Links, Meaning, and Contexts: Making Sense & Using Logic

Michael Buckland

International UDC Consortium Seminar: Classification & Authority Control: Expanding Resource Discovery
Lisbon, 29 October 2015
(089.7)“329.302”
"Classification & Authority Control: Expanding Resource Discovery.”
http://seminar.udcc.org/2015/programme.php

Linked data practices and techniques have opened new possibilities in exploiting controlled vocabularies and improving resource discovery.

Authority data held in library systems often includes classification schemes. These knowledge structures now have the potential for being shared across the linked data environment.

The objective of this conference is to explore such potential, expanding the value and use of classification as an authority controlled vocabulary, from a local perspective to the global environment.

What interests me – as problems or opportunities.

**Group 1** entities are defined as the products of intellectual or artistic endeavours that are named or described in bibliographic records: *work, expression, manifestation*, and *item*. [= DOCUMENT]

**Group 2** entities are those responsible for the intellectual or artistic content, the physical production and dissemination, or the custodianship of the Group 1 entities: *person, corporate body, and family*. [= CREATOR]

**Group 3** entities represent an additional set of entities that serve as the subjects of *works: concept, object, event, and place*. [= TOPIC]

---

**Functional Requirements for Subject Authority Data (FRSAD)**, 2010.

THEMA [= Topic: e.g. Physical object, conceptual entity, event]

NOMEN [= Name of topic: Subject heading, classification no., code, etc.]

Questionable: Readers interpret *items*. Manifestations interpret *Expressions*. Creator’s intention not always known. But *Topics* assigned to *Works*, not to *Manifestations* or *Items*. 
MULTIPLE RELATIONSHIPS AMONG THEMAS AND NOMENS

THEMAS = Topics (What a Work is about and/or of), e.g. Physical object, conceptual entity, event; i.e. anything sensed, perceived, imagined [= PHENOMENA].

NOMENS = Names of topics: Subject headings, Classification numbers, Ontology units, Category codes, Keywords, Tags, etc.

NOMENS are names (nominations), hence language acts. Languages are largely composed of names that are related.

VOCABULARY = a set of names, is sometimes controlled for Preferred forms and/or Semantics: equivalence (synonyms), inclusion (hierarchy), other relationships (see also).

Linking NOMENS in different languages (VOCABULARIES) is “mapping”.
LINKS, CONTEXTS, LOGIC

Links between names in different languages are necessarily links between names in different contexts.

Links express relationships
-- Links are logical statements
-- But many relationships are not logical ("material relations").

SYMPOSIUM THEME WITH TWO ASPECTS
-- Performance: How best to combine links and vocabularies for resource description and discovery.
-- Exploratory: What can be said about relationships between phenomena, names, and links?
  What are the limits to linking?
  Can we cross (or change) these limits in productive ways?
MY ASSUMPTIONS

1. Learning, knowing, and understanding constitute how we live, so Documentation (by whatever name) is a form of cultural engagement.

2. Documentary systems are full of links of many kinds, including subject indexes, syndetic structures, search term recommender services, query-to-retrieved set relations, as well as “linked data” in sense of Linked Open Data. Any relationship potentially has a link.

3. There is a tension between logic (system) and language (names), between (hyper)rationality and making sense (reasoning).

4. How to combine the expressive power of language, the cultural complexity of our environment, and use of hyper-rational tools?

5. Probabilistic methods are useful in a complex, unstable world.

6. Where the limits? Limits are challenges and opportunities.

7. What does all this signify for our field?
HOMMAGE TO PAUL OTLET (1868-1944)

1892: Collective action for “the creation of a kind of artificial brain by means of cards containing actual information or simply notes of references”. “... a careful arrangement of its nomenclature ... would thus permit the creation of very practical links.”

In contrast: LUDWIK FLECK (1896-1961): Local cultural context is important for sense and understanding:
- Writer, text, and author’s habits / culture.
- Reader, text, and reader’s habits / culture.
- Differences in habits / culture hinder understanding.

We each live in a “small world” (Elfreda Chatman), in the “World of Where and When” (Stephen Toulmin).
Fleck’s insistence on the uniqueness of local personal contexts means that convenient formal relationships across contexts are not reliable. This is subversive of Otlet’s modernist, global vision.

In Documentation:
Large collections serve heterogeneous users – and include diverse materials from specialized sub-domains.
Therefore, a single vocabulary (SKOS, classification) designed for the entire collection will not be the best for many (?most) users -- or for all material.
In a pre-digital environment there was no other possibility, but now . . . ?
NAMES: UNFAMILIAR VOCABULARIES (from outside our small world!)
*Hand-to-hand fighting, oriental, in motion pictures.* (Former LCSH for Kung Fu films).

*HS 847120: Digital auto data proc mach contng in the same housing a CPU and input & output device* [Sic !] (= Computer. International Harmonized Commodity Classification).

Search terms for automobiles include:
- 629.331 (Universal Decimal Classification)
- PASS MOT VEH, SPARK IGN ENG (US Federal Import/Export statistics)
- TL 205 (Library of Congress Classification)
- 180/280 (US Patent classification)
- 3711 (Standard Industrial Classification)
- etc., etc.

Increased connectivity means:
-- more use of unfamiliar vocabulary, so
-- increased difficulty in effective and efficient discovery, and
-- greater need for helpful links.
LINKS FROM FAMILIAR TO UNFAMILIAR NAMES, e.g.

Dewey *Decimal Classification* 1876

Railroads 385 -- indicates equivalence.

*Decimal Classification* 1899: Varies “in different connections” (contexts).

Railroads architecture 725

corporations 385

engineering 625

travel 614.863
LINKS TO MORE THAN TWO VOCABULARIES

Combinatorial increase in direct links.

Or, use one vocabulary (e.g. UDC) as pivot (switching language):
Each name is mapped to it and so indirectly to every other.

www.udcc.org/udcsummary/php/index.php

<table>
<thead>
<tr>
<th>English</th>
<th>Esperantoi</th>
<th>Português</th>
<th>Deutsch</th>
</tr>
</thead>
</table>

Mapping “by hand” is difficult, complex, expensive, and obsolescent. e.g. Unified Medical Language System www.nlm.nih.gov/research/umls

Probabilistic mapping can generate search term recommender services rapidly and economically if suitable data is available as a “training set”. (Also called “Classification clustering” (Ray Larson 1991, 1992) and “Instance-based matching”). Easily updated by making a new one.
Metathesaurus Concepts

- **Concept** (~ 1.5M) **CUI**
  - Set of synonymous concept names

- **Term** (~ 5.5M) **LUI**
  - Set of normalized names

- **String** (~ 6.1M) **SUI**
  - Distinct concept name

- **Atom** (~ 7.4M) **AUI**
  - Concept name in a given source

---

<table>
<thead>
<tr>
<th>A0066000</th>
<th>Headache (MeSH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A0065992</td>
<td>Headache (ICD-10)</td>
</tr>
<tr>
<td>S0046854</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A0066007</th>
<th>Headaches (MedDRA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A12003304</td>
<td>Headaches (OMIM)</td>
</tr>
<tr>
<td>S0046855</td>
<td></td>
</tr>
<tr>
<td>L0018681</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>A0540936</th>
<th>Cephalodynia (MeSH)</th>
</tr>
</thead>
<tbody>
<tr>
<td>S0475647</td>
<td></td>
</tr>
<tr>
<td>L0380797</td>
<td></td>
</tr>
<tr>
<td>C0018681</td>
<td></td>
</tr>
</tbody>
</table>
PROBABILISTIC MAPPING FOR SPECIALIZED MATERIAL IN A COLLECTION

Multiple indexes (search term recommender services, mappings) for different specialties in the same collection. Use specialized (biased!) training sets using INSPEC records. Query “Galileo”:

- A collection-wide index recommended: “Jupiter” then “Planetary sciences”
- An Information Science index: “Reservation computer systems” then “Travel industry”
- A Biotechnology index: “History”
- A Water Resources index: “Planetary atmospheres”

All different! The first is from the space probe named Galileo then seeking evidence of water on the planet Jupiter and its moons. The second is from the Galileo online ticketing system then used by the travel industry. The third recognized an historical name, Galileo Galilei. The fourth also was derived from the Galileo space probe. Each valid in its context! The collection-wide index was good for Water Resources but not for other specialties.
## PROBABILISTIC MAPPING FOR DIFFERENT SPECIALIST USERS

Different specialists with same query want different documents from same collection: MeSH queries for “Cardiac arrest” (Vivien Petras, 2006).

<table>
<thead>
<tr>
<th>Anaesthesiologists</th>
<th>Drug therapists</th>
<th>Geriatricians</th>
</tr>
</thead>
<tbody>
<tr>
<td>5. Heart attack, induced</td>
<td>5. Arrhythmia</td>
<td>5. Cardiovascular agents</td>
</tr>
</tbody>
</table>
NOT ONLY SPECIALIST INDEXES BUT PERSONALIZED?

In the context of your own personal knowledge, for A see B.

A good teacher or good librarian will try to do this.

A logical consequence of changing attention from Otlet to Fleck.

Different from the traditional practice of one single, collection-based index for all users which was the only affordable method before digital technology.

Difficult because each person’s “small world” is complex.

Online marketing tries to do this.
LINK RELATIONS

“SAME AS”
Commonly in the form of A <sameAs> B. Strictly, this is always incorrect because no two different entities items can ever be the same (Patrick Hayes, 2011).

“Same” means acceptable alternative: Similar enough for a purpose. Situational.

SYNONYMS MAY NOT BE EQUIVALENT
Alcoholism (English), Alkoholismus (German), and Alcoholismo (Spanish) are associated with different discourses in their different cultural contexts. Not equivalent as search terms.
“FUNCTIONAL” RELATIONSHIPS Example: Biogas

Biogas

Water hyacinths

Pig manure

Water hyacinths and pig manure are important ingredients for biogas, so if interested in one, probably interested in the other two.

Song: “Love and marriage go together like horse and carriage. You can’t have one without the other!”

Related terms, but not a semantic relation. And not always. Easy to provide using co-occurring topic names.
“DUAL NAMING”

Links between different kinds of vocabulary in the same facet.

WHERE: Dual naming: *Place* (cultural) and *space* (physical: latitude and longitude; georeferencing).

*Place name list* ("gazetteer") = bilingual *place – space* dictionary.

WHEN: Dual naming: *Event* (cultural) and *time* (physical: Calendar time). We speak and write using events to express *when*.

*Chronology* (analogous to place name list) = bilingual dictionary of *event* and *time*.

WHERE and WHEN need each other!

Because *places* and *events* are cultural, a place name list should include *when* the record is valid for – and a chronology should indicate *where* the event existed.
WHERE: Dual naming: *Place* (cultural) and *space* (physical).
Example: Lecture Tours of Emma Goldman (Ryan Shaw)
http://metadata.berkeley.edu/emma/
Emanuel Goldberg, b. Moscow, 1881; son of Grigorii Goldberg; Univ. of Moscow, 1900-04; Ph.D w. Robert Luther, Leipzig Univ., 1906; Assistant, Adolf Miethe, TU Charlottenburg, 1906-07; Prof, Akad. f. graphische Künste, Leipzig, 1907-17; ICA, Zeiss Ikon, Dresden, 1917-1933; Kinamo cine camera, 1921; microdots, 1925; search engine, 1927; Contax 35 mm camera 1932; kidnapped by Nazi SA; refugee in Paris, 1933-37; Laboratory, Palestine, Israel, 1937; d. 1970.

Designed for print environment. For digital environment: Normalize terminology in mark-up; add links to resources; interface to generate search queries.
Emanuel Goldberg, b. Moscow, 1881; son of Grigorii Goldberg; Univ. of Moscow, 1900-04; Ph.D w. Robert Luther, Leipzig Univ., 1906; Assistant, Adolf Miethe, TU Charlottenburg, 1906-07; Prof, Akad. f. graphische Künste, Leipzig, 1907-17; ICA, Zeiss Ikon, Dresden, 1917-1933; Kinamo cine camera, 1921; microdots, 1925; search engine, 1927; Contax 35 mm camera, 1932; kidnapped by Nazi SA; refugee in Paris, 1933-37; Laboratory, Palestine, Israel, 1937; d. 1970.

WHERE?
Emanuel Goldberg, b. Moscow, 1881; son of Grigori Goldberg; Univ. of Moscow, 1900-04; Ph.D w. Robert Luther, Leipzig Univ., 1906; Assistant, Adolf Miethe, TU Charlottenburg, 1906-07; Prof, Akad. f. graphische Künste, Leipzig, 1907-17; ICA, Zeiss Ikon, Dresden, 1917-1933; Kinamo cine camera, 1921; microdots, 1925; search engine, 1927; Contax 35 mm camera, 1932; kidnapped by Nazi SA; refugee in Paris, 1933-37; Laboratory, Palestine, Israel, 1937; d. 1970.
WHO: Most links are logical, but most relationships are not logical.

Non-familial relations: <studied at> <studied with>, <assistant to>, <kidnapped by>, <taught at>, <invented>, and <became refugee in>.

Use natural language A <English: co-worker> B.
Use a classification, maybe A <UDC: 622.33> B, for A was a coal miner with B.

Maybe algorithmic processing on such statements.

Ontologies can in theory have unlimited relations.

How practical?

Move from an artificial vocabulary to an artificial language.
FACETS, “FACET GENRES”, AND CONTEXTS

Conventional to divide vocabularies by facet. Arrange reference works by facet: Biography, geography, etc. But, internally, they have other facets.

-- Library subject headings often have a geographic subdivision, a chronological subdivision, or a form designation, and/or a personal name.

-- Place name list (gazetteer) have place name, then geographical feature type, latitude and longitude, etc.

-- Time period directory: Period name as heading, qualified by type of period or event, time markers (calendar time), and where that period or event occurred.

-- A biographical dictionary will be arranged by personal name, followed by multiple instances of activity, date, other persons, and locations (e.g. “b. Moscow, 1881; son of Grigorii Goldberg”, etc.).
Library subject headings

**What** – **Geographic subdivision** – **Chronological subdivision** – **Person**

Place name gazetteer

**Place name** – **Type** – **Space (Lat & long)** – **When** – **Person**

Time Period Directory

**Period name** – **Type** – **Time (Calendar)** – **Where**

Biographical Dictionary

**Who** – **Activity type** – **Time** – **Where** – **Who else**

Same facets in different facet orders.

If we make the facet order the same . . .
Facet genres with facets realigned

Links are usually within one facet within one facet genre. Expand!

WHAT (LCSH)

WHERE (Place name list)

WHEN (Time period dir.)

WHO (Biographical dict.)

From LCSH “Lighthouses” to NGA place name list Geographic Description Code “Lthse” (Lighthouse). Place name list give locations of examples. Catalog gives literature about lighthouses.

Vertical mappings extend semantic links to new vocabularies. Horizontal links provide additional context.

Like using a pre-digital reference library. An agenda for discovery!
CONCLUSIONS

The UDC Consortium Seminar theme is rich at multiple levels:

1. Exciting new tools for combining classification, authority control, and linking for improved research discovery; and

2. An invitation to examine more deeply the fundamental challenge of using formal rational tools in contexts that resist them: phenomena, language, culture, and knowledge.

3. Our field is compromise: Logical tools in an complex world for people who try to make sense. Typical!

Thank-you!