Knowledge engineering and the Web

Guus Schreiber
VU University Amsterdam
Computer Science, Network Institute
Overview of this talk

• Web data representation
  – a meta view
• Knowledge for the Web: categories
  – key sources, alignment, conversion
• Using knowledge: visualization and search
My journey
knowledge engineering

• design patterns for problem solving
• methodology for knowledge systems
• models of domain knowledge
• ontology engineering
My journey
access to digital heritage
My journey
Web standards

Chair of
• RDF 1.1
• OWL Web Ontology Language 1.0
• SKOS model for publishing vocabularies on the Web
• Deployment & best practices
A few words about Web standardization

• Key success factor!
• Consensus process actually works
  – Some of the time at least
• Public review
  – Taking every comment seriously
• The danger of over-designing
  – Principle of minimal commitment
Example: W3C RDF 1.1 group

- 8K group messages (publicly visible)
- 2K messages about external comments
- 125+ teleconferences
- 200 issues resolved
WEB DATA REPRESENTATION
Caution

• Representation languages are there for you

• And not the other way around ....
HTML5: a leap forward

Rationale

- Consistent separation of content and presentation
- Semantics of the structure of information

Typical new elements

<article>
<section>
<aside>
<header>
<footer>
RDF: triples and graphs

RDF is simply labeling resources and links
RDF: multiple graphs

www.example.org/bob

Alice is a friend of Bob who is a Person born on 14 July 1990. Bob is interested in The Mona Lisa, which was created by Leonardo Da Vinci. The Mona Lisa is about La Joconde à Washington.
RDF syntaxes

• Human-readable: Turtle/TriG
• Line-based: N-Triples/N-Quads
• HTML embedding: RDFa
• JSON-based: JSON-LD
• XML-based: RDF XML
Data modeling on the Web

RDF
- Class hierarchy
- Property hierarchy
- Domain and range restrictions
- Data types

OWL
- Property characteristics
  - E.g., inverse, functional, transitive, …
- Identify management
  - E.g., same as, equivalent class
- ..........

I prefer a pick-and-choose approach
Writing in an ontology language does not make it an ontology!

- Ontology is vehicle for **sharing**
- Papers about your own idiosyncratic “university ontology” should be rejected at conferences
- The quality of an ontology does not depend on the number of, for example, OWL constructs used
SKOS: making existing vocabularies Web accessible

Rationale

• A vocabulary represents distilled knowledge of a community
• Typically product of a consensus process over longer period of time

Use

• 200+ vocabularies published
• E.g.: Library of Congress Subject Headings
• Mainly in library field
The strength of SKOS lies in its simplicity.

Baker et al: Key choices in the design of SKOS
Beware of ontological over-commitment

- We have the understandable tendency to use semantic modeling constructs whenever we can.
- Better is to limit any Web model to the absolute minimum.
KNOWLEDGE ON THE WEB: CATEGORIES
The concept triad

Musical Instrument
(identifies)

any device capable of tuneful, melodious, harmonious sounds
(definition)

Musical Instrument

Concept

(instance)

Categorization

• OWL (Description logic) takes an extensional view of classes
  – A set is completely defined by its members
• This puts the emphasis on specifying class **boundaries**
• Work of Rosch et al. takes a different view
Categories (Rosch)

• Help us to organize the world
• Tools for perception
• Basic-level categories
  – Are the prime categories used by people
  – Have the highest number of common and distinctive attributes
  – What those basic-level categories are may depend on context
## Basic-level categories

<table>
<thead>
<tr>
<th>Superordinate</th>
<th>Basic Level</th>
<th>Subordinate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Furniture</td>
<td>Chair</td>
<td>Kitchen chair</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Living-room chair</td>
</tr>
<tr>
<td>Table</td>
<td></td>
<td>Kitchen table</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Dining-room table</td>
</tr>
<tr>
<td>Lamp</td>
<td></td>
<td>Floor lamp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Desk lamp</td>
</tr>
<tr>
<td>Tree</td>
<td>Oak</td>
<td>White oak</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Red oak</td>
</tr>
<tr>
<td>Maple</td>
<td>Silver maple</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sugar maple</td>
<td></td>
</tr>
<tr>
<td>Birch</td>
<td>River birch</td>
<td></td>
</tr>
<tr>
<td></td>
<td>White birch</td>
<td></td>
</tr>
</tbody>
</table>
FOAF: Friend of a Friend
Dublin Core: metadata of Web resources
Iconclass categorizing image scene

ICONCLASS Subject Classification

Search using: "baring the head, lifting one's hat" (it has 101 hits)

Show Kw | SPACES | Hide Codes | de | fr | it | fi | Top |

3 Human Being, Man in General
33 relations between individual persons
33A non-aggressive relationships
33A1 saluting
33A11 baring the head, lifting one's hat
schema.org categories for TV programs
schema.org issues

• Top-down versus bottom-up
• Ownership and control
• Who can update/extend?
• Does use for general search bias the vocabulary?
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”
Category alignment vs. identity disambiguation

- Alignment concerns finding links between (similar) categories, which typically have no identity in the real world.
- Identity disambiguation is finding out whether two or more IDs point to the same object in the real world (e.g., person, building, ship).
- The distinction is more subtle than “class versus instance.”
Alignment techniques

• Syntax: comparison of characters of the terms
  – Measures of syntactic distance
  – Language processing
    • E.g. Tokenization, single/plural,

• Relate to lexical resource
  – Relate terms to place in WordNet hierarchy

• Taxonomy comparison
  – Look for common parents/children in taxonomy

• Instance based mapping
  – Two classes are similar if their instances are similar.
Alignment evaluation

Please check this mapping:
(still 44 concepts to go) Save intermediate results
detail panel
http://e-culture.multimedian.nl/ns/rkd.thesaurus.subject#autobus
[3639] autobus
Example usage:

[Image of example usage]

detail panel
http://e-culture.multimedian.nl/ns/corneotto/wordnet/d_n-10563

(target concept 1 out of 1 mappings:
   Approve  Reject exact match
   Should be broader  narrower  related
   I'm not sure
   apparaat
   inrichting
   differentieel
   motorrijtuig
   autobus

[Image of detailed panel]
Limitations of categorical thinking

Preservationists "Those engaged in making changes in an object to prevent further deterioration."

Conservators "Those engaged in preventive care for long term safe-keeping"

Restorers "Those engaged in changes to an object that will approximate original state"

Renovator

Refinisher

AAT concept(s)

WordNet concept
Be modest! Don’t recreate, but enrich and align

- Knowledge engineers should refrain from developing their own idiosyncratic ontologies
- Instead, they should make the available rich vocabularies, thesauri and databases available in an interoperable (web) format
- Techniques: learning, alignment
Europeana ThoughtLab

The Amsterdam Museum Linked Open Data set is now a part of the Europeana ThoughtLab, a showcase of the technologies developed by the Europeana partners. The Amsterdam Museum Linked Open Data is a prime example of Linked Open Data of the type that Europeana will be producing.

More info at: www.europeana.eu/portal/thoughtlab_linkedopendata.html

Apps for Amsterdam

AM as LOD is one of the datasets available for the Apps for Amsterdam challenge. Apps for Amsterdam is an initiative to open up and use data about the city of Amsterdam. The apps-design competition runs up until 4 may 2011.

More info at: www.appsforamsterdam.nl/

Dataset added to CKAN

Victor. 17 Feb 2011

About AM data in EDM/RDF

The Amsterdam Museum (AM) data was retrieved from an XML-dump of the museum’s Adlib collection database. The data was converted to RDF compliant with the Europeana Data Model [pdf] (EDM). The AHM EDM data is published as Linked Open Data. The conversion software is available as open source. Europeana Data Model [pdf]
Issues in conversion to Linked Data

• Keep original info as much as possible
• Ensure standard datatypes for value conversion
• Choice of concept URL

• Vocabulary URL: our strategy
  – Large organizations: use their own
  – Small organizations: purl.org URL
Selecteer een fragment om het spel mee te starten

De tiet zal uit útwiezen
29 keer gespeeld
hoogste score: 2.915

Burgemeester viekken w
2 keer gespeeld
hoogste score: 35

Een week later
4 keer gespeeld
hoogste score: 475

Wat sleept u met u mee?
0 keer gespeeld
hoogste score: 0

Spul wat ze inspuiten
0 keer gespeeld
hoogste score: 0

Hoe werkt dit?
Selecteer een video en probeer zoveel mogelijk woorden in te voeren die beschrijven wat u ziet of hoort. Als anderen binnen tien seconden hetzelfde woord invoeren heeft u een ‘match’. Matches leveren punten op. U kunt hierboven een fragment selecteren, of één uit de wachtlijst en direct beginnen met spelen. Om uw punten te kunnen bewaren dient u geregistreerd en aangemeld te zijn. Om mee te kunnen doen moet u de Microsoft Silverlight plugin geïnstalleerd hebben.

Win leuke prijzen!
Deelnemers
1 Michiel 2 bikorr

Hoe houdt u de stemming erin? nog 00:15

510 1 / 1
Uw ingevulde woorden:
glasses +5
bebbeltbox +5

515 1 / 2
Uw ingevulde tags:
- jenever +5
- head +150 match met head van bikorr. Jij hebt dit woord geïntroduceerd
- man +5
- woman +5
- beard +150 match met beard van bikorr. Jij hebt dit woord geïntroduceerd
- meeting +5
- bar +25 geografische naam
- wine +150

Puntentotaal
515

8 tags zonder match? 40

1 geografische naam? 25

0 persoonsnamen? 0

3 matches met medespelers? 150

3 pioniersmatches? 300

Kijk hiernaast in het overzicht van 'Uw ingevoerde tags' voor een specifieker overzicht per match.

Meer punten verdienen?
Daag vrienden uit om dit spel te spelen en vergroot uw eigen kans op een hogere score?
Improve video search for (1) fragment retrieval & (2) within video navigation by using crowdsourcing for (1) including time-based annotations & (2) bridging the vocabulary gap of searcher & cataloguer
Nichesourcing: finding the right minority vote

“the frog stands for the rejected lover”

“The frog indicates the profession of the women”

“women”

“Sexy ladies!”
USING KNOWLEDGE:
VISUALIZATION AND SEARCH
Mobile museum tour
Visualising piracy events
Extracting piracy events from piracy reports & Web sources

2010-10-23

23.10.2010: 1235 UTC:
Posn 04:14.05 – 041:19.0E
Around 98 nm east of Mombasa, Kenya, Off Southern Somalia.
Armed pirates attacked and hijacked a LPG tanker underway. Further details awaited.
Enriching description of search results

**Woman with a Fan**

[Image of Picasso's painting]

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creator</td>
<td>Picasso, Pablo</td>
</tr>
<tr>
<td>Location</td>
<td>Paris; The Hermitage Museum, St. Petersburg</td>
</tr>
<tr>
<td>Material</td>
<td>canvas; oil paint</td>
</tr>
<tr>
<td>Measurements</td>
<td>152 x 101 cm</td>
</tr>
<tr>
<td>Title</td>
<td>Woman with a Fan</td>
</tr>
</tbody>
</table>
Sample graph search algorithm

From search term (literal) to art work
• Find resources with matching label
• Find path from resource to art work
  – Cost of each step (step when above cost threshold)
  – Special treatment of semantics: sameAs, inverseOf, transitive, ....
• Cluster results based on path similarities
Graph search

- Works in museum (2)
  - Self-Portrait
    - Miró, Joan
  - Portrait of a Spanish
    - Miró, Joan

- Works created by (92)
  - Rembrandtesque
    - Picasso, Pablo
  - Reservoir at Horta
    - Picasso, Pablo
  - Still Life with
    - Picasso, Pablo
  - Glass, Dice, and
    - Picasso, Pablo
  - Seated Old Man
    - Picasso, Pablo

- Works with style/period Surrealist also used by artist (1)

- Works with style/period Cubist also used by artist (1)

- Works by professionally related artist (31)
  - Fruit Dish, Ace of Clubs
    - Braque, Georges
  - Man with a Violin
    - Braque, Georges
  - Bottle, Newspaper,
    - Braque, Georges
  - Still Life BACH
    - Braque, Georges
  - Black Fish
    - Braque, Georges
Example of path clustering

Issues:
- number of clusters
- path length
Using alignment in search

“Tokugawa”

AAT style/period
Edo (Japanese period)
Tokugawa

AAT is Getty’s
Art & Architecture Thesaurus

SVCN period
Edo

SVCN is local in-house ethnology thesaurus
Location-based search: Moulin de la Galette

relatively easy
Relation search: Picasso, Matisse & Braque
This is a research prototype of Europeana's semantic search engine. Enter a search term, for example: Egypt, Rembrandt, window.

Collections

<table>
<thead>
<tr>
<th>Collection</th>
<th>Artworks</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rijksmuseum</td>
<td>46,038</td>
</tr>
<tr>
<td>RKD</td>
<td>82,781</td>
</tr>
<tr>
<td>Louvre</td>
<td>11,327</td>
</tr>
</tbody>
</table>
Acknowledgements

• Long list of people
• Projects: COMMIT, Agora, PrestoPrime, EuropeanaConnect, Poseidon, BiographyNet, Multimedian E-Culture