Understanding the standards gap

LIDA 2009 - Libraries in the Digital Age
Dublin Core Tools Community workshop
30 May, Zadar, Croatia

Dan Brickley
Dublin Core RDF tools in theory & practice....

Standards Gap =
what you want
minus
what you get

What RDF’s standards do “out of the box”
...and what we have to build ourselves.

A brief introduction to Semantic Web thinking,
and a proposal for role of the DC Tools group
If you buy, install or build RDF/DC “standards-compliant” metadata tools, what’s still missing? how to fill that gap?

Two conceptual tools:

1. Understand the RDF & Semantic Web style -
   Information-linking using Web technology

2. Understand Dublin Core history: DC as element set vs DC as a community meeting place -
   from 15 DC metadata elements as ‘the’ solution, to DC community as place to share & find solutions
The Semantic Web project

A true story about linked information systems:

From many small pieces of information ("claims")...

...to the total universe of information ("The Web").
Claims? as properties, relationships and attributes...

W3C Resource Description Framework (RDF):

simple factual claims in the Web.

(Q: are these claims useful for anything unless you know who made them?)
(Q: how do these claims relate to documents? to their creators? to hyperlinks? provenance?)
The world according to a document: claims as triples
http://www.bristol.ac.uk/
Acronym Guide part 1:

RDF/XML, RDFa, GRDDL, N3/Turtle standards

These are all ways of encoding simple 3-part claims in various Web documents.

That is all that they do.

So different documents can make different claims?

Yes.
What the computers see...

3 sets of claims
Acronym Guide part 2:

SPARQL - an RDF query language

SPARQL lets you ask factual questions against a database of these claims. That is all that it does.

And the queries can talk about who made each claim?

Yes.
What school does Alice say she went to?

Who has something to say about Alice’s schooling?
- does Alice’s school agree that she attended it? Does Bob’s?
- when sources disagree, how to prioritise? who to believe?

This is RDF’s approach - understand it and you’ll understand what you get from RDF tools, RDF data, RDF people...
SKOS in the Web: each concept gets a page...
SKOS is a set of terms for making claims about subjects/topics, their properties and relationships.

(Q: how might employers, academia or individuals use it to make Web-based claims about expertise?)
Beyond Toy Examples

Each Library of Congress Subject Heading has an RDF/SKOS page

... 6.5M bib records, ~200k auth records Swedish Union Catalog

Many thesauri. All of Wikipedia (dbpedia.org). Yahoo & Google.

From http://www.gutenberg.org/wiki/Gutenberg:Feeds

<dc:subject>
  <rdf:Bag>
    <rdf:li><dcterms:LCSH><rdf:value>Epic poetry, Greek -- Translations into English</rdf:value></dcterms:LCSH></rdf:li>
  </rdf:Bag>
</dc:subject>

To using “http://id.loc.gov/authorities/sh2007100520”
People care about metadata standards for many reasons:

- long term archival
- integration with other collections
- lower cost of systems development
- reduce lock-in to commercial or opensource systems
- potential for alternative UIs
- consistency of interface for users

None of us get all we need from off-the-shelf tools...
So ... DC tools & the standards gap...

(the difference between what you want and what you get...)

Typical RDF off-the-shelf toolkit offers:

- Parsers, XML & SQL mappings, query
- and rule systems, databases.

That’s about it. But solid, standardised, well understood and with many independent implementations...

What more could you want?
Full text search, autocompletion systems, sophisticated ranking algorithms, tagging, search and browse interfaces. Tools to check and validate conformance to ‘application profiles’ (ie. particular patterns of descriptions & claims). Conversion and mapping tools. Analytics. Stats. Tools that compensate for inconsistent, semi-chaotic data. Tools that help authors and publishers. [...] 

... & .... documents that describe the available tools and their use cases. 

*Dublin Core is not a technical answer to all questions, but a place where practitioners can share answers, experiments, ideas*...
Back to the future?

TODO.txt

1. (re-)read “Information Management: A Proposal” (TimBL, 1989)

2. Document your own standards gap in the DC Tools community: what do you need that isn’t ready yet? What have you tried? What works, what doesn’t?
Questions?

mailto:danbri@danbri.org