Semantic Interoperability in Digital Library Systems

Task 3: Semantic Interoperability

WP5: Knowledge Extraction and Semantic Interoperability
DELOS2 Network of Excellence in Digital Libraries

Manjula Patel, Traugott Koch, Martin Doerr, Chrisa Tsinaraki

Background

Author Affiliations

Manjula Patel, UKOLN, University of Bath, UK

Traugott Koch, Netlab Knowledge Technologies Group, Lund University, Sweden

Martin Doerr, ICS FORTH, Crete, Greece

Chrisa Tsinaraki, Technical University of Crete, Greece

State of the Art Report

Integrate views from overlapping communities:

- Semantic Web, AI, KR, ontology, library and information science, computer science etc.

Main Issues

- Why is Semantic Interoperability (SI) important in Digital Library systems (DLs) and how can it be used in DLs?
- Analysis of different types or levels of SI
- Clarify relationship between syntactic and semantic interoperability
- Describe relevant methodologies, prerequisites, standards and tools
- How can SI in DLs be enhanced?
- Identify gaps and open issues

NB: We are still in preliminary stages of writing the report

We would like to use this workshop to engage others in contributing to the report

Report Outline

- Overview
- Introduction and definition of SI
- Importance of SI in DLs
- Theoretical Considerations
- Prerequisites to enhancing SI
- Methods and processes to enhance SI in DLs
- SI in DL Services

Interoperability

- Internet and Web making digital resources widely available
- Finding and assembling information and resources is largely a manual task
- Interoperability provides potential for automation
- *Interoperability in general is concerned with the capability of differing information systems to communicate. This communication may take various forms such as the transfer, exchange, transformation, mediation, migration or integration of information.*
- Ouksel and Sheth identify several types of heterogeneity:

  System: incompatibilities between hardware, operating systems etc.

  Syntactic: differences in encodings and representation

  Semantic: inconsistencies in terminology and meanings

Syntactic Interoperability

- Digital processing of information requires it to be encoded in a machine-processable format e.g. MARC, XML, RDF, OWL etc.
- Syntactic heterogeneity is concerned with differences in the representation and encoding of data
- Achieved when compatible forms of encoding and access protocols are used to allow information systems to communicate
- Identification and naming schemes are important for pulling together related information

Semantic Interoperability (SI)

- Identified as being of primary importance in DL research by NSF Post Digital Libraries Futures Workshop: Wave of the Future, June 2003, http://www.sis.pitt.edu/%7Edlwkshop/
- Goal is to facilitate complex and more advanced, context-sensitive query processing over heterogeneous information resources
- Characterised by the capability of different information systems to communicate information consistent with the intended meaning of the encoded information:
  - *processing of shared information so that it is consistent with the intended meaning*
  - *encoding of queries and presentation of information so that it conforms with the intended meaning regardless of the source of information*

Importance of SI in DLs

- Improving precision of search; enabling advanced types of search
- Enable reasoning over document collections and knowledge bases
- Integration of heterogeneous resources
- SI varies in importance in different parts of the information life-cycle and its management:

1. Creation, modification
2. Publication
3. Acquisition, selection, storage, collection building
4. Cataloguing (metadata, identification/naming, registration), indexing, knowledge organisation, KR, modelling
5. Integration, brokering, linking, update and merging
6. Mediation (user interface, personalisation, recommendation)
7. Access, search and discovery
8. Use, sharing, scholarly communication, annotation, evaluation, reuse
9. Maintenance
10. Archiving and preservation

Theoretical Considerations (1)

- Clarifying and selecting terminology
  - Ontology and Vocabulary
  - Language and Vocabulary
  - Schema, Data Model and Conceptual Model
  - Mapping and Cross-Walks
- Standardization versus Interpretation

  Standardization may comprise:
  - the form and meaning of metadata and content schemata
  - shared concepts defined in KOS
  - use of names and construction of identifiers for concepts and real world items

  Mechanical interpretation may comprise:
  - the mapping of metadata and content schemata (sometimes called crosswalks)
  - correlation of concepts defined in KOS (sometimes called cross-concordances)
  - translation of names and reformatting of identifiers for concepts and real world items.

Theoretical Considerations (2)

- Levels of SI in DL Environments

Three levels of information, that are treated in a distinct manner and give rise to distinct methods to address SI:

Data structures**:** be it metadata, content data, collection management data, service description data.

Categorical data**:** i.e. data that refer to universals, such as classification, typologies and general subjects. Theoretically, one can regard all numbers to belong to this category.

Factual data**:** i.e. data that refer to particulars, such as people, items, places.

Prerequisites to enhancing SI (1)

- Information structure, language and identifiable semantics
- Standards and consensus building
- Role of foundational and core ontologies
- Knowledge Organisation Systems (KOS)
  - Terminology and Taxonomy of KOS
  - Number and size of KOS and NKOS
  - Methods and processes applied
  - Availability, Rights
  - Examples of usage of KOS in Semantic Interoperability applications

Prerequisites to enhancing SI (2)

- Syntactic interoperability and encoding systems
- Role of Semantic Registries
  - Metadata Registries
  - Metadata Schema Registries
  - Registries of Crosswalks or Mappings
  - Ontology Servers
  - Other Terminology Services
- Role of tools and architectures
- Rights Issues

Prerequisites: Semantic Registries (1)

Semantic “look-up” services play an important role by supporting the following types of functionality:

- disclosing concepts, terms and semantic relationships
- promoting consistent use of vocabularies
- publication of semantics
- providing examples of use and best practice
- making accessible information relating to provenance, currency, authoritativeness, deduction and reasoning processes

Prerequisites: Semantic Registries (2)

- The types of entities that are used to determine semantic proximity and that support semantic reconciliation include:

  Vocabularies –describing and organising discrete physical elements or intellectual concepts

  Classifications –efficiently locate objects and arrange them into broad classes

  Taxonomies – record generalisation and specialisation type relationships

  Thesauri –define specific semantic relationships, e.g. broader term, narrower term, synonym, hyponym etc

- Role of Metadata Registries, Metadata Schema Registries, Registries of Crosswalks or Mappings, Ontology Servers, Other Terminology Services

Enhancing SI in DLs

- Standardization of metadata schemas, mediation and data warehousing
  - Schema integration and modular approaches
  - Usage of foundational and core ontologies
  - Data integration
  - KOS compatibility with core ontologies
- Knowledge Organisation Systems (KOS)
  - KOS Standardization and Interlinguas
  - KOS schema standardization
  - KOS upper level compatibility
  - KOS Transformation
  - KOS Correlation
- URI and identifier generation rules

SI in DL Services

- Searching, browsing and navigation
  - Advanced queries
  - Cross-searching and Cross-browsing
  - Brokerage services
- Information tracking
- Transformation
- User interfaces
- Automatic indexing and classification
- Taxonomy services
- Mapping services
- Translation support
- Data and text mining

Omissions, Questions, Comments, Discussion …