Semantic Interoperability in Cultural Heritage

Shenghui Wang

STITCH project
http://www.cs.vu.nl/STITCH/
Vrije Universiteit Amsterdam, NL

SIKS 2008
Outline

1. Introduction
   - Interoperability Problems
   - Cultural Heritage and Semantic Web

2. Towards interoperability in Cultural Heritage
   - Porting thesauri to the Semantic Web
   - Thesaurus alignment

3. Summary
Background

- **CATCH @ NWO**
  - Continuous Access To Cultural Heritage (CATCH)
  - 10 computer science projects applied to CH, including Personalisation of access, image/text/audio analysis, *etc.*
  - Integration of projects in CH institutes (museums, archives)

- **STITCH**
  - SemanTic Interoperability To access Cultural Heritage
  - Goal: Cultural Heritage metadata interoperability
    - Build semantic links between the vocabularies
    - Develop theory, methods and tools
  - Vrije Universiteit Amsterdam (VU), Koninklijke Bibliotheek (KB) and Max Planck Institute (MPI)
About Cultural Heritage collections

- Representation of objects and knowledge about them
  - Pointing at collection artifacts: books, paintings, ...
  - Describing them by creating metadata, using
    - specific metadata structures (metadata schemes)
    - controlled vocabularies (e.g., thesauri)
- Accessible through metadata and thesauri
KB Illustrated Manuscripts – Iconclass

Browse by subject

using the iconclass classification system

Religion and Magic

Nature

earth, world as celestial body

animals

groups of animals

mammals

birds

reptiles

amphibians

fishes

lower animals

mis-shapen animals; monsters

fabulous animals (sometimes wrongly called 'grotesques'); 'Mostri' (Ripa)

animals (+ anatomy of animals)

animals (+ animal(s) in motion; positions, expressions of animals)

Human Being, Man in General

Society, Civilization, Culture
KB Illustrated Manuscripts – Iconclass

Iconography:
Results 1-5 of 114 for: ICONCLASS codes: ("25f6**")

1. The Hague, KB, KA 16
   Gladius maris (swordfish)
   Fol. 106vb1: column min.
   35x60
   iconclass: 25F62(SWORD-FISH)

2. The Hague, KB, KA 16
   Pister
   Fol. 108va2: column min.
   35x55
   iconclass: 25F62("PISTER")

3. The Hague, KB, KA 16
   Synacus
   Fol. 109vb1: column min.
   40x55
   iconclass: 25F62("SYNACUS")

4. The Hague, KB, KA 16
   Tignus
   Fol. 110vb1: column min.
   35x55
   iconclass: 25F62("TIGNUS")

5. The Hague, KB, KA 16
   Thonitus (tunny fish)
   Fol. 110vb2: column min.
   30x55
   iconclass: 25F62("TUNNY")

© Koninklijke Bibliotheek National Library of the Netherlands
Interoperability problem in Cultural Heritage

Goal: Simultaneous access to different collections, e.g.,
- The European Library (www.theeuropeanlibrary.org)
- Memory of the Netherlands (www.geheugenvannederland.nl)
- e-culture (e-culture.multimedian.nl)

Difficulties
- Different metadata schemes
- Different thesauri
  - “classical ruins” vs. “landscape with ruins”
  - “the Virgin Mary” vs. “Saint Mary”

However, a universal thesaurus is not favoured, as different thesauri are designed for different domains, applications, etc.

Practical consequence
- searching for “the Virgin Mary” misses “Saint Mary”
Interoperability problem in Cultural Heritage

- **Goal**: Simultaneous access to different collections, *e.g.*,
  - The European Library ([www.theeuropeanlibrary.org](http://www.theeuropeanlibrary.org))
  - Memory of the Netherlands ([www.geheugenvannederland.nl](http://www.geheugenvannederland.nl))
  - e-culture ([e-culture.multimedian.nl](http://e-culture.multimedian.nl))

- **Difficulties**
  - Different metadata schemes
  - Different thesauri
    - “classical ruins” vs. “landscape with ruins”
    - “the Virgin Mary” vs. “Saint Mary”

- However, a universal thesaurus is not favoured, as different thesauri are designed for different domains, applications, etc.

- **Practical consequence**
  - searching for “the Virgin Mary” misses “Saint Mary”
## Interoperability problem in Cultural Heritage

- **Goal**: Simultaneous access to different collections, e.g.,
  - The European Library (www.theeuropeanlibrary.org)
  - Memory of the Netherlands (www.geheugenvannederland.nl)
  - e-culture (e-culture.multimedian.nl)

- **Difficulties**
  - Different metadata schemes
  - Different thesauri
    - “classical ruins” vs. “landscape with ruins”
    - “the Virgin Mary” vs. “Saint Mary”

- However, a universal thesaurus is not favoured, as different thesauri are designed for different domains, applications, etc.

- **Practical consequence**
  - searching for “the Virgin Mary” misses “Saint Mary”
Introduction

Towards interoperability in Cultural Heritage

Interoperability Problems

Interoperability problem in Cultural Heritage

- **Goal:** Simultaneous access to different collections, *e.g.,*
  - The European Library (www.theeuropeanlibrary.org)
  - Memory of the Netherlands (www.geheugenvannederland.nl)
  - e-culture (e-culture.multimedian.nl)

- **Difficulties**
  - Different metadata schemes
  - Different thesauri
    - “classical ruins” vs. “landscape with ruins”
    - “the Virgin Mary” vs. “Saint Mary”

- However, a universal thesaurus is not favoured, as different thesauri are designed for different domains, applications, *etc.*

- **Practical consequence**
  - searching for “the Virgin Mary” misses “Saint Mary”
Interoperability problems
Goal of STITCH
Towards interoperability in Cultural Heritage

Interoperability Problems

Two important steps towards interoperability

- Representing Cultural Heritage vocabularies (thesauri)
  - semantics formally defined, compatible with the Semantic Web
- Thesaurus alignment
  - providing semantic links between thesauri for the accessibility across collections
Cultural Heritage vs. Semantic Web

A simple Semantic Web

- Pointers to resources: documents, knowledge objects, etc.
- Enabling structured assertions
  - *i.e.*, metadata about entities presented on the Web
- Using vocabularies with defined semantics
  - Ontologies: formal definitions of shared conceptual vocabularies
  - RDFS/OWL
Introduction

Towards interoperability in Cultural Heritage

Cultural Heritage and Semantic Web

Similarity between Cultural Heritage and Semantic Web

- Categorising/classifying objects
- Structuring descriptions
- Web-based Approach

Mutual benefits

- Cultural Heritage leverages the advances of the Semantic Web
- Real applications in Cultural Heritage boost the improvements of current Semantic Web techniques
Introduction

Towards interoperability in Cultural Heritage

Cultural Heritage and Semantic Web

Similarity between Cultural Heritage and Semantic Web

- Categorising/classifying objects
- Structuring descriptions
- Web-based Approach

Mutual benefits

- Cultural Heritage leverages the advances of the Semantic Web
- Real applications in Cultural Heritage boost the improvements of current Semantic Web techniques
Towards interoperability in Cultural Heritage

Two main tasks of STITCH

- Porting thesauri to the Semantic Web
- Aligning thesauri
Porting thesauri to the Semantic Web

Thesauri and ontologies: similarities

- Both ontologies and thesauri bring concept hierarchies
- Both give the intended meaning of a vocabulary through links between their items

Correspondences:

- “concept/term” ≈ owl:class
- “broader” ≈ rdfs:subClassOf
- “scope notes” ≈ rdfs:comment
Porting thesauri to the Semantic Web

Thesauri and ontologies: similarities
- Both ontologies and thesauri bring concept hierarchies
- Both give the intended meaning of a vocabulary through links between their items

Correspondences:
- “concept/term” $\approx$ owl:class
- “broader” $\approx$ rdfs:subClassOf
- “scope notes” $\approx$ rdfs:comment
Thesauri and ontologies: differences

Thesauri are designed for humans, without formal interpretations

**Context of: 94M24**

- 9: Classical Mythology and Ancient History
- 94: the Greek heroic legends (I)
- 94M: (story of) Theseus
- 94M2: love-affairs of Theseus
- **94M24**: Perigune, the daughter of Sinis, hides in a thicket of asparagus-plants after Theseus had killed her father; she falls in love with Theseus

How do we interpret a thesaurus in RDFS/OWL?!
Thesauri and ontologies: differences

Thesauri are designed for humans, without formal interpretations

Context of: 94M24

- 9  Classical Mythology and Ancient History
- 94  the Greek heroic legends (I)
- 94M (story of) Theseus
- 94M2 love-affairs of Theseus
- 94M24 Perigune, the daughter of Sinis, hides in a thicket of asparagus-plants after Theseus had killed her father; she falls in love with Theseus

How do we interpret a thesaurus in RDFS/OWL?!
STITCH task1

Representing thesauri using SKOS (Simple Knowledge Organisation System)

- Core model for representing thesauri, classification schemes, subject heading lists, taxonomies, folksonomies, and other types of controlled vocabulary.
- An RDF application in the Cultural Heritage domain
- Within the frame of the Semantic Web
Porting thesauri to the Semantic Web

Example: SKOS building blocks

- skos:ConceptScheme
  - rdf:type
  - http://www.iconclass.nl/
- skos:Concept
  - rdf:type
  - http://www.iconclass.nl/s_11F
- skos:prefLabel
  - "the Virgin Mary"@en
  - "la Vierge Marie"@fr
- skos:broader
  - http://www.iconclass.nl/s_11
- skos:inScheme
  - http://www.iconclass.nl/s_11F
SKOS building blocks

- Classes *Concept* and *ConceptScheme*
- Lexical properties
  - prefLabel
  - altLabel
- Semantic properties
  - broader, narrower
  - related
- Properties for notes and comments
  - scopeNote
  - definition
SKOS

Benefits

- An RDF application in Cultural Heritage domain and within the frame of the Semantic Web
- Enhancing re-usability/interoperability of application components, e.g., browsing, query reformulation

However

- Not everything can be represented in SKOS, e.g., for Iconclass, difficulty to represent all types of auxiliaries
- Ongoing work — see http://www.w3.org/TR/2008/WD-skos-primer-20080221/
SKOS

**Benefits**

- An RDF application in Cultural Heritage domain and within the frame of the Semantic Web
- Enhancing re-usability/interoperability of application components, e.g., browsing, query reformulation

**However**

- Not everything can be represented in SKOS, e.g., for Iconclass, difficulty to represent all types of auxiliaries
- Ongoing work — see [http://www.w3.org/TR/2008/WD-skos-primer-20080221/](http://www.w3.org/TR/2008/WD-skos-primer-20080221/)
STITCH task 2: Thesaurus alignment

Cultural Heritage Interoperability Problem

- Problem: different databases/metadata schemes/vocabularies
- Solution:
  - Syntactically:
    - using common format: XML (RDF)
    - using common vocabulary model (SKOS)
  - Semantically, how do we solve problems caused by conceptual heterogeneity?

Cultural Heritage domain also benefits from techniques developed for the interoperability problem in the Semantic Web, e.g., ontology alignment techniques.
Cultural Heritage Interoperability Problem

- Problem: different databases/metadata schemes/vocabularies
- Solution:
  - Syntactically:
    - using common format: XML (RDF)
    - using common vocabulary model (SKOS)
  - Semantically, how do we solve problems caused by conceptual heterogeneity?

Cultural Heritage domain also benefits from techniques developed for the interoperability problem in the Semantic Web e.g., ontology alignment techniques
Towards interoperability in Cultural Heritage

STITCH task2: Thesaurus alignment

Cultural Heritage Interoperability Problem

- Problem: different databases/metadata schemes/vocabularies
- Solution:
  - Syntactically:
    - using common format: XML (RDF)
    - using common vocabulary model (SKOS)
  - Semantically, how do we solve problems caused by conceptual heterogeneity?

Cultural Heritage domain also benefits from techniques developed for the interoperability problem in the Semantic Web e.g., ontology alignment techniques.
Introduction

Towards interoperability in Cultural Heritage

Summary

STITCH task2: Thesaurus alignment

Cultural Heritage Interoperability Problem

- Problem: different databases/metadata schemes/vocabularies
- Solution:
  - Syntactically:
    - using common format: XML (RDF)
    - using common vocabulary model (SKOS)
  - Semantically, how do we solve problems caused by conceptual heterogeneity?

Cultural Heritage domain also benefits from techniques developed for the interoperability problem in the Semantic Web e.g., ontology alignment techniques
STITCH task2: Thesaurus alignment

- STITCH aims to align thesauri (semi-)automatically, i.e., to find correspondences between thesaurus concepts, e.g.,
  - "Diabetes mellitus" – "suikerziekte"
  - "the Virgin Mary" – "Saint Mary"
- Applying alignment techniques developed in the Semantic Web to the Cultural Heritage domain
  - techniques already investigated there
**Introduction**

Towards interoperability in Cultural Heritage

**Summary**

**Thesaurus alignment**

### Representation of alignments

- **Equivalence/specialisation links for properties and classes**
  - `myVoc:auteur rdfs:subPropertyOf dc:creator`
  - `myVoc:Article owl:equivalentClass yourVoc:Artikel`

- **Identity links between individuals**
  - `vu:swang owl:sameAs kb:ShenghuiWang`

- (yet unstable) SKOS mapping links between subjects
  - `Iconclass:birds skos:exactMatch swd:vogel`
  - `GTT:Cultuur skos:broadMatch Brinkman:cultuurgeschiedenis`
**Representation of alignments**

- Equivalence/specialisation links for properties and classes
  - `myVoc:auteur rdfs:subPropertyOf dc:creator`
  - `myVoc:Article owl:equivalentClass yourVoc:Artikel`

- Identity links between individuals
  - `vu:swang owl:sameAs kb:ShenghuiWang`

- (yet unstable) SKOS mapping links between subjects
  - `Iconclass:birds skos:exactMatch swd:vogel`
  - `GTT:Cultuur skos:broadMatch Brinkman:cultuurgeschiedenis`
Automatic alignment techniques

- **Lexical**
  - labels and textual information of entities

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

- **Extensional**
  - statistical information of instances, *i.e.*, objects indexed with entities

- **Background knowledge**
  - using a shared conceptual reference to find links indirectly
Lexical methods

- Edit distance
- String matching

- Vector space model using textual information of concepts

**Thesaurus alignment**

- GTT:Cultuur
- Brinkman:cultuurgeschiedenis

broader
**Instance-based approach for aligning two thesauri in KB**

- **Scientific collection**: 1.4 M books
- **Depot collection**: 1M books

250K books

**Indexed by**

- **GTT**: (35K concepts)
- **Brinkman**: (5K concepts)
Instance-based approach for aligning two thesauri in KB

Scientific collection
1.4 M books
indexed by GTT
(35K concepts)

250K books
concept alignment?

Depot collection
1M books
indexed by Brinkman
(5K concepts)
Concept mappings from instance similarities

- Directly measuring overlap of instances from 250K dually indexed books
  - Simple similarity measure, e.g.,
    \[
    \text{Jaccard similarity} = \frac{A \cap B}{A \cup B}
    \]

- Some results

<table>
<thead>
<tr>
<th>GTT</th>
<th>Brinkman</th>
</tr>
</thead>
<tbody>
<tr>
<td>“Schilderijen”</td>
<td>“schilderkunst”</td>
</tr>
<tr>
<td>“Kwaliteitszorg”</td>
<td>“kwaliteitsmanagement”</td>
</tr>
<tr>
<td>“Personeelsmanagement”</td>
<td>“personeelsbeleid”</td>
</tr>
<tr>
<td>“Diabetes mellitus”</td>
<td>“suikerziekte”</td>
</tr>
</tbody>
</table>

- Limitation: common instances are necessary, i.e., some book instances are indexed by both thesauri
Concept mappings from instance similarities II

- Predicting concept mappings from the similarity of metadata of individual books, using all books in both collections
  - Assumption: similarity between individuals is informative of similarity between concepts
  - Methods: classification problem using probabilistic learning approach, evolutionary strategy, etc.
  - Limitation: good learning examples and the ground truth are necessary
Problems when using existing techniques

- Thesauri are normally too large for current tools to handle
- Although alignment links can be created somehow, the semantics of those links are not clear

similarity measure → relatedness
→ exactMatch / broadMatch / narrowMatch / ... ?
Introduction
Towards interoperability in Cultural Heritage

Summary

Thesaurus alignment

Problems when using existing techniques

- Thesauri are normally too large for current tools to handle
  - Although alignment links can be created somehow, the semantics of those links are not clear

  similarity measure $\rightarrow$ relatedness
  $\rightarrow$ exactMatch / broadMatch / narrowMatch / ... ?
Problems when using existing techniques

- Thesauri are normally too large for current tools to handle
- Although alignment links can be created somehow, the semantics of those links are not clear

similarity measure $\rightarrow$ relatedness

$\rightarrow$ exactMatch / broadMatch / narrowMatch / ... ?
Problems when using existing techniques

- Thesauri are normally too large for current tools to handle
- Although alignment links can be created somehow, the semantics of those links are not clear

similarity measure $\rightarrow$ relatedness

$\rightarrow$ exactMatch / broadMatch / narrowMatch / ... ?
Towards interoperability in Cultural Heritage

Introduction

Thesaurus alignment

Summary

Problems when using existing techniques

- Thesauri are normally too large for current tools to handle
- Although alignment links can be created somehow, the semantics of those links are not clear

similarity measure $\rightarrow$ relatedness

$\rightarrow$ exactMatch / broadMatch / narrowMatch / ... ?
Problems when using existing techniques

- Thesauri are normally too large for current tools to handle.
- Although alignment links can be created somehow, the semantics of those links are not clear.

similarity measure $\rightarrow$ relatedness

$\rightarrow$ exactMatch / broadMatch / narrowMatch / ... ?
Problems when deploying alignments to real applications

- Different scenarios have different requirements and use alignments in different ways
  - book **retrieval** (concepts used for same books do not necessarily mean the same)
    e.g., GTT: “Opgravingen” – Brinkman: “archeologie; Nederland”
  - book **reindexing** (post-coordination rules)
    e.g., GTT: “Ouderen” + “Sociale relaties” + “Samenlevingsvormen” – Brinkman: “ouderen; maatschappij”
  - thesaurus **merging** (“broadMatch” and “narrowMatch” alignments are still missing from current tools)

...
What can Cultural Heritage offer to the Semantic Web?

- Huge data collections, extremely heterogeneous data sources and versatile applications form a big challenge for Semantic Web techniques,
  - performance, ontology alignment, ...

- A perfect real-world evaluation platform, e.g., OAEI (http://oaei.ontologymatching.org/)
What can Cultural Heritage offer to the Semantic Web?

- Huge data collections, extremely heterogeneous data sources and versatile applications form a big challenge for Semantic Web techniques,
  - performance, ontology alignment, . . .
- A perfect real-world evaluation platform, e.g., OAEI (http://oaei.ontologymatching.org/
Summary

- Cultural Heritage domain leverages the advances of the Semantic Web
  - Representation of collection metadata and thesauri
  - Alignment techniques for interoperability problem
- Cultural Heritage domain conversely provides real applications and an evaluation platform to the Semantic Web community.
Links

- STITCH
  http://www.cs.vu.nl/STITCH/
- SKOS
  http://www.w3.org/TR/2008/WD-skos-primer-20080221/
- OAEI (Ontology Alignment Evaluation Initiative)
  http://oaei.ontologymatching.org/

Related projects
- Museum Finland
  http://www.museosuomi.fi/
- e-culture
  http://e-culture.multimedian.nl/
- The European Library
  http://www.theeuropeanlibrary.org/
- Memory of the Netherlands
  http://www.geheugenvannederland.nl/