Preservation

Notes from a break-out session

DC-2012 Special Session on Vocabularies and Alignment
21 September 2011, The Hague

- Access mechanisms to vocabularies: technologies will change
  - Currently using DNS
  - Future will be something else
  - How do we migrate?
- URIs
  - Identification of the object and access mechanism: same identifier
- Use persistent identifier that is embedded for now in
  - URL: what it resolves to can change
  - With persistent ID, object can move around more easily than with DNS redirects

- Split persistence
  - Organizational/political
  - Technical
- With highly specific systems, lack expertise to maintain.
  - With persistence, maintenance part is the most difficult.
  - Becomes less and less useful.
  - Persistence not just keeping available, but that.

- Mechanism for mapping old vocabularies to new.
- In sciences, ad-hoc vocabularies all the time.
  - Folksonomies that become vocabularies, get abandoned, what's left is useless.
- Organizations should commit
  - But organizations can collapse, so need failr safe mechamisn:
- When you start, need exit stragegy
  - Wider community that has guarantees survival that is strongly supported.

- Three-year funding cycles.
- Be explicit, state up-front.
- People can go back to it, re-map.
- Need to be clear what needs to be preserved.
  - Maintenance of resolution system.
  - But last resort might be paper or PDF.
- Heart the problem is a business model
  - Language resources area. Build up knowledge bases from producers of data. How can an organization commit resources long-term? Governments are still the best safeguard, but must be learn.

- LOCKSS just preserves bits
  - But long-term, may need to be migrated.
- In addition, need technical environment (access and maintenance)
  - Organizations building ontology libraries
  - But cannot get off-the-shelf, commercial systems – tehcnical specs that limit interoperability.
  - Build national repositories for vocabularies.
  - Like to have international system that harvests SKOS-based.

- Search engine harvesting SKOS vocabularies.
  - Internationally centralized? Should be locally decided instead of
  - Would not be expensive service to run.
- For preserving vocabularies, also need documentation: structured descriptions
- What to preserve?
  - Documentation
  - RDF schemas resolvable – but not “preservation”

- Any scientific dataset: 5 years half-life.
  - If you don't want the data, you don't want the vocabulary.
  - Need to preserve some vocabularies long-term.
- Vocabularies need to evolve.
- Changing tool used for classification – millions of books in the old system.
- Snapshots: cheap to make
  - Periodically refresh cache of the Subversion repository of a vocabulary

- Libraries that prioritize
- Not expensive to keep a frozen copy.
- Europeana as repository of vocabularies?
- Scientific communities: each one has its own terminologies.
  - Most not formalized (SKOS, etc)
  - We need to go into their communities, understand which vocabularies used in which workflows.

- First ring: developing vocabularies
- Need to preserve environment (hardware, software)?
- Need to preserve access (long-term preservation).
- Not just LOCKSS, but several, e.g. IRODS: same general functionality.
  - IRODS: local client can specify what to ingest
  - If all systems fail simultaneously… but better
  - More an archival solution.

- A vocabulary that is not maintained will not remain useful.
  - Whether a preserving organization
  - SKOSification as part of the long-term preservation process
    - Makes it easier to move
- Levels of preservation?
  - What are the risk factors.

- Vocabularies are dynamic, unlike publications.
- Need to maintain system that keeps it useful.
- Example showing importance or danger of vocabulary preservation?
- Risk factors, Preservation levels – needed for justifying funding.
- Kevin Clair, Jain Qin, Shigeo, Raju, Juha, Bob Boelhouvwer, Doug Moncur