Proposals for

principles of knowledge engineering

In the 21st century

Guus Schreiber
VU University Amsterdam
Knowledge engineering in the 20th century

- Closed systems
- Growing importance of knowledge patterns
  - Focus on patterns of problem-solving tasks
- The great divide between knowledge-engineering and knowledge-representation communities
- Protégé is prime descendant of KAW breeding ground of knowledge-engineering research
Knowledge engineering in the 21\textsuperscript{st} century

• Open Web systems
• Rich availability of (new) knowledge sources
• New programming paradigms
• Ontologies have become “en vogue”
Knowledge engineering and the Semantic Web Project

• The Semantic Web is not a research discipline, but an application domain

• Knowledge-engineering research has been and still is a key driver for the Semantic Web Project

• Knowledge engineering flourishes through the multi-disciplinary cooperation within the Semantic Web Project
Hypothesis

- Semantic Web technology is in particular useful in knowledge-rich domains

or formulated differently

- If we cannot show added value in knowledge-rich domains, then it may have no value at all
This talk

Can we formulate principles for knowledge engineering in the 21st century?

Knowledge-engineering case study: Distributed heritage collections
This is a research prototype of Europeana's semantic search engine. Enter a search term, for example: Egypt, Rembrandt, window.

Collections

- Rijksmuseum: 46,038 artworks
- RKD: 82,781 artworks
- Louvre: 11,327 artworks

powered by ClioPatria 1.0 beta 2 (18/12/2008)
SWI Prolog 5.7.10-28-gaf37c11
eculture.cs.vu.nl
The Web: resources and links

URL → Web link → URL
The Semantic Web: typed resources and links

Painting
“Woman with hat
SFOMOA

Dublin Core
creator

ULAN
Henri Matisse

URL
Web link
URL
Related People or Corporate Bodies:

- apprentice was Jolin, Einar 1911-1913
  ----------------------------- (Swedish painter, 1890-1990) [500014093]
- parent of Duthuit, Marguerite Matisse
  ----------------------------- (French painter, born ca. 1900) [500075813]
- patron was Barnes, Dr. Albert C.
  ----------------------------- (American collector, 1872-1951) [500057478]
- student of Cormon, Fernand
  ----------------------------- (French painter and teacher, 1845-1924) [500115385]
- student of Moreau, Gustave
  ----------------------------- (French painter, 1826-1898) [500115776]

Roles:

- artist (preferred)
- painter
- printmaker
- sculptor
- designer
- writer

Gender: male

Birth and Death Places:

- Born: Le Cateau-Cambrésis (Nord, Nord-Pas-de-Calais, France) (inhabited place)
- Died: Nice (Alpes-Maritimes, Provence-Alpes-Côte d'Azur, France) (inhabited place)
Vernacular Display | English Display

Click the icon to view the hierarchy.
Check the boxes to view multiple records at once.

- Top of the TGN hierarchy (hierarchy root)
- World (facet)
- Europe (continent)
- Netherlands (nation)
- [view physical features]
- Aarkanal (canal)
- Afsluitdijk (dam)
- Alblasserwaard (general region)
- Altena, Land van (general region)
- Amstelland (general region)
- Amsterdam Rijn Kanaal (canal)
- Aruba (dependent state) [N]
- Bernisse Molen (mill center)
- Biesbos (general region)
- Brouwersdam (dam)
- Calandkanaal (canal)
- Delfland (general region)
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
Power of (simple and partial) vocabulary alignments

“Tokugawa”

AAT style/period
Edo (Japanese period)
Tokugawa

AAT is Getty’s
Art & Architecture Thesaurus

SV CN period
Edo

SV CN is local in-house ethnology thesaurus

AAT is Getty’s Art & Architecture Thesaurus
Knowledge engineering activities for distributed heritage collections

Vocabulary interoperability
Vocabulary alignment
Metadata schema interoperability
Metadata enrichment

Semantic search
Semantic annotation
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
Vocabulary interoperability:
an ad for SKOS

Simple Knowledge Organisation System (SKOS)

SKOS Core | SKOS Mapping | SKOS Extensions

This page: Specifications | RDF Vocabularies | Development


SKOS is an area of work developing specifications and standards to support the use of knowledge organisation systems (KOS) such as thesauri, classification schemes, subject heading lists, taxonomies, other types of controlled vocabulary, and perhaps also terminologies and glossaries, within the framework of the Semantic Web.

There are three RDF vocabularies under active development: SKOS Core | SKOS Mapping | SKOS Extensions. There is also the SKOS API, a web service API for interacting with a KOS data source.

SKOS Specification Development

The following specifications are under development within the W3C Semantic Web Best Practices and Deployment Working Group:

- SKOS Core Guide

  This document is a guide using the SKOS Core Vocabulary, for readers who already have a basic understanding of RDF concepts. It is the authoritative guide to recommended usage of the SKOS Core Vocabulary at the time of publication.

- SKOS Core Vocabulary Specification

  This document gives a reference-style overview of the SKOS Core Vocabulary as it stands at the time of publication. It is the authoritative human-readable account of the SKOS Core Vocabulary at the time of publication. It also describes the policies for ownership, naming, versioning and change to which the SKOS Core Vocabulary is organized.
Multi-lingual labels for concepts

prefix ex: <http://www.example.com/concepts#>
prefix skos: <http://www.w3.org/2004/02/skos/core#>
Semantic relation: broader and narrower

- No subclass semantics assumed!
Issues in specification of SKOS semantics

• SKOS should cover a large range of “vocabularies”, “thesauri”, “terminologies”, “classification schemes”, etc.
• Therefore: objective was to define the minimal semantics
• Leave hooks for specializations
• See SKOS Primer for examples
skos:exactMatch is disjoint with each of the properties skos:broadMatch and skos:closeMatch.

Characteristics: exact
- Functional
- Inverse functional
- Transitive
- Symmetric
- Asymmetric
- Reflexive
- Irreflexive

Description: exactMatch
- Domains (intersection)
- Ranges (intersection)
- Equivalent object properties
- Super properties
- Inverse properties
Example requirement

• Being able to define relations between labels
  – “WHO” is an acronym of “World Health Orgnization” (in English)
  – “WGO” is an acronym of “Wereldgezonheidsorganisatie” (in Dutch)

• Treat lexical labels as resources with URI?
  – But many simple vocabularies don't needs this
  – Would be burden
Large organizations have adopted SKOS

How it works

Users and machines simply request the URI of interest over HTTP. For example, to access the data value "World Wide Web" in the Library of Congress Subject Headings, one would request this URI:

▷ http://id.loc.gov/authorities/sh95000541#concept

When requesting this URI, users have mechanisms for specifying how they want to serialize the data they wish to access. These include common RDF serializations carrying Simple Knowledge Organization System (SKOS) metadata, and Javascript Object Notation (JSON).

See the Technical Center for more details.

Benefits

For users (whether human or machine):

▷ Access to data at no cost.
▷ Granular access to individual data values.
▷ Ability to download entire controlled vocabularies and the values within them in numerous formats.
Metadata schema interoperability

• Cultural heritage has an abundance of metadata format standards
  – Dublin Core, VRA (images), MARC, ....

• Current practice: XSLT transformations (and similar)

• `owl:EquivalentProperty` and `rdfs:subPropertyOf` are well suited for defining partial alignments between schemata
Aligning VRA with Dublin Core

• VRA is specialization of Dublin Core for visual resources
• VRA properties “material.medium” and “material.support” are specializations of Dublin Core property “format”

vra:material.medium  
rdfs:subPropertyOf dc:format .

vra:material.support  
rdfs:subPropertyOf dc:format .
Strong point of OWL

“For collection X the range of dc:creator is a value from the ULAN thesaurus”

=> Define an owl:Restriction for resources in X which specifies a corresponding local range restriction for the dc:creator value
Built-in overcommitment in OWL DL

Is `dc:creator` an `owl:DatatypeProperty` or an `owl:ObjectProperty`?

Answer: depends on the context!

The minimal commitment is:

`dc:creator rdf:type rdf:Property`.
Metadata enrichment

<inn:Record>
  <inn:NUMMER>6</inn:NUMMER>
  <inn:TITEL>Delftse Bijbel...</inn:TITEL>
  <inn:TITEL_EN>Delft Bible...</inn:TITEL_EN>
  <inn:MAKER>Yemantszoon, Mauricius : d</inn:MAKER>
  <inn:OBJECT>tekstbladzijde</inn:OBJECT>
  <inn:TECHNIEK>boekdruk</inn:TECHNIEK>
  <inn:DATERING>10 jan. 1477</inn:DATERING>
  <inn:CLASSIFICATIE>D</inn:CLASSIFICATIE>
  <inn:ORIGINEEL>Bijbel. Oude Testament...</inn:ORIGINEEL>
  <inn:REPRODUCTIE>
  <inn:TWOND>typografische vormgeving</inn:TWOND>
  <inn:TWOND>bijbels</inn:TWOND>
  <inn:TWGEO>Delft</inn:TWGEO>
  <inn:OMSCHRIJVING>Eerste bijbel die in het Nederlands verscheen...</inn:OMSCHRIJVING>
  <inn:OMSCHRIJVING_EN>The first Bible to appear in the Dutch language...</inn:OMSCHRIJVING_EN>
  <inn:AFMETINGEN>27 x 20 cm</inn:AFMETINGEN>
  ...
</inn:Record>
Replace strings with concepts:
quality issues of automatic extraction
Hot issue: event modelling
“what is happening on an image?”
Vocabulary alignment

• 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</th>
<th>GS</th>
</tr>
</thead>
<tbody>
<tr>
<td>edgar degas</td>
<td>0.0699</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>edouard manet</td>
<td>0.0548</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>pierre-auguste renoir</td>
<td>0.0539</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>morisot, berthe</td>
<td>0.0393</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>gogh, vincent van</td>
<td>0.0337</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>cassatt, mary</td>
<td>0.0318</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>cezanne, paul</td>
<td>0.0302</td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>
Results of automatic alignment vary in quality

![Venn diagram showing the overlap of Baseline, Lexical, and Falcon in aligning semantic alignments.]

- Baseline: 12 (100%)
- Lexical: 647 (35%)
- Falcon: 244 (44%)
- Intersection: 77 (93%)
- Intersection: 511 (59%)
- Intersection: 94 (96%)
- Total: 2765 (59%)
Partial human engineering and/or evaluation is often time/cost effective
Semantic search: clustering and cluster-order principles
Research topic: semantic patterns which increase recall without sacrificing precision
Semantic annotation: granularity level

<table>
<thead>
<tr>
<th>Who</th>
<th>Historical persons</th>
</tr>
</thead>
<tbody>
<tr>
<td>person</td>
<td></td>
</tr>
<tr>
<td>Oldenbarnevelt, Johan van</td>
<td>x</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>What</th>
<th>Iconclass (en), WordNet (en), events (nl)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(mythological) concept, object or event</td>
<td>beheading</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Where</th>
<th>Name of place or region</th>
</tr>
</thead>
<tbody>
<tr>
<td>geographical place</td>
<td>Den Haag</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>When</th>
<th>Date, year or period</th>
</tr>
</thead>
<tbody>
<tr>
<td>enter date</td>
<td></td>
</tr>
</tbody>
</table>

Print with a scene

update  cancel

done | cancel
Autocompletion and disambiguation issues
Principles for knowledge engineering on the Web
Principle 1: Be modest!

- Ontology 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.
- Initially, only add the originally intended semantics.
Principle 2: Think large!

Doug Lenat

"Once you have a truly massive amount of information integrated as knowledge, then the human-software system will be superhuman, in the same sense that mankind with writing is superhuman compared to mankind before writing."
Principle 3: Develop and use patterns!

- Don’t try to be (too) creative
- Ontology engineering should not be an art but a discipline
- Patterns play a key role in methodology for ontology engineering
- See for example patterns developed by the W3C Semantic Web Best Practices group http://www.w3.org/2001/sw/BridgePractices/
- SKOS can also be considered a pattern
Principle 4: Don’t recreate, but enrich and align

• Techniques:
  – Learning ontology relations/mappings
  – Semantic analysis, e.g. OntoClean
  – Processing of scope notes in thesauri
  – Manual evaluation sometimes key
Principle 5: Beware of ontological over-commitment!
Principle 6: Specifying a data model in OWL does not make it an ontology!

- Papers about your own idiosyncratic “university ontology” should be rejected at conferences
- The quality of an ontology does not depend on the number of OWL constructs used
Principle 7: Required level of formal semantics depends on the domain!

• In our semantic search we use three OWL constructs:
  – owl:sameAs, owl:TransitiveProperty, owl:SymmetricProperty

• But cultural heritage has is very different from medicine and bioinformatics
  – Don’t over-generalize on requirements for e.g. OWL
Thank you!

*Acknowledgments*: slides and ideas from many co-workers within VU, Amsterdam and KE and SW communities, in particular Lora Aroyo, Michiel Hildebrand, Antoine IsaacJacco van Ossenbruggen, Anna Tordai, Jan Wielemaker.