Accessing Cultural Heritage Collections Using Semantic Web Techniques

Antoine ISAAC
STITCH Project

DE Conferentie
November 10th 2005
(Some) Needs for managing CH Digital Collections

- Representation of artefacts and knowledge about them
  - Pointing at collection objects
  - Describing them (creating metadata) according to specific structures and vocabularies
    - Metadata schemes
    - Controlled vocabularies, e.g. thesauri

- Accessing them using metadata and viewpoint information
  - Data commonly searched/presented via the MD schemes and the vocabularies
  - Sometimes ‘advanced’ features such as search using the information contained in thesauri
Accessing Cultural Heritage Collections Using Semantic Web Techniques
Semantic Web Main Features

Knowledge representation mechanisms to create machine-readable metadata, allowing systems and users to deal with the growing number of accessible documents on the web

- Pointing at documents/knowledge objects (URIs)
- Enabling structured assertions
  RDF: factual knowledge in the form of subject-property-object triples
- Building those assertions on top of shared and structured views
  Ontologies: declarative specifications of concepts and roles/property (building blocks of assertions) in a given domain
- Controlling existing facts and producing new ones
  RDFS/OWL encoding of ontologies, used by inference engines
  Hierarchical knowledge, domains/ranges of relations, etc.

Using reasoning knowledge amounts to delegate part of the exploitation tasks from the user to the system
**Fundamental Common Features**

- Need to categorize/classify things
  - Both ontologies and thesauri bring concept hierarchies giving the intended meaning of a vocabulary through definition and use links between its items
- Need to structure representations
  - Resorting to MD schemes is similar to using relations
  
  *The representations/description of CH artefacts can be encoded by SW facts*

- Turning to ‘Ontological’ commitments to allow for an intelligent (i.e. proper) way to exploit the information
  
  *SW formal control and inference mechanisms can be considered as a realization of the information management vision that grounded the introduction of controlled descriptions*
Sesame Query Example

Evaluate a SeRQL-select query

Your query:

```
select Xid, Xtitle
from (X) rdf:type {ims:Record}, (X) rdf:type {ic:not_49_b_17_1}, (X) rdf:type {ic:not_34},
(X) ims:record-id {Xid}, (X) ims:title {Xtitle}
using namespace

ims = <http://www.cs.vu.nl/STITCH/pp/ims-schema#>,
ic = <http://www.cs.vu.nl/STITCH/pp/ic8>
```

Response format: HTML

Query results:

<table>
<thead>
<tr>
<th>Xid</th>
<th>Xtitle</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;11356&quot;</td>
<td>&quot;St. George on horseback fighting the dragon; the princess nearby with a dog (?); the king and queen watching the scene from the city-walls&quot;</td>
</tr>
</tbody>
</table>

1 results found in 10 ms.
Differences

• CH documentary systems do not use explicit formal meaning
  • Controlled vocabularies are not ontologies
  • Qualitative gap regarding ‘meaningful exploitation’ enabled
• SW techniques bring more freedom regarding structure expression
  • Semi-structured data vs. structured data
• Consequence: dealing with SW formal knowledge is more complex for machines and designers, but also for users
  • How can we transpose existing information sources (databases and vocabularies)
  • How may formal reasoning knowledge benefits exceed the price of turning to SW solutions?
Main Research Problems

- Ontological representation of vocabularies and collection data
  - Vocabularies: relevant encoding ontologies (cf. W3C SKOS)
  - MD structures: ontologies providing proper collection-related properties/relations
- Reasoning to help exploitation
  - Transposition of hierarchical/associational knowledge encoded in CH resources
  - Design of application-related reasoning knowledge
- Projects
  - UvA/VU: annotation, transposition of thesauri and MD schemes
  - MuseumFinland: browsing/recommendation system, central repository for various collections
**CH Interoperability Problems**

- Current CH trend: portals that build on heterogeneous collections
  - Different databases/vocabularies/MD schemes

- Syntactic interoperability problem being solved?
  - Access can be granted

- The semantic interoperability problem is still to be addressed
  
  Collections usually come with a point of view/commitment which is lost during the aggregation of descriptions
  - Structure: imposed common MD schemes for storage and/or search
  - Content: full-text storage and search, ignoring the semantics of initial vocabularies

- It requires solving difficult problems
  - Representation of different points of view in the same system
  - Creation of new alignment knowledge that would enable the system to work in a seamless way
SW and Interoperability

SW propose solutions to the CH world regarding semantic interoperability:

- « Meta-Language » standardization
  - For metadata (using RDF triples)
  - For metadata scheme/vocabulary (using ontologies in RDFS/OWL)
- Alignment of description languages/points of view
  - Some (often controversial) upper-level and reference ontologies
  - Growing interest for tools matching ontologies without enforcing standardization
Automatic Ontology Matching Techniques

Recognizing equivalence or subsumption links between ontology elements

- **Lexical**
  Labels of entities and textual definitions

- **Structural**
  Structure of the formal definitions of entities, position in the hierarchy

- **Statistical**
  Objects, instantiation of the concepts

- **Shared background knowledge**
  Using a conceptual reference to deduce correspondences

Indeed, most of the mapping tools use a mix of such approaches. E.g., lexical string matching can ignite a structural alignment process.
Applying Ontology Mapping for CH (1)

- Concepts come from thesauri, not from full-fledged formal ontologies
  Only BT/RT, limits semantic/structure-based matching?
- Difficulty to obtain objects indexed against two points of view
  Disqualifies instance-based matching?
- Several problems in a single one
  - Metadata scheme mapping
  - Content vocabulary mapping
  - Alignment between MD elements and vocabulary elements
- We need useful situated mappings, rather than ‘strictly ontological’ mappings
  Focusing on associative RT links rather than definitional ones
Applying Ontology Mapping for CH (2)

Different possible solutions, taking into account the different cases

- MD scheme mapping and thesaurus mapping can use structural techniques
  - BT/NT
  - MD organization
- MD scheme mapping can use background knowledge
  - Existing alignments between MD schemes
  - Dedicated SW-flavored initiatives (CIDOC-CRM)
- Thesaurus/content vocabulary mapping can use lexical techniques
  - Language information found in thesauri (synonymy/UF)

All reasonable combinations have to be tested

*How automatic can be the matching between two CH collections?*
Conclusion

- Applying SW techniques to CH collections is feasible
  - Representation and exploitation
  - Special case of semantic integration
- A research agenda well filled!
- There is room for cross-domain collaboration
- STITCH/CATCH
Accessing Cultural Heritage Collections Using Semantic Web Techniques

**Links**

- **SW at W3C**
  - [http://www.w3.org/2001/sw/](http://www.w3.org/2001/sw/)

- **SW at VU**

- **DE talk by Lynda Hardman**

- **Sesame RDF/S repository**
  - [http://www.openrdf.org](http://www.openrdf.org)

- **Medieval Illuminated Manuscripts**
  - [http://www.kb.nl/kb/manuscripts/](http://www.kb.nl/kb/manuscripts/)

- **Other projects**
  - SKOS, [http://www.w3.org/2004/02/skos/](http://www.w3.org/2004/02/skos/)
  - MuseumFinland, to follow!