Reconciling Vocabularies on the Semantic Web

Amy Lucker
Institute of Fine Arts, New York University
Meaning is expressed via heterogeneous sets of terms, some controlled, some not. If the terms don’t relate each to the other, then there is no way for machines to collocate or differentiate information. One of the main features of the semantic web is that connections are made automatically, machines make inferences and assumptions but without structured information it will be meaningless.
Authority records/vocab systems can be national, international, local, general, subject specific
Examples include: Subject (topical): e.g., LCSH, AAT; Form/genre: AAT, TGM; Personal names: LCNAF, ULAN; Geographic names: TGN, LCSH; Uniform titles: LC; Corporate names: LCNAF, VIAF vocabularies, keywords, and thesauri: keywords can be controlled or not; vocabulary v thesaurus: has to do with whether or not relationships between terms exist. If not, basic vocabulary (esp. in local systems). Still "controlled vocabulary" though. Thesaurus includes relationships, which can be hierarchical (AAT, MeSH) or not (LCSH). Easy to do your own controlled vocab list, not easy to do your own thesaurus, yet the thesaurus if used well can be extremely helpful for users.
Benefits of controlled vocabularies

- Include:
  - Collocation
  - Differentiation
  - Precision and Relevance or Quality vs. Quantity

Collocation: pull together items which on the surface might not discuss same topic. Where keyword alone wouldn’t join them. Also allows for pulling together items on conceptual basis; requires analysis of contents.
Differentiation: deal with homonyms (bear/bear), synonyms (see also concatenation), variant spellings (especially names, e.g., Raphael in ULAN).
Quantity v quality, or the problem with Google
“Alternate Labels” = “see” references
Broader/narrower terms also part of syndetic structure of vocab.
Visual representation of the syndetic structure – seen in earlier presentations – here in it’s most simple form
Source code indicates some of the limitations of LCSH, e.g., language (all English terms); LCSH terms are often expressed in pre-coordinated strings. LC now committed to using SKOS for presentation of its vocabularies: “The data model used by the Authorities and Vocabularies service is influenced entirely by SKOS. As a result, most of the metadata associated with our offerings are disseminated using properties and classes from SKOS.” Allows for dissemination and use of LC systems in wider variety of ways, and for better integration into the (eventual) semantic web.
 Getty Vocabularies

What is Getty Vocabularies? Getty Vocabularies contain structured terminology for art, architecture, decorative arts, and other cultural materials, visual art, and bibliographic materials. Consistent with international standards, Getty vocabularies provide authoritative information for art historical description and can be used to enhance access to databases and object sites. Getty Vocabularies are open access resources.

The Getty Thesaurus of Geographic Names (TGN)

TGN is a structured vocabulary, including names, geographies, and other information for places important to art and architecture.

The Union List of Artist Names (ULAN)

ULAN is a structured vocabulary, including names, biographies, and other information for artists important to art and architecture.

The Art & Architecture Thesaurus (AAT)

AAT is a structured vocabulary, including terms, descriptions, and other information for generic concepts related to art and architecture.

The Cultural Objects Name Authority (CONA)
AAT – better inclusion of other languages
Variety of languages, sources broader than LC. All terms like LC have id number. BUT … All Getty vocabs are in strict hierarchies. … not expressed in SKOS format, relies on more descriptive information than LC, so harder in some ways to integrate into S.W. Doesn’t use pre-coordination though, so no strings to deal with.
The web does not contain a dictionary that translates all languages into all other languages; terms in one language do not necessarily relate to specific terms in another, so even if such a dictionary could be constructed the semantic part — the subtle (or not so subtle) meanings embedded in things like subject terms — would be difficult for a machine to create. It’s not enough to indicate that word A = word B, rather that IDEA A = IDEA B, and ideas are the realm of humans, not machines.

Post- v. pre-coordination — again, the variety of ways in which terms can be combined is infinite. But pre-coordinated terms do not translate well. Keywords are harder to control, but easier to translate.

Machines are not able to translate the broader/narrower/use for/used for relationships without specific instruction, even within a single language. Again these relationships are those of ideas and concepts, not just words.

Faceted hierarchies are more flexible, which makes them easier to work with although again allow for far more interpretations than do rigid hierarchies. Very difficult to match up one rigid hierarchy with another.

Moving everything to the “lowest common denominator”, e.g., Dublin Core puts everything on the same plane, but loses things like syndetic relationships.

---

So what’s the problem?

- INTEROPERABILITY of:
  - Languages
  - Post v. Pre-coordination
  - Strings v. Keywords
  - Maintaining syndetic relationships
  - Faceted v. rigid hierarchies
Must deal with 2 issues: structural (e.g., term v. class, hierarchy v. flat file, XML v. text, etc.) and conceptual (same ideas, different terms or same terms, different ideas). Can approach with:
Alignment: creating equivalencies between concepts, including “broader”, “narrower”. Requires subject expertise, and is a manual process; could start to automate if they are structurally similar. Could also automate based on finding terms that all relate to the same object, and then aligning the terms – done with project in Amsterdam (ARIA).
Translation: literally translating one term to another to another – in effect what has been done with VIAF, and what you see in ULAN and other Getty vocabs.
Cross-walk: similar to alignment, create document which indicates that A = B = C. The Getty Metadata Standards Crosswalk is a good example of this. Again, very labor intensive, and better for things like structural schema and not broad-based hierarchies like vocabularies.
Harvesting: using some/all of the above techniques, harvest source data and output it dynamically into other formats, e.g., SKOS.
In the end, no one technique or tool is going to solve all the problems, but rather different approaches will need to be combined.
Compare to ULAN – again multiple languages in essence translated. VIAF however broader even than ULAN in approach and is a coordinated collaborative effort:

“VIAF, implemented and hosted by OCLC, is joint project of several national libraries plus selected regional and trans-national library agencies. The project’s goal is to lower the cost and increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web.” from the website. Now, names and work titles; presumably will coordinate with CONA for work titles, since it includes Getty Research Institute.
Very visual, shows who has “mapped” or translated themselves to the others.
Wikipedia-ish, more grass-roots than VIAF. But also prospective, not retrospective, and less useful for cultural heritage description. Again, names, not concepts.
Concept matching/linking
Integrated access to BnF Mandragore and KB Illuminated Manuscripts Collection

Introduction

Starting 2019, STITCH has collaborated with the French National Library to build a demonstration web browser that would illustrate how semantic ontologies and parallel description vocabularies can help integrating collections coming from different backgrounds.

The object of this experiment consisted of two collections of illuminated manuscripts:

- the Mandragore database from the Manuscript Department at the French National Library (BnF)
- the Medieval Illuminated Manuscripts at the National Library of the Netherlands (Koninklijke Bibliotheek)

Disclaimer: This prototype that is presented here is the result of preliminary investigations. Especially, the semantic correspondences it relies on have not yet been validated and may require revision. This work is still ongoing research, and shall not be seen as official access means to the collections.

Help

System URL

The illumination collections of both BnF and KB can be accessed from their respective portal using two options:
### Collection

<table>
<thead>
<tr>
<th>Sous-collection</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Musée de la civilisation</td>
<td>Patrimoine de la civilisation française (1750)</td>
</tr>
</tbody>
</table>

### Collection de la bibliothèque nationale de France (BNF)

<table>
<thead>
<tr>
<th>Local</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOCAL (14)</td>
<td>ARTS (10) BOTANIQUE (10) CHIMIE (10) ÉCONOMIE (10) GEOGRAPHIE (10) INFORMIČATION (10) LITTÉRATURE (10) MATHEMATIQUES (10) PHILOSOPHIE (10) PHYSIQUE (10) PSYCHOLOGIE (10) SCIENCE (10) TECHNOLOGIE (10) TOPOGRAPHIE (10) VÉHICULES (10)</td>
</tr>
</tbody>
</table>