Open Web standards and classification: Foundations for a hybrid approach

UDC Seminar: Classification at a Crossroads
The Hague, 30 October 2009

Dan Brickley, Vrije University Amsterdam
Enough about me:

• Confession - not a librarian.
• (not a classification expert either)
• Accidental technologist -
  • Subject Gateways, eLib, SOSIG & DESIRE
• Web technologies for ‘knowledge sharing’
• Semantic Web standards: RDF, SKOS, OWL
Principles not Technology

• Foundations for a hybrid approach?

• introduction - historical ‘Web’ view.

• principles: obvious, idealistic, and inevitable.

• some examples from TV & today.

• practical suggestions for using Web...
The Palais Mondial
Rooms of the Documentary Encyclopedia
Paul Otlet, the secretary of the Brussels International Bibliographic Institute, estimates the number of printed books since the invention of printing to January, 1900, at 12,163,000 works, and the number of periodicals at between 15,000,000 and 18,000,000.
1000 small answers

• tools, datasets, strategies, techniques
• individual pieces are fragmented, isolated
• closed datasets, software systems
• most classification work today is hidden from most of humanity
Back to basics

• origins of the Web
• origins of subject classification
• (deceptively?) simple suggestions
Let's Share What We Know

World Wide Web
Need for hybrid models:

• 1,000,000,000,000 pages; 7,000,000,000 people
• Information reproducing faster than we are!
• Beware false oppositions:
  • tags ‘vs’ KOS
  • automation ‘vs’ cataloguing
  • ontology ‘vs’ classification
  • Web 2.0 ‘vs’ tradition
‘Hybrid’ means sharing

- New combinations of tools and data
- Using standards (W3C SKOS, ISO, ...)
- Letting data flow between systems
Varieties of knowledge

- ‘Bibliographic facts’, disciplinary views.
- Classifications and thesauri; linked.
- Increasing quantities of base factual data.
- Maps, biographies, Wikipedia, Data.gov, ...
- Ontologies can model books or the world.
- We seem to need both...
Hybrid models

- human classification mixed with automatic
- professional schemes mixed with user tags
- thing-oriented ontologies mixed with discipline-oriented subject classifications
Common aspects

- much innovation underway in each area
- each combines different datasets and systems
The Proposal:

A fundamental strategy for faster innovation

PUT OUR DATA IN THE WEB

(SO OTHERS FIND IT & USE IT & IMPROVE IT)
Once again.

• We need to put our data in the Web.
• In public. Downloadable. Linkable.
• Using standard formats.
• Enthusiastically! loudly! and soon!
What kind of data?

• Thesauri, classification systems, KOS.

• Controlled lists of names and named entities.

• Autoclass configs. Tag cloud linkage. Mappings. Translations. Local improvements.

• Anything someone else might build on.
Business Models

• there will be no “Netscape moment” for us
• a gentle, peaceful decline in usage
• Open Data communities are exploring many license and funding options - MusicBrainz, Wikipedia, Data.gov, data.gov.uk, OpenStreetMap, ....
• No magic answer, but the Web is inevitable
Interlude: Television
Classification & TV

• Inspirations: Douglas Adams’ *Hyperland*
• BBC and the ‘London Classification’
• Obama, Iran, CNN & Debategraph
• Freebase Parallax ‘sets and relationships’
Douglas Adams, “Hyperland, a fantasy documentary” (1990)
BBC Lonclass

- An adapted UDC, used since 1960s
- Used internally to catalogue BBC materials
- A cultural heritage artifact in its own right
LETTER TO MRS THATCHER FROM SIR AUSTIN PEARCE, TELLING MOMENTS INVOLVING MARGARET THATCHER (PRIME MINISTER), THATCHER IN THE TARDIS (ALEXEI SAYLE'S STUFF COMEDY SKETCH) THATCHER FOUNDATION, RESIGNATION OF MARGARET THATCHER AS PRIME MINISTER, RESIGNATION OF MARGARET THATCHER AS MP, THATCHER MEETING WITH LECH WALESA (GDANSK) (NOVEMBER 1988), THATCHER'S REJOICE STATEMENT AFTER THE BRITISH RECAPTURE OF SOUTH GEORGIA, DAVID OWEN IS MRS THATCHER IN A TROUSER SUIT (CATCHPHRASE), DEATH OF DENIS THATCHER, WHOLE REGIMENT OF CLICHES MARCHING INTO THE SUNSET (MARGARET THATCHER QUOTE)
Examples

• R672:32.007(47) YELTSIN:342.518.1 THATCHER “TWO SHOTS OF MARGARET THATCHER AND BORIS YELTSIN”

• [BRITISH AEROSPACE].007.11 PEARCE: 656.881:342.518.1 THATCHER “LETTER TO MRS THATCHER FROM SIR AUSTIN PEARCE”

• 656.881:301.1 62.721:32.007 THATCHER: 654.192.731 TV-AM “MARGARET THATCHER'S LETTER OF APOLOGY TO TV AM”
Compositional Semantics

- 656.881:301.162.721 “LETTERS OF APOLOGY”
- 656.881 “LETTERS (POSTAL SERVICES)”
- 656.881:06.022.6 “RESIGNATION LETTERS”
- 654.192.731 TV-AM “TV AM (TELEVISION AM)”
Then & Now

• 19th century: before modern linguistics, computers, the Internet or Web.

• 2009 - software engineers want to know how to parse these codes and store/query them.

• how to modularise, link out (eg. more info about TV-AM, Mrs Thatcher, artists, topics...).
From classification to argumentation wikis?

“I want this show to stir the global conscience. I have witnessed so much that it’s time to start making real sense of it all”

Debategraph: How to respond to Iran’s nuclear ambitions?

News in Context?

debategraph.org
Amanpour, CNN

Is Iran shifting from a Theocracy to a Military State?

Pro-democracy Iranian refugees fleeing to Turkey?

Does Iran intend to acquire nuclear weapons?

What does Iran want?

What does the international community want from Iran?

Responding to Iran's nuclear ambitions
(3) Iran must take concrete steps in good time

(2) Concrete steps to show that nuclear intentions are peaceful

History of Western involvement in Iran
History of Western involvement in Iran

- Timeline?
- Does the international community want from Iran?
- Change of international attitude towards Iran
- Related maps
- Obama's Speech in Cairo

http://id.loc.gov/authorities/sh85067887
The 1953 coup deposed the democratically elected Mosaddeq government.
History of Western involvement in Iran

http://id.loc.gov/authorities/sh85067887
Meanwhile, in LCSH...

Iran--Civilization--Western influences

URI: <http://id.loc.gov/authorities/sh85067887#concept>

Type: Geographic Name

Alternate Labels: Iran--Civilization--Occidental influences

Broader Terms:

- Iran--Civilization
- Civilization, Western

Created: 1986-02-11

Last Modified: 2007-11-09 07:53:59
a ‘SKOS page’ describing every LCSH concept:

```xml
<rdf:Description rdf:about="http://id.loc.gov/authorities/sh85067887#concept">
  <skos:inScheme rdf:resource="http://id.loc.gov/authorities#conceptScheme"/>
  <skos:inScheme rdf:resource="http://id.loc.gov/authorities#geographicNames"/>
  <owl:sameAs rdf:resource="info:lc/authorities/sh85067887"/>
  <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#Concept"/>
  <skos:broader rdf:resource="http://id.loc.gov/authorities/sh85026485#concept"/>
  <skos:altLabel xml:lang="en">Iran--Civilization--Occidental influences</skos:altLabel>
  <dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime">1986-02-11T00:00:00-04:00</dcterms:created>
  <skos:prefLabel xml:lang="en">Iran--Civilization--Western influences</skos:prefLabel>
</rdf:Description>
```

Sunday, November 1, 2009
Freebase Parallax

Filter Results

3216 topics

* embed these topics as thumbnails • as tiles

**Skyscraper**

**Bank of China Tower**
The Bank of China Tower (abbreviated BOC Tower) is one of the most recognisable skyscrapers in Central, Hong Kong. It houses the headquarters for the Bank of China (Hong Kong) Limited. The building is located at 1 Garden Road, in Central and Western District on Hong Kong Island. Designed by I. M. Pei, the building is 305 m (1,000.7 ft) high with two masts reaching 367.4 m (1,205.4 ft) high. It

**Skyscraper**

**CN Tower**
The CN Tower, located in downtown Toronto, Ontario, Canada, is a communications and observation tower standing 553.33 metres (1,815.4 ft) tall. It surpassed the height of the Ostankino Tower while still under construction in 1975, becoming the tallest free-standing structure on land in the world. On September 12, 2007, after holding the record for 31 years, the CN Tower was surpassed in

**Skyscraper**

**Chrysler Building**
The Chrysler Building is an Art Deco skyscraper in New York City, located on the east side of Manhattan in the Turtle Bay area at the intersection of 42nd Street and Lexington Avenue.

**Skyscraper**

**Central Plaza, Hong Kong**
Central Plaza is the second tallest skyscraper in Hong Kong. With a height of 374 m (1,227 ft), Central Plaza is only surpassed by 2 IFC (415 m / 1,362 ft) in Central. The
Skyscraper →

**Empire State Building**

The Empire State Building is a 102-story landmark Art Deco skyscraper in New York City at the intersection of Fifth Avenue and West 34th Street. Its name is derived from the nickname for the state of New York. It stood as the world's tallest building for more than forty years, from its completion in 1931 until construction of the World Trade Center's North Tower was completed in 1972.

Skyscraper →

**United Nations Headquarters**

The United Nations Headquarters is a
Hypertext Television?

• Every show or scene linked to context
• ...to dialogue, to debate, to literature
• So long as the basic structures are in place
  • from librarians? journalists? users? everyone?
( End of TV intermission )
LCSH is in the Web

http://id.loc.gov/authorities/sh85067887#concept

All of LCSH, officially, in modern machine formats, is part of the public Web.
Not just LCSH

- AGROVOC, RAMEAU, Wikipedia in SKOS
- Dewey.info and UDC experimentation
- A dozen or more thesauri in SKOS
- Described and shared in the Web...
Developer perspective

• Imagine you are a 20 year old programmer.
• You don’t even remember before the Web, never mind the Mundaneum.
• You’re asked to make a useful, usable site from data full of UDC or DDC codes - numbers and punctuation. Help!!!
• You’re told it’s “some library thing...”.
• How can you write software to handle it? Who will help? Whose work can you build on? What linked datasets can you exploit?
“Tekkies think that electronic documents and the World Wide Web are something completely new and that they own it, exactly the way every generation of teenagers thinks they've invented sex and it's their secret.

But it's not new and they don't own it. Word processing and the World Wide Web are not intrinsically new. They are literature.

What is literature? Literature is (among other things) the study and design of documents, their structure and connections. Therefore today's electronic documents are literature, electronic literature, and the question is what electronic literature people really need.”

Ted Nelson, transliterature.org
Not just coders...

• Interface designers
• Information Retrieval experts
• Information Architects
• Data mining researchers
• Ontologists
• Other people like you
UDC - an invisible Web?

• A huge community of practice links libraries across Europe and the World

• Dozens of systems, datasets, mappings, enrichments

• Seminars are a great way to share experiences

• .... and the Web is a great way to share and link data.
In 5 years time...

- What data and tools can we show the developers who make Firefox (35% European web users use it)
- Or Drupal 7 (which powers 1000s of sites)
- Or OpenOffice (where docs are written)
- Or Google?
Warning signs

• Q: Why don’t you show labels on the codes?
• A: Because we’d have to buy a license.
For discussion

• The road to relevance in the modern Web is through developers, designers and content creators.

• Without them, we’ll drown in data.

• Without our data, they won’t know how to help ... or realise that they want to?
Let’s Share What We Know

World Wide Web
Acknowledgements & Links:

- http://www.w3.org/Illustrations/LetsShare.ai graphic design by Robert Cailliau

- Lonclass visuals generated by Max Froumentin's Textorizer - http://lapin-bleu.net/software/textorizer/

- 'Palais Mondial, Rooms of the Documentary Encyclopedia' from W. Boyd Rayward, "The Universe of Information - The work of Paul Otlet for documentation and international organisation" (FID 520), Moscow : Published for International Federation for Documentation (FID) by All-Union Institute for Scientific and Technical Information (VINITI), 1975. p.237. http://people.lis.illinois.edu/~wrayward/otlet/PAUL_OTLET_REFLECTIONS_ON_BLOG.HTM

- Paul Otlet estimates on number of printed books, from Shiner gazette. (Shiner, Tex.), August 31, 1904. via http://chroniclingamerica.loc.gov/lccn/sn86090270/1904-08-31/ed-1/seq-7/

- Google Blog, "We knew the web was big..." 25 July 2008, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html

- Debategraph project, http://www.debategraph.org/


- Freebase Parallax by David Huynh, http://www.freebase.com/labs/parallax/

- Many thanks to my NoTube, VU and FOAF collaborators for ideas and brainstorming on Hypertext TV scenarios; in particular regarding UDC/Lonclass at the BBC, Vicky Buser, Libby Miller, Patrick Sinclair, Brendan Quinn, Yves Raimond and others have been exploring these issues; at VU this work is in collaboration with Guus Schreiber, Lora Aroyo, Antoine Isaac, Ronald Siebes and Chris van Aart; also many thanks to Aida Slavić for help understanding the nature and structure of UDC.