Optimizing Data Resources for Reuse: Site-Based Data Curation

Carole L. Palmer
Center for Informatics Research in Science & Scholarship
Graduate School of Library & Information Science
University of Illinois at Urbana-Champaign

CyberGIS Brown Bag
11 February 2014
“... make data resources “accessible, discoverable, and usable to fuel entrepreneurship, innovation, and scientific discovery.”

Online Scientific Data Curation, Publication, and Archiving (Gray, et al., 2002)

- managing proliferation of data, derived products
- capture “ephemera” otherwise data “incomprehensible, useless”
Leverage data into “major downstream products”

Need data to become fundamentally collective, shared resource

- diverse data formats
- collecting and standardizing metadata
- disciplinary cultures

Open and Closed Evidential Cultures (Collins, 1998)

What qualifies a releasable result? wave or “coincidence”

Who takes responsibility for validity and meaning? lab or research community
Data practices research

Data Conservancy

Oceanography
Climate science - modern
Climate science - paleo
Soil ecology
Volcanology
Stratigraphy
Mineralogy
Microbiology
Sensor network science
Environmental engineering
Photonics

Site-Based Data Curation at YNP

Geology
Geochemistry
Microbiology

Curation Profiles Project

Earth and Atmospheric
Anthropology
Plant sciences
Kinesiology
Speech and Hearing
Data production, use, & sharing across fields

- variations in curation requirements
- re-use value

What factors are associated with value?

How are dependencies among research communities evolving around data resources?

- interdisciplinary research; need for data from outside fields
- curation expectations for sharing vs. reuse
Formalized levels of data products

**NASA data levels**

- **L0**: Raw instrument data
- **L1A**: Reconstructed, raw instrument data
- **L1B**: Geolocated and calibrated
- **L2**: Products derived from L1B
- **L3**: Gridded and quality controlled
- **L4**: Model output; derived variables

<table>
<thead>
<tr>
<th>Raw data</th>
<th>Telemetry data with data embedded</th>
<th>Little use to most of science community, except radio science</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 0</td>
<td>Edited</td>
<td>Corrected for errors, split into dataset for instrument; tagged with time &amp; location</td>
</tr>
<tr>
<td>Level 1A</td>
<td>Calibrated</td>
<td>Corrected and expressed proportional to some physical unit</td>
</tr>
<tr>
<td>Level 1B</td>
<td>Resampled</td>
<td>Resampled and possibly calibrated; can’t be reconstructed</td>
</tr>
<tr>
<td>Levels 2-5</td>
<td>Derived</td>
<td>General way information transferred</td>
</tr>
</tbody>
</table>

Heterogenous earth science data

Geobiology “dataset”
rock & water chemistry, microscopy images, annotated field photos, microbial genomic data

Sharing conventions:
personal contact after publication

<table>
<thead>
<tr>
<th>Designated community</th>
<th>Geobiology, Microbiology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Potential communities</td>
<td>Chemistry, Evolutionary biology, Bioprospecting, U.S. Park Service, Public Health</td>
</tr>
<tr>
<td>Reuse application</td>
<td>Microbial data – disease presence, extent</td>
</tr>
</tbody>
</table>

“... a lot of geochemical work that relies less on field context. ... somebody who knows about isotopes can take my data and do a whole different series of investigations.”
Releasable products, meaning for reuse

version willing to release vs. best for reuse

complex sets vs. usable parts

Curation Profiles Project

Data sharing is about data producers
Data reuse is about data consumers
Value indicators

- Reputation of data collector
- Longitudinal coverage
- Multiple sources
  - triangulation & context
- Workflows and provenance

**Site factors:**
unique conditions
rarely studied,
politically volatile
permitting

Climate modeling
Ocean modeling
Soil Ecology
Volcanology
Stratigraphy
Sensor Engineering
Site-based Data Curation at Yellowstone

Model for curation processes and policies that capitalize on data from scientifically significant sites.

Collaboration with experts in:

- Data curation & archiving – Sayeed Choudhury, JHU
- Geobiology – Bruce Fouke, IGB
- NPS resource management – Ann Rodman, YNP
**Problem**

**Specimen data** – NPS reporting requirements
- taxonomic identifications
- number of specimens
- location collected with UTM coordinates
- date/time collected
- method of preservation or preparation
- current location

**Digital data** – no requirements, but value recognized
- multiple investigators at same geologic features over time
- opportunities for more integrative, comparative science

Clear benefits for both scientists and site management in advancing preservation, access, and utility.
Yellowstone as exemplar

Leader in policy for scientific research.

Research ranges from origins of life on Earth to life on other planets.

Protected natural system - indicator of world events.

- 12,000 diverse thermal features with distinct boundaries.

- Ease of access (but not publically exploited).
  - refine questions & data collection methods
  - verification of system-wide hypotheses & results
  - patterns and trends over time

- Accurate and stable sampling.
Initial focus on systems geobiology

- Representative of heterogeneous “long tail” research & data types.
- Individualized programs of research - not dominated by large team projects.
- Aim of data integration among physical, chemical, biological.

- Microbes — 55%
- Geochemistry — 10%
- Geothermal — 10%
- Geology — 5%
- Viruses/Prions — 5%
- Hydrology — 4%
- Geomorphology — 3%
- Atmosphere — 2%
- Seismology — 2%
- Stratigraphy — 2%
Methods and goals

Workshop, focus groups, interviews, worksheets

Active YNP researchers: geologists, geochemists, & microbiologists

YNP personnel: research permitting & reporting, GIS, library and archive

Data assessments: inventories, curation requirements, transfer workflows

Curation processes and policies

- Criteria for reusable datasets and descriptions
- Collections / key groupings of data over time
- Coordination with permitting & reporting

Workshop report:
http://hdl.handle.net/2142/47070
Reuse dependent on record of sampling

- detailed, consistent description of data collection

**Minimal description**

1. Investigator Name(s)
2. Date and Time
3. Location - at vent if possible - 3X different types of input coordinates
4. Site Name – NPS feature names
5. Sample ID, Code numbering scheme - reference codes
6. Altitude (Elevation) – meters derived from topo maps or GPS systems
7. Field photograph – embedded digital information on sample sites
8. Temperature – with brief description of techniques
9. pH – with description of techniques

**Additional requirements for**

geological, water, weather, and biological
Photos key to data organization

- Explicit relationships among all data from a given trip to field
  - reliable capture of context and conditions
  - tracking changes over time
  - support browsing

Core metadata +
geological, water, weather, biological elements
Value of YNP data in aggregate

Science aims – access to reusable data sets

- Longitudinal studies across features
- Exploration of dynamic systems
- Site evolution over time
- Suites of metabolic function & relationship to environment
- Evolution of microbial communities in space and time

YNP aims – metadata functionality

- Ecological impacts of research at features
- Coordination & transparency of data collection activities
  - trends and connections across projects
  - permitting to avoid replication of data
Broader impacts and strategic science

Scientists starting to emphasize broader impacts of data sharing

Improving **scale and efficiencies** of science

- Broader, more systematic & synthetic inquiry.
- Assess feasibility & next steps of research projects and agendas based on existing resource base.
- Foster collaboration among investigators.

Public data with adequate context to benefit

 science & data literacy
public perceptions of science
Next steps

• Procedures and workflows for packaging, describing, and transferring data
• Assess compatibility with other NPS systems
• Curation awareness materials for permit packages
• Demonstration with data across YNP geologic features
• Test, extend model for other site-based science
Acknowledgements

Site-Based Data Curation

Project team:


JHU: Tim Dilauro

Science advisors:

John Peters
John Spear
Cristina Takacs-Vesbach
David Mogk
Bill Inskeep
Gary Olsen

YPN advisors:

Christie Hendrix
Colleen Curry
Anne Foster
Bob Fuhrmann
Henry Heasler
Susan Kelly

Data Conservancy

PI: Sayeed Choudhury, JHU

Illinois Data Practices Team

Tiffany Chao
Nic Weber
Karen Baker
Andrea Thomer
Melissa Cragin
(now at NSF)
“... active and on-going management of data through its life cycle of interest and usefulness ...

... enables data **discovery** and **retrieval**, maintains **quality**, adds **value**, provides for **re-use** over time.”

(Lord & MacDonald, 2003)

---

managing and promoting use from point of creation enrichment & updating to keep **fit for purpose** discovery and re-use.

(Lord & MacDonald, 2003)
Methods developed for data service applications

**Lead scientists** - research context, sharing, access, discovery, re-use

1) Pre-interview worksheets
2) Semi-structured interviews
3) Follow-up sessions with selected participants

**Researchers managing data** - stages, versions, standards, tools

4) Data deposit & sharing worksheet
5) Data samples, related documentation

**Rapid reporting framework**
- quick turnaround of key observation for JHU infrastructure team
Specialization in Data Curation

Required Courses

- Foundations of Data Curation
- Digital Preservation
- Metadata in Theory & Practice

Selected Electives – require 2, recommend 4

- Systems Analysis & Management
- Information Modeling
- Foundations of Information Processing
- Digital Libraries: Research & Practice
- Representing & Organizing Info Resources
- Ontologies
- Information & Collaboration in Science

* Note: <br>Switch in Metadata and Systems Analysis starting this year.
Program of research and education

Data Conservancy

Information & Discovery in Neuroscience

Site-Based Data Curation at YNP

Profiling & Curating Enterprise Data

Digital Collections & Content

EDUCATION

Data Curation in the Sciences 2006-11

Biological Information Specialists 2006-09

Summer Institutes in Data Curation 2008-11

Data Conservancy Education Initiatives 2009-12

Data Curation Education in Research Centers 2010-14

Data Curation in the Humanities 2008-12

2008-12

2006-11

2006-09