Long term identifiers for terms in a crowd-sourced vocabulary

DataONE Preservation and Metadata Working Group (PAMWG)

John Kunze, California Digital Library
Greg Janee, UC Santa Barbara
Christopher Patton UC Davis
Agreeing on terms: a different take

• Most metadata standards controlled by experts
• Change by committee is costly and slow
• Progress: Dublin Core was designed to break with the past, but still room to improve
  • 5 years to agree on same 15 terms we had at 1 year
  • highly divergent, non-interoperable local use with frequent reference to many external ontologies
Metadata Vision

One dictionary, one namespace

- Crowd sourced plus lightly supervised canon
- Anyone can look up terms
- Any domain, any part of “metadata speech”
  - Names, values, units, relationships, ...
- Anyone can propose and refine their terms
- Strong terms rise, weak terms decline

What have we learned from Wikipedia, the Internet-Draft/RFC process, and the American Heritage Dictionary?
Metadictionary

A crowd sourced metadata dictionary. Search for terms, upvote useful ones.
<table>
<thead>
<tr>
<th>Term</th>
<th>Votes</th>
<th>Consensus</th>
<th>Class</th>
<th>Contributed by</th>
<th>Last modified</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hello</td>
<td>-1</td>
<td>0%</td>
<td>vernacular</td>
<td>Chris Patton</td>
<td>16/7/2013 21:28</td>
</tr>
<tr>
<td>identifier</td>
<td>0</td>
<td>50%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>16/7/2013 1:32</td>
</tr>
<tr>
<td>structured datum</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>16/7/2013 1:32</td>
</tr>
<tr>
<td>publisher</td>
<td>3</td>
<td>100%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>16/7/2013 1:21</td>
</tr>
<tr>
<td>CHL</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>Greg Janée</td>
<td>15/7/2013 15:30</td>
</tr>
<tr>
<td>great</td>
<td>1</td>
<td>100%</td>
<td>vernacular</td>
<td>Nassib Nassar</td>
<td>15/7/2013 13:07</td>
</tr>
<tr>
<td>creator</td>
<td>1</td>
<td>100%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>14/7/2013 23:02</td>
</tr>
<tr>
<td>talus</td>
<td>0</td>
<td>50%</td>
<td>vernacular</td>
<td>Angela Murillo</td>
<td>14/7/2013 22:58</td>
</tr>
<tr>
<td>data</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>14/7/2013 22:44</td>
</tr>
<tr>
<td>description</td>
<td>1</td>
<td>100%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>14/7/2013 22:44</td>
</tr>
<tr>
<td>metadata</td>
<td>1</td>
<td>100%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>14/7/2013 21:58</td>
</tr>
<tr>
<td>hydraulic gradient</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>Angela Murillo</td>
<td>14/7/2013 15:19</td>
</tr>
<tr>
<td>talus slope</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>Angela Murillo</td>
<td>12/7/2013 22:00</td>
</tr>
<tr>
<td>metadatum</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>12/7/2013 21:16</td>
</tr>
<tr>
<td>structured data</td>
<td>0</td>
<td>0%</td>
<td>vernacular</td>
<td>John Kunze</td>
<td>9/7/2013 23:59</td>
</tr>
</tbody>
</table>

Long term identifier: MISSING!
Browse dictionary

high score | recent | volatile | stable | alphabetical

C
- canine contributed by Deb
- CHL contributed by Greg
- creator contributed by John

D
- data contributed by John
- datum contributed by John
- datum contributed by Chris
- description contributed by John

H
- hydraulic gradient contributed by Angela
Persistent *term* identifiers

- Persistent official *concept* name
- Disambiguation of natural language terms
- ARK ids (expressed as URLs) from CDL’s EZID service, suitable for linked open data apps
- Work with Research Data Alliance PIT group
  - scheme-agnostic resolution (N2T resolver)
  - added extended-term “fake” ids
Questions?

Try out the Sealce prototype at
http://seaice.herokuapp.com
https://github.com/cjpatton/seaice

Try out EZID at http://n2t.net/ezid