Panel Discussion presentation: "Value-based Indicators for Reuse & Their Implications for Data Curation"

Nic Weber
University of Illinois at Urbana-Champaign

Follow this and additional works at: https://escholarship.umassmed.edu/escience_symposium

Part of the Library and Information Science Commons

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License.

https://escholarship.umassmed.edu/escience_symposium/2013/program/1

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in University of Massachusetts and New England Area Librarian e-Science Symposium by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Value-based indicators for reuse & their implications for data curation

03 April 2013
UMass + New England Area Librarians eScience Symposium

Nic Weber

Tiffany Chao, Karen Baker
Andrea Thomer & Dr. Carole Palmer
Overview

I. The Data Practice Working group
   - What we talk about when we talk about Value

II. Some research findings
   - Qualitative Case: The Data Conservancy
   - Quantitative Case: NCAR’s Research Data Archive
Data Practices Research Group

Dr. Melissa Cragin, Tiffany Chao, Karen Baker, Andrea Thomer & Dr. Carole Palmer

Qualitative & Quantitative Studies of Data production and use

**Aims:** Inform development of curation services and systems

**Focus:** Long-tail, heterogeneous, ‘small’ data-intensive science

<table>
<thead>
<tr>
<th>Range</th>
<th>$300,000 - $38,131,952</th>
<th>$579 - $300,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Grants</td>
<td>2405</td>
<td>9621</td>
</tr>
<tr>
<td>Total dollars</td>
<td>$1,747,957,451</td>
<td>$1,117,431,154</td>
</tr>
</tbody>
</table>

12,025 NSF grants awarded in 2007 = $2,865,388,605
Value-based Indicators

(Some of our working assumptions)

If long-term preservation is the goal, research libraries and data centers want to make targeted investments in high value data collections.

The value of data is a socio-technical phenomenon.

The view of data value is a relational one.

Value is not necessarily dependent on quality, size, scale, support, rarity or expense.

The value of data increases with use.
Qualitative Studies of Re-Use
Earth System Science (ESS)

Macro-perspective of Earth Science work: Data Communities, Evidential Cultures.

"...the global earth environment can be understood only as an interactive system embracing the atmosphere, oceans, and sea ice, glaciers, and ice-sheets, as well as marine and terrestrial ecosystems"

Asrar, Kaye & Morel, 2001
## Sub-disciplinary Profiles

<table>
<thead>
<tr>
<th>Study approach</th>
<th>Soil Ecology</th>
<th>Volcanology</th>
<th>Stratigraphy</th>
<th>RS Engineering</th>
<th>C/O Modeling</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Study approach</strong></td>
<td>Biotic and abiotic properties of soil</td>
<td>Chemical and textural properties of rock samples combined with geospatial data</td>
<td>Range of signals compared to refine the geological time scale</td>
<td>Prototyping and designing field sensors to optimize field data collection</td>
<td>Computational or mathematical modeling of aquatic dynamics.</td>
</tr>
<tr>
<td><strong>Kinds of data used</strong></td>
<td>Physical soil samples, maps (paper &amp; digital), biological species inventory, lab-based outputs</td>
<td>Whole rock samples; thin slices of rock samples on glass slides; chemical data; maps</td>
<td>Numerical data and graphs pulled from papers; physical samples; chemical, radioactive isotope, and astronomical cycle data</td>
<td>Autonomous field measurement of sensor and environmental data recorded on data loggers or transferred directly to a database</td>
<td>Water sample, meteorological, and remote sensing data downloaded; diverse models’ output at many spatial &amp; temporal scales</td>
</tr>
<tr>
<td><strong>Patterns of data use</strong></td>
<td>Systematic review of data for quality where values are checked against multiple sources</td>
<td>Iterative reference to &amp; comparison of data sources, including chemical data, field notes, papers &amp; maps</td>
<td>Highly iterative comparison of datasets and modeling of signals of time</td>
<td>Regular review of data for investigating various sensor configurations and contexts of data collection</td>
<td>Irregular patterns of use, based on need for model calibration or benchmarking for reliability</td>
</tr>
<tr>
<td><strong>Norms of data re-use</strong></td>
<td>Informal sharing of processed data and methods, though perceptions on re-use vary</td>
<td>High expectation of data re-use, particularly with physical samples and thin sections</td>
<td>Moderate expectation of re-use aiming to find new ways of determining geological time scales for re-use</td>
<td>Diverse, informal re-uses: optimizing sampling design; providing data to project researchers; or for public posting</td>
<td>Informal sharing of data inputs and software code; Informal and formal mechanisms for re-use and sharing of model</td>
</tr>
</tbody>
</table>
Value types: Frequent Data Re-Users

“...you have to go back to the data gatherer and ask them, “What’s this (cell) value? This doesn’t seem to be right.? Do you remember what happened? Did a shark hit your boat or something?” ...the quality control doesn’t exist really well. So one has to work back and forth with the data collector.” Ocean Modeler

Value types observed: Verification, Depth of Description, Equivalence

Implications for Systems & Services Development:

Enable users and curators to trace provenance and context of production.

Data change in value based on the context of communities of practice- and participation in communities of practice are more dynamic than we often assume.

Identifying data producers (authorship ?) is burgeoning issue of importance for meaningful re-use. We have to come up with sound guidelines, and be able to establish persistent ways of tracking data producers (ORCID IDs!)
"We have people who are participating in triathlons...and they want to know about the water temperature and want to know about patterns of temp change. We've had Search and Rescue teams download our data to be able to predict what will be going on...fishermen will request data to look at trends...We also have industry people, need to know what the typical water level will be so they can get their boat in there."
Sensor Engineer

Value types observed: Regenerative, Malleability

Implications for Systems & Services:

Design infrastructures to recapture secondary products, serve flexible / shifting client base. Discovery is still ad-hoc, back channel.

Find ways for signals of value to be consumed by both curators and re-users.
Types of Value

**Re-users** (How they describe valuable data for their own work)

- Verification: This data helps me trust / refute existing data source
- Depth of description: This data adds to basic understanding of existing data source
- Equivalence: This dataset is the same (content) as that data source

**Producers** (How they imagine their data having value)

- Regenerative: Do data have “reach” beyond original intention or application?
- Malleability: How flexible or fragile are data to new application, new domain or new method?
Quantitative Studies of Re-Use
Data Citation @ NCAR

Scientist

NASA data archive

Data set 1

DOI 1

Data set 2

DOI 2

NCAR data archive

Data set 3

DOI 3

Data set 4

DOI 4-9

Data set 5

DOI 10
Holdings > 1.3 PB, static and dynamic datasets including...

Atmospheric and oceanographic observational data, weather prediction model output, gridded analyses and reanalyses, climate model output, and satellite derived data

2012 served ~1 PB to ~1500 unique users from 127 different countries.
The Challenge

“2012 served ~1 PB to ~1500 unique users from 127 different countries”

Impressive, but not very meaningful.

At best it’s an incomplete picture of the RDA’s curation work, and it’s impact on Earth Science domain.
Is the RDA less helpful now?

* Subsets decreased
* Assisted users downloading decreased
* Advanced users downloading increased
The Data Usage Index

Originally developed for Biodiversity database (GBIF) by Ingwersen and Chavan (2011)

Takes suite of archive-user interaction metrics

Uses these as indicators or proxy of impact, that can be later combined in unique ways to demonstrate value

Indicators are standard across datasets, allows for comparisons across different data types, and time periods.
<table>
<thead>
<tr>
<th>Indicator</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unique Users (UU)</td>
<td>Unique users that downloaded data during a time window</td>
</tr>
<tr>
<td>Unique Users - Advanced</td>
<td>UUs that accessed data programmatically</td>
</tr>
<tr>
<td>Unique Users - Assisted</td>
<td>UUs that accessed data via GUI or Service</td>
</tr>
<tr>
<td>Number of Datasets</td>
<td>Number of Datasets assigned DS number</td>
</tr>
<tr>
<td>Files DS</td>
<td>Number of files in Dataset per time window</td>
</tr>
<tr>
<td>Download Frequency</td>
<td>Total number of files downloaded per time window</td>
</tr>
<tr>
<td>Download Frequency - Advanced</td>
<td>Files downloaded by Advanced users</td>
</tr>
<tr>
<td>Download Frequency - Assisted</td>
<td>Files downloaded by Assisted users</td>
</tr>
<tr>
<td>Homepage Hits</td>
<td>Dataset Homepage Hits per time window</td>
</tr>
<tr>
<td>Homepage Hits - Direct Access</td>
<td>Dataset Homepage Hits per time window by users with direct access (link not indexed or retrieved by search)</td>
</tr>
<tr>
<td>Homepage Hits - With Link</td>
<td>Dataset Homepage Hits per time window by users with link (from indexed list or retrieved by search)</td>
</tr>
<tr>
<td>Subset Requests</td>
<td>Subsets Requests per time window</td>
</tr>
<tr>
<td>Download Density</td>
<td>Average number of files downloaded per UU</td>
</tr>
<tr>
<td>Usage Impact</td>
<td>Total number of downloaded files over total files in dataset</td>
</tr>
<tr>
<td>Usage Impact - Advanced</td>
<td>“</td>
</tr>
<tr>
<td>Usage Impact - Assisted</td>
<td>“</td>
</tr>
<tr>
<td>Interest Impact</td>
<td>Total homepage hits per number of files in dataset</td>
</tr>
<tr>
<td>Usage Balance</td>
<td>Files downloaded by number of homepage hits per time window</td>
</tr>
<tr>
<td>Subset Ratio</td>
<td>Number of subset requests over total number files downloaded per time window</td>
</tr>
<tr>
<td>Secondary Interest Impact</td>
<td>Homepage over UU</td>
</tr>
</tbody>
</table>
Impact indicators

Tell a more complex story using combination of metrics

Allows us to convey to funding agencies long-term influence of introducing new services

And most important, can give long-term view of value.
Lessons learned

Metrics are a (painful!) craft process:

Start with a baseline (Data Usage Index)

Adapt for the specificities of your domain and your archive.

Find weird patterns, and explore (Science!)
Thank you.

nmweber@illinois.edu

@nniiicc
Papers where this work appears

Qualitative


Quantitative
