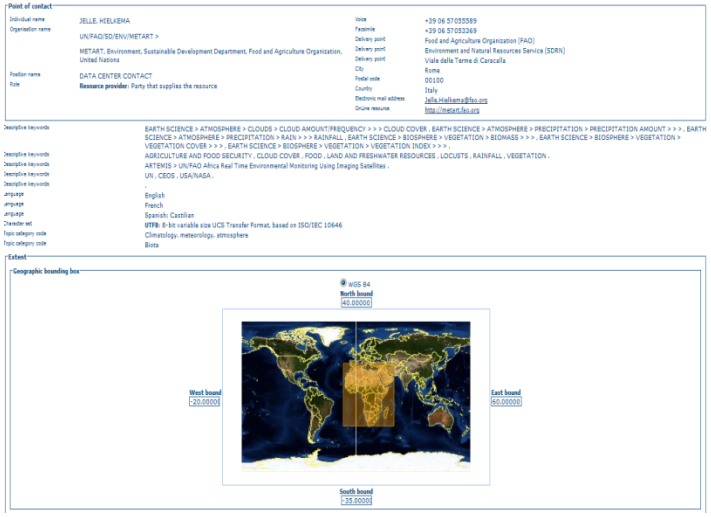
A Knowledge Mining Approach

- Goal:
  - Identify latent semantic associations rather than manually built-up
  - Search based upon meaning rather than appearance
  - Bottom-up approach: let the data speak
- Methodology:
  - Latent Semantic Analysis (LSA)

Intro to LSA

- Mathematical approach for computer modeling and simulation of the MEANING of words and paragraphs
- Identify semantic structure of domain knowledge residing in the metadata files.
  - Concept with similar meanings
  - Similar metadata documents
- Linear Algebra:
  - SVD: Singular Value Decomposition
  - Lower-rank estimation

Intro to LSA



Geospatial metadata in XML

	d1	d2	..	..	..	dn
k1						
k2						
.						
.						
km						

Inverted index: Term-Document Matrix
k: keyword
d: documents

c1: The *geospatial* Web: how *geo*-browsers, social software and the Web 2.0

c2: *Geospatial semantics*: capture meanings of *spatial* information

c3: A *semantic search* engine for *spatial* Web portals

c4: Google's *spatial search* tools in the Marine *Environment* - Decision Support

m1: Darcy's *law* on *hydrology*

m2: *Hydrology* and Water *Law* - Bridging the Gap

m3: *Hydrology*: an *environmental* approach

m4: *Environmental law*: Hazardous wastes and substances

r(geo search) = -0.3333
r(geo law) = -0.4472

S =

[2.7395	0	0	0	0	0	0	0
0	2.3709	0	0	0	0	0	0
0	0	1.6454	0	0	0	0	0
0	0	0	1.2380	0	0	0	0
0	0	0	0	1.0000	0	0	0
0	0	0	0	0	0.7963	0	0
0	0	0	0	0	0	0.0000	0]

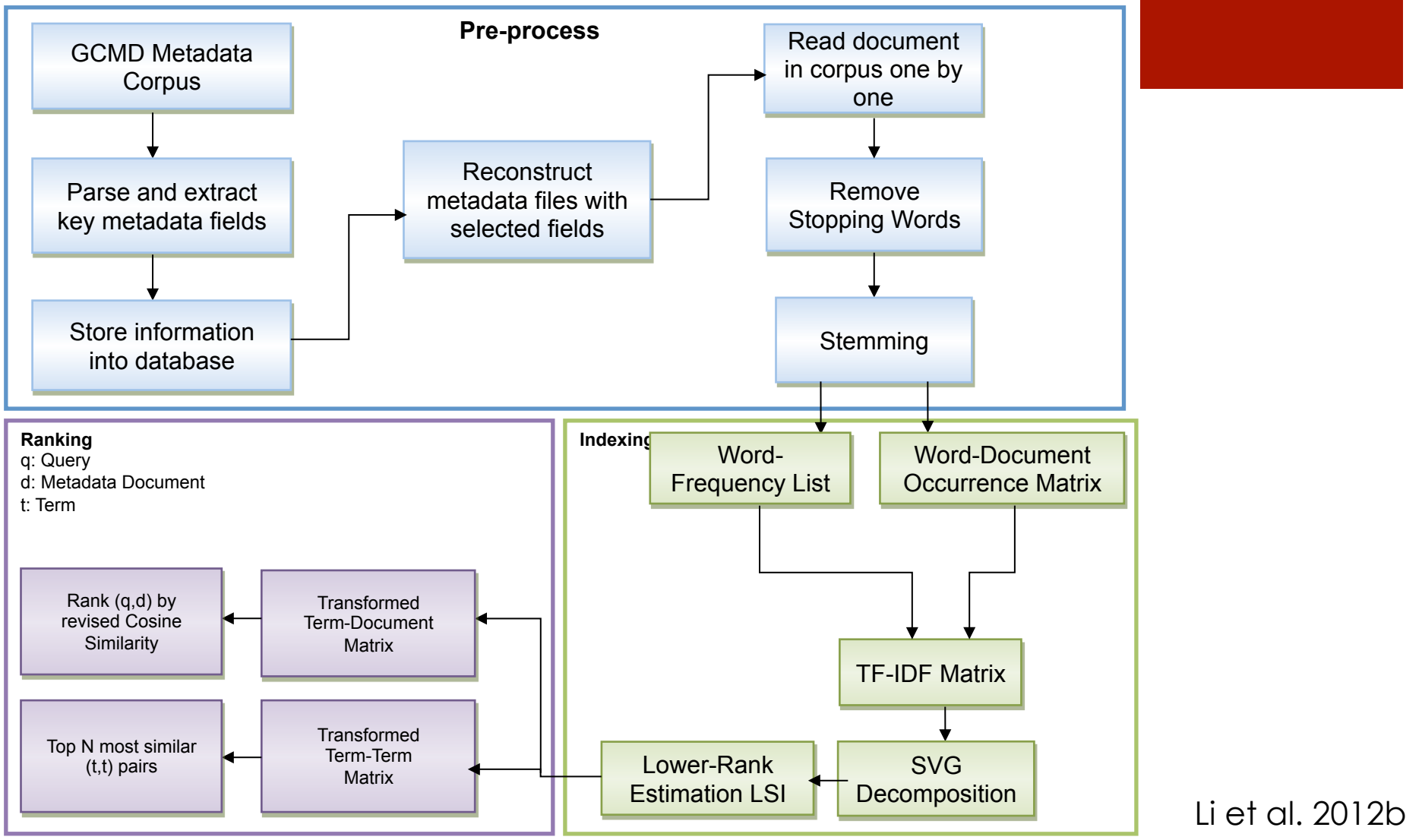
The Term-by-Document Matrix A.

Matrix A	c1	c2	c3	c4	m1	m2	m3	m4
Geo	1	1	0	0	0	0	0	0
spatial	1	1	1	1	0	0	0	0
semantic	0	1	1	0	0	0	0	0
search	0	0	1	1	0	0	0	0
Environment(al)	0	0	0	1	0	0	1	1
law	0	0	0	0	1	1	0	1
hydrology	0	0	0	0	1	1	1	0

Matrix $\hat{A}$	C1	C2	C3	C4	M1	M2	M3	M4
Geo	0.3833	0.5307	0.5322	0.4226	-0.1075	-0.1075	-0.0160	-0.0160
spatial	0.7852	1.0837	1.1007	0.9342	-0.0709	-0.0709	0.0965	0.0965
semantic	0.4459	0.6170	0.6206	0.5008	-0.1051	-0.1051	-0.0013	-0.0013
search	0.4019	0.5529	0.5685	0.5116	0.0366	0.0366	0.1124	0.1124
Environment(al)	0.1697	0.2210	0.2773	0.4591	0.5449	0.5449	0.5055	0.5055
law	-0.0892	-0.1417	-0.0697	0.2553	0.7951	0.7951	0.6700	0.6700
hydrology	-0.0892	-0.1417	-0.0697	0.2553	0.7951	0.7951	0.6700	0.6700

r(geo search) = 0.9961
r(geo law) = -0.9655

System Architecture



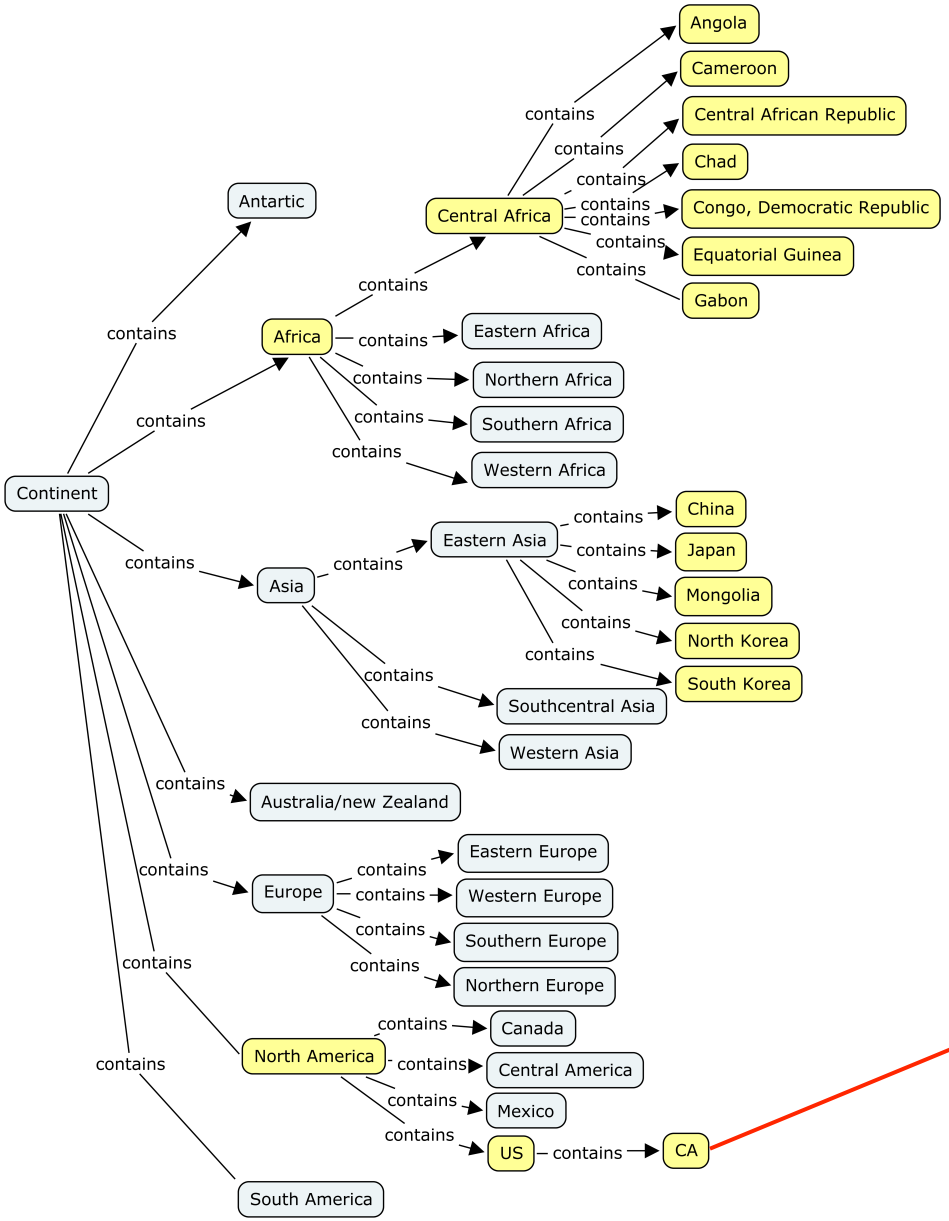
Li et al. 2012b

Geospatial Taxonomy Aided Semantic Search

- Role:
  - Location annotation
    - Emphasize association of science keywords
    - Better handle spatial query with location with keywords
  - Placename detection

```
- <gmd:descriptiveKeywords>
  - <gmd:MD_Keywords>
    - <gmd:keyword>
        <gco:CharacterString>CONTINENT > EUROPE > SOUTHERN EUROPE > SPAIN > GIBRALTAR > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>CONTINENT > EUROPE > SOUTHERN EUROPE > GREECE > > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>CONTINENT > NORTH AMERICA > GREENLAND > > > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>OCEAN > ATLANTIC OCEAN > NORTH ATLANTIC OCEAN > CARIBBEAN SEA > GRENADA > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>OCEAN > ATLANTIC OCEAN > NORTH ATLANTIC OCEAN > CARIBBEAN SEA > GUADELOUPE > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>OCEAN > PACIFIC OCEAN > CENTRAL PACIFIC OCEAN > GUAM > > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>CONTINENT > NORTH AMERICA > CENTRAL AMERICA > GUATEMALA > > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>CONTINENT > AFRICA > WESTERN AFRICA > GUINEA > > </gco:CharacterString>
      </gmd:keyword>
    - <gmd:keyword>
        <gco:CharacterString>OCEAN > ATLANTIC OCEAN > NORTH ATLANTIC OCEAN > CARIBBEAN SEA > HAITI > </gco:CharacterString>
      </gmd:keyword>
    </gmd:MD_Keywords>
  </gmd:descriptiveKeywords>
```

GCMD Location Taxonomy




```
−<boundingBox>
  −<southWest>
      <latitude>32.534290</latitude>
      <longitude>-124.409622</longitude>
    </southWest>
  −<northEast>
      <latitude>42.009460</latitude>
      <longitude>-114.130836</longitude>
    </northEast>
  </boundingBox>
```

Experimental Settings

- Benchmark

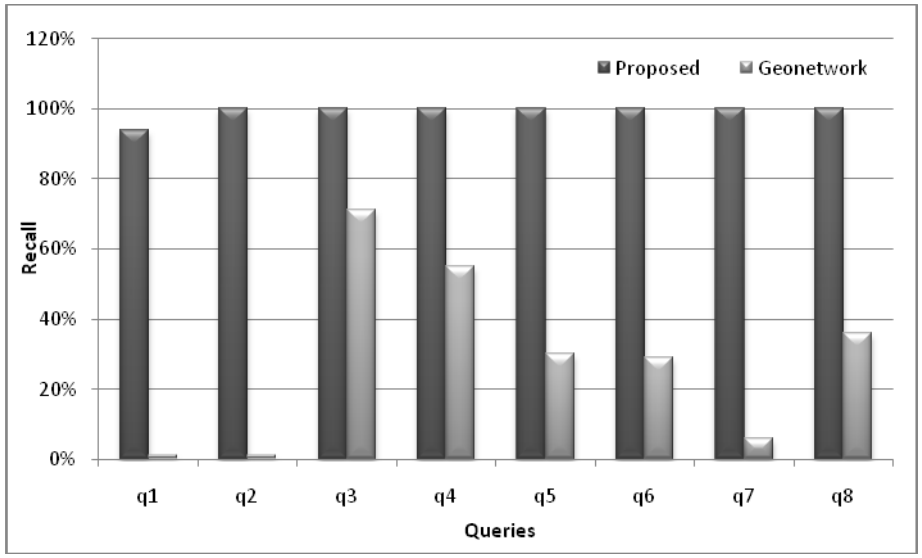
$$\text{Recall} = \frac{|\{\text{all relevant records}\} \cap \{\text{all retrieved records}\}|}{|\{\text{all relevant records}\}|}$$

$$\text{Precison} = \frac{|\{\text{all relevant records}\} \cap \{\text{all retrieved records}\}|}{|\{\text{all retrieved records}\}|}$$

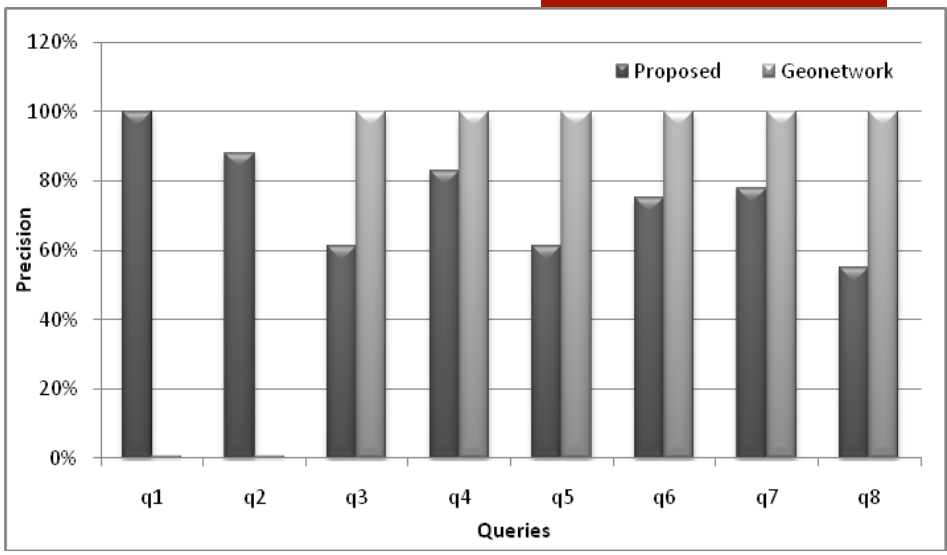
Query Type	Query	Keyword
1	Q 1.1	natural disaster death
	Q 1.2	disaster population impact
	Q 1.3	natural disaster damage
	Q 1.4	wildlife distributions by species
	Q 1.5	global climate change pollution
	Q 1.6	China agriculture food sustainability
	Q 1.7	census housing condition
	Q 1.8	Africa poverty statistics
2	Q 2.1	Colorado population
	Q 2.2	California population dynamics in the United States
	Q 2.3	wild life habitat of Costa Rica
	Q 2.4	China County level population data
	Q 2.5	Puerto Rico census data

Results

Recall



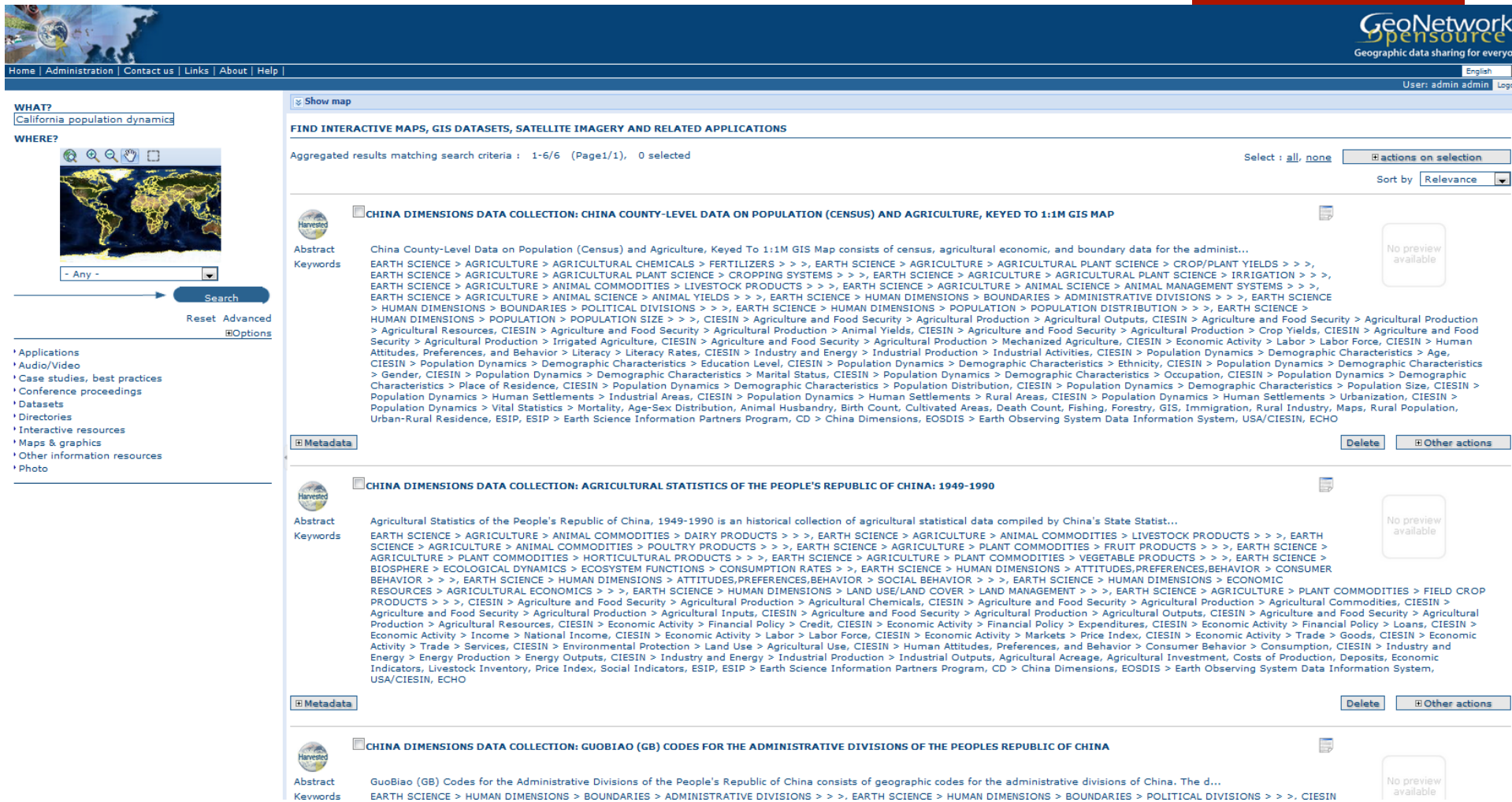
Precision



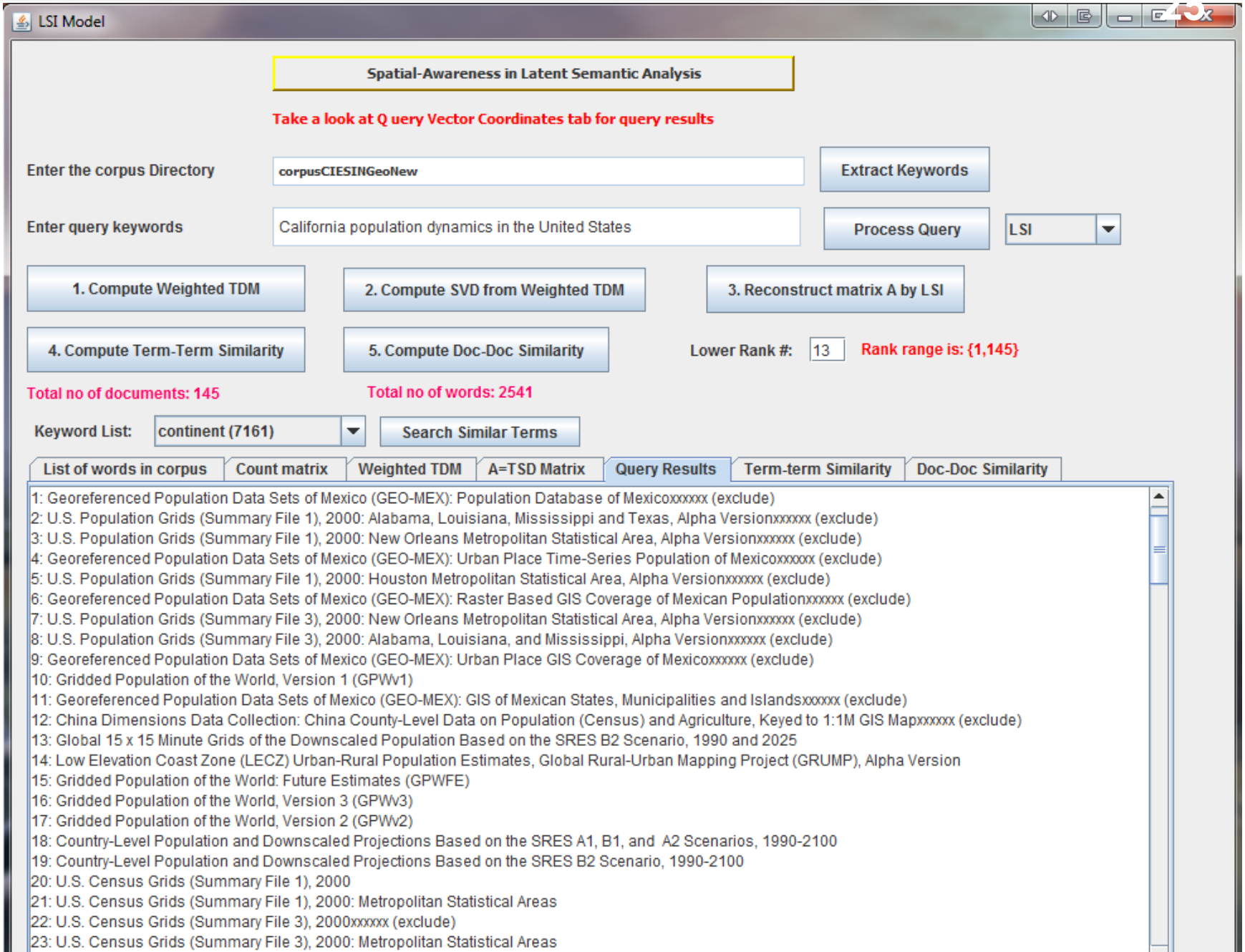
- Q1.1 natural disaster death
  - Global earthquake/flood/volcano/drought/landslide/cyclone mortality
- Q1.5 global climate change pollution
  - SIR: 33 (20 related); Geonetwork: 6
  - Exception: "Global Multi-hazard total economic loss", "environmental protection" -> pollution

    cause: "hazard"

Q 2.2 California population dynamics in the United States



Q 2.2 California population dynamics in the United States



Summary

- A data mining approach to improve geospatial data discovery
- LSA outperforms full-text search
- An alternative approach to establish domain ontology to complement the top-down approach

Future Work

- Integration with Geonetwork
- Refine algorithm

References

- W. Li, Automated Data Discovery, Reasoning and Ranking in Support of Building an Intelligent Geographic Search Engine, Ph.D. Dissertation. George Mason University, August 2010.
- W. Li, C. Yang, D. Nebert, R. Raskin and H. Wu, 2011. Semantic-based service chaining for building a virtual Arctic spatial data infrastructure. Computers & Geosciences. 37(11), 1752-1762.
- W. Li, R. Raskin and M.F. Goodchild, 2012a. Semantic similarity measurement based on knowledge mining: an artificial neural network approach. International Journal of Geographic Information Science, 26(8), 1415-1435.
- W. Li, M.F. Goodchild and R. Raskin, 2012b. Towards geospatial semantic search: exploiting latent semantic relations in geospatial data. International Journal of Digital Earth, DOI:10.1080/17538947.2012.674561.