Semantic interoperability of data values, use and matching of ontologies and unstructured vocabularies

Jacco van Ossenbruggen
VU/CWI Amsterdam

Semantic interoperability of data values, use and matching of ontologies and unstructured vocabularies

Jacco van Ossenbruggen
VU/CWI Amsterdam
Tearing Down Walls & Building Bridges

Steps towards a Culture Web

Interoperability: tearing down the walls between collections

- Musea have increasingly nice websites
- But: most of them are driven by stand-alone collection databases
- Data is isolated, both syntactically and semantically
- If users can do cross-collection search, the individual collections become more valuable!
The Web: “open” documents and links

The Semantic Web: “open” data and links

Painting
“Green Stripe (M™ Matisse)”
Royal Museum of Fine Arts, Copenhagen

Painter
“Henri Matisse”
Getty ULAN

URL
Web link
URL

creator
Dublin Core
Levels of interoperability

- **Syntactic interoperability**
  - using data formats that you can share
  - XML family is the preferred option
- **Semantic interoperability**
  - How to share meaning / concepts
  - Technology for finding and representing semantic links
Multi-lingual labels for concepts

Prefix ex: <http://www.example.com/concepts#>
Prefix skos: <http://www.w3.org/2004/02/skos/core#>

```
ex:shrubs
  skos:prefLabel 'shrubs'@en
  skos:altLabel 'arbuste'@fr

ex:bushes
  skos:prefLabel 'bushes'@en
  skos:altLabel 'buisson'@fr
```
Important Update Regarding the XML format of the NASA Taxonomy - Jan 9, 2007

The next version of the NASA taxonomy will be in the **SKOS format**.

The SKOS Core is a model and an RDF vocabulary proposed by the W3C for expressing the basic structure and content of concept schemes such as thesauri, classification schemes, subject heading lists, taxonomies, other types of controlled vocabulary.

The SKOS Core Vocabulary is an application of the Resource Description Framework (RDF), that can be used to express a

---

**Principle 1: semantic annotation**

Description of web objects with “concepts” from a shared vocabulary
Principle 2: semantic search

- Search for objects which are linked via concepts (semantic link)
- Use the type of semantic link to provide meaningful presentation of the search results

Term disambiguation is key issue in semantic search

- Post-query
  - Sort search results based on different meanings of the search term
  - Mimics Google-type search
- Pre-query
  - Ask user to disambiguate by displaying list of possible meanings
  - Interface is more complex, but more search functionality can be offered
Principle 3: vocabulary alignment

“Akugawa”

AAT style/period
Edo (Japanese period)
Tokugawa

SVCN period
Edo

AAT is Getty’s
Art & Architecture Thesaurus

SVCN is local in-house
ethnology thesaurus

The myth of a unified vocabulary

- In large virtual collections there are always multiple vocabularies
  - In multiple languages
- Every vocabulary has its own perspective
  - You can’t just merge them
- But you can use vocabularies jointly by defining a limited set of links
  - “Vocabulary alignment”
- It is surprising what you can do with just a few links
Common approaches to ontology matching

- Linguistics & structure
- Shared vocabulary
- Instance-based matching
- Shared background knowledge

Learning alignments

- Learning relations between art styles in AAT and artists in ULAN through NLP of art historic texts
  - “Who are Impressionist painters?”

<table>
<thead>
<tr>
<th>Artist Name</th>
<th>IS</th>
<th>In GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>edgar degas</td>
<td>0.0699</td>
<td>1</td>
</tr>
<tr>
<td>edouard manet</td>
<td>0.0548</td>
<td>1</td>
</tr>
<tr>
<td>pierre-auguste renoir</td>
<td>0.0539</td>
<td>1</td>
</tr>
<tr>
<td>morisot, berthe</td>
<td>0.0393</td>
<td>1</td>
</tr>
<tr>
<td>gogh, vincent van</td>
<td>0.0337</td>
<td>0</td>
</tr>
<tr>
<td>cassatt, mary</td>
<td>0.0318</td>
<td>1</td>
</tr>
<tr>
<td>cezanne, paul</td>
<td>0.0302</td>
<td>1</td>
</tr>
</tbody>
</table>
Thesaurus schema mapping

Meta-Data schema mapping

Meta-Data mapping

Thesaurus alignment

Example textual annotation

<imRecord>
  <im.NUMMER>6</im.NUMMER>
  <im.TITEL>DeIfte Bijbel</im.TITEL>
  <im.TITEL.EN>Delft Bible</im.TITEL.EN>
  <im.MAKER>Renantszoen, Mauritius</im.MAKER>
  <im.OBJECT>tekstbladzijde</im.OBJECT>
  <im.TECHNIEK>boekdruk</im.TECHNIEK>
  <im.DATERING>10 jan. 1477</im.DATERING>
  <im.CLASSIFICATIE>D</im.CLASSIFICATIE>
  <im.ORIGINEEL>Bijbel. Oude Testament</im.ORIGINEEL>
  <im.REPRODUCTIE/>
  <im.TWOND>typografische vormgeving</im.TWOND>
  <im.TWOND>bijbel</im.TWOND>
  <im.TWEDT>Delft</im.TWEDT>
  <im.OMSCHRIFTING>Eerste Bijbel die in het Nederlands verscheen</im.OMSCHRIFTING>
  <im.OMSCHRIFTING.EN>The first Bible to appear in the Dutch language</im.OMSCHRIFTING.EN>
  <im.AFMETINGEN>27 x 20 cm</im.AFMETINGEN>
  ...
</im.Record>
Perspectives

- Openness seems contagious
  - sometimes cannot be open in only a part of your workflow

- Social barriers have to be overcome!
  - “open door” policy
  - Involvement of general public => issues of “quality”
Caveats for museum software

- Be wary of Flash
  - Accessibility
- Make sure you can connect others and other can connect to you
  - “Don’t buy software which does not support standard open API’s”
- Export facilities to common formats (XML, RDF, DC, SKOS, ...)

MoveMyData.org

Because if you can’t move it it’s not really yours

http://movemydata.org/
http://e-culture.multimedian.nl

- Part of the Dutch knowledge-economy project MultimediaN
- Partners: VU, CWI, UvA, DEN, ICN
- People:
  Alia Amin, Lora Aroyo, Mark van Assem, Victor de Boer, Lynda Hardman, Michiel Hildebrand, Laura Hollink, Marco de Niet, Borys Omelayenko, Marie-France van Orsouw, Jacco van Ossenbruggen, Guus Schreiber Jos Taekema, Annemiek Teesing, Anna Tordai, Jan Wielemaker, Bob Wielinga
- Artchive.com, RKD, Rijksmuseum Amsterdam, Dutch ethnology musea (Amsterdam, Leiden), National Library (Bibliopolis)