Morning Address, Part 2: "UC San Diego Curation Pilots: Planning for the Future"

David Minor
University of California - San Diego

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

Part of the Library and Information Science Commons

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

Repository Citation

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.
Table of Contents

• What did we do?
• Findings
• Recommendations
• Implementation
Data Curation defined in planning documents

The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and other published materials.

Source: *From Data Deluge to Data Curation*
Philip Lord, Alison Macdonald, Liz Lyon, and David Giaretta

- appraisal
- accession
- arrangement
- description
- storage
- preservation
- access
The activity of managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and reuse. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and other published materials.

Source: *From Data Deluge to Data Curation*
Philip Lord, Alison Macdonald, Liz Lyon, and David Giaretta
Curation Pilot Story

The curation pilot goals:

• Investigate what it means to make a variety of research data discoverable and reusable
• Investigate current UC San Diego tools for accomplishing this work
• Learn how researchers, information technologists, and librarians work together with data
• Recommend production services
• Develop budget and cost models
The pilot participants ...
Preserve and curate the digital version of the brain of patient HM, the most studied neuropsychological patient in modern medicine.
• Existing web site pointing to some of the collection – what is relation to our work?
• A number of display and organization requests for the data
• Complex display needs – 3D, multiple layers
• Data set included related materials such as published papers, videos, interviews, emails
OpenTopography facilitates community access to high-resolution, Earth science-oriented, topography data, and related tools and resources.
• Existing user interface that showed derived data products – didn’t need to be recreated
• Primary source data was sitting on hard drives on desks, not preserved, not accessible, not organized across collections
• OpenTopo is grant-funded, with a finite funding life span
• Wanted DOIs to reference data
• Had a number of ownership and access issues
• This data set used the fewest curation services
Levantine Archaeology Laboratory

Focuses on archaeological investigations concerning the evolution of societies in the southern Levant from the Neolithic to Islamic periods.
• Data were stored in a number of locations
• Metadata differed depending on context and collection method
• Mix of digital and physical objects
• A number of discoverability requests in terms of display and access
• Complex system of object relationships
• This discipline has a variety of standards for data organization and metadata
The Sediment Core collection contains samples collected from as early as 1916. The Cored Sediment Collection is a growing archive of sea-floor samples and associated data supporting a diverse variety of scientific research.
• Extensive help with data organization
• Inconsistent metadata because of several decades of various collection methods
• Local data are shared with a national database that organizes, references, and creates metadata differently
• Data are used by multiple communities who have different metadata for similar objects
• Used several methods for working with their data
• Complex relationship between physical and digital objects
The Laboratory for Computational Astrophysics

Dedicated to advancing the state-of-the-art of astrophysical simulation through the development and dissemination of community codes, and through large-scale simulations of astrophysical and cosmological systems.
• Data needed better organization
• Internal operation data was mixed with useful outward facing data
• A library application (Archivists’ Toolkit) was used to create a finding aid for the data set
• The way information is presented in the finding aid is similar to what we have seen in data papers
• This is the most curated data set
What does discoverable and reusable look like?
Discovery and Access Resources

• The **UC San Diego Library's Digital Asset Management System** (DAMS) is a digital object and linked data access and discovery tool. In existence for nearly ten years, the newest version (informed by RCI Curation) has been extended to support complex research data.

• The **California Digital Library’s Online Archive of California** (OAC) provides free public access to detailed descriptions of primary resource collections maintained by contributing institutions throughout California, including the 10 University of California (UC) campuses. The OAC contains more than 20,000 online collection guides and 220,000 digital images and documents.
Research Data Curation Program

Collection: The Brain Observatory

About the Research Data Curation Program

ResearchCyberinfrastructure (RCI) is a UCSD-sponsored program that offers campus researchers facilities, storage, data curation, computing, and networking to facilitate their research using shared cyberinfrastructure services across campus.

http://rci.ucsd.edu/

Browse By

- Collection
- Format
- Topic
As other collections are added, they will be listed here. Cross-collection discoverability is key.

Complex research collections will be “mixed in” with regular digital collections.
Complex components and imagery

Data at Redshift=3.0 (RD0009) Components

- Parameters
  - Simulation parameters
  - Boundary condition parameters
  - Grid hierarchy and metadata
  - Grid metadata
  - Grid hierarchy statistics
- Complete data set
  - RD0009 data set, including parameters and grid data
  - RD0009 data set checksum
- Derived data
  - Halo statistics
    - Summary statistics for all halos
    - Radial profiles of all halos
    - Summary statistics at the virial radius for all halos
    - Sunyaev-Zel’dovich effect decrement for all halos
- Projections
  - Projections of various fields along the x-axis for the back half of simulated field

Processing Information: The project lead collected, on the Triton Resource at the San Diego Supercomputer Center, all data generated by the Santa Fe Light Cone Simulation project deemed essential to representing the simulation project and facilitating reuse of the data. Data files were categorized and arranged to represent each snapshot (Data at redshift) comprising the simulation. The files for each snapshot include files specifying the parameters for each snapshot, binary data files constituting the results of applying the parameters, and derived data products generated from processing of the results. Files deemed irrelevant to representation of the project and / or use of the data were removed from the data set. In addition to data files, scripts necessary for processing the data were added to the collection, as were products generated using the scripts. The former are included in the component labeled “Data Processing Tools,” whereas the latter are typically included in a sub-component labeled “Derived Products” for each of the primary “Data at redshift” components. Finally, a variety of project files, primarily proposals and project status reports, have been incorporated and are listed in the component labeled “Historical documents.” The Santa Fe Light Cone simulation files were then transferred from the SDSC server to the Research Data Storage data storage space. The transfer of all files were monitored for accuracy. The entire collection was arranged into thirty-one components and described completely using the Archivists’ Toolkit application. Component and sub-component descriptions were linked to digital object records composed in the AT and containing links to the files constituting the data set, or snapshot. The AT description was used to generate an Encoded Archival Description (EAD) document for the complete set of files for the Santa Fe Light Cone Simulation project data set and a METS document for each primary component. The EAD is to be uploaded to the Online Archive of California (OAC), whereas the METS records and the digital content files they reference are to be uploaded to the UC San Diego Digital Asset Management System (DAMAS). A researcher will thus be enabled to access the data files through the OAC or the UCSD DAMAS. Finally, all files and descriptive records for the simulation project are to be deposited in the Chronopolis digital preservation network for long-term preservation management.

Rick Wagner, 2012.

PROJECT BACKGROUND

The Santa Fe Light Cone Simulation project was the result of several converging efforts in computational astrophysics, namely: --L7 simulation designed as the final tile; a demonstration of “AMR everywhere” --Halfman computed the necessary size to use only the L7 by stacking state LUSB3D During the 1990s, observational cosmology became “big science” involving expensive instruments (e.g., the Hubble Space Telescope) and large teams (e.g., the Sloan Digital Sky Survey [SDSS]) attacking fundamental questions about the origin and evolution of the universe. Progress was astonishing and included the discovery of the accelerating universe (Riess et al., 1998; Perlmutter et al., 1999); precision measurements of the global geometry, age, and composition of the universe (de Bernardis et al., 2000); and deep images of galaxies at the dawn of time (Beckwith et al., 2004). These and other observations have narrowed the range of acceptable theoretical models for cosmological structure formation to a single model called the concordance model (Bahcall et al., 1995), for which key parameters are now known to high precision (Spergel et al., 2003). Cosmology thus finds itself in a place not unlike particle physics, where the goal going forward is to refine and test the standard model with yet higher precision measurements. Fundamental science questions driving the field include the nature of dark energy and dark matter, the formation and evolution of galaxies and quasars, and how and when the intergalactic medium was re-ionized. Future progress requires ambitious observational
Title:
Projections of various fields along the x-axis for the back half of simulated volume.

Technical Requirements:
Binary data file in HDF5 file format.

From:
Data at Redshift=3.0 (RD0009)

Collection:
Santa Fe Light Cone Simulation research project files

Contributing Institution:
UC San Diego, Research Data Curation Program

Download c8b56kpc-FID28.octet-stream
Download RD0009.x.r2.back.project
“Same Data” presented in different format.
Point Reyes, CA: Landscape Response to Tectonics

DOI: 10.5066/G9NK3BZH
OT Collection ID: OT.052012.28910.1
Dataset Name: Point Reyes, CA: Landscape Response to Tectonics
Short Name: CA09_Morell
Collection Platform: Airborne LiDAR

Dataset Overview:
NCALM Seed PI: Kristin Morell, Penn State University. The survey area is located 26 kilometers northwest of San Francisco, California in the Point Reyes National Seashore. The survey was flown as a part of Seed Money Survey Campaign that took place over two days: Sept 8, 2009 and Sept 9, 2009. LiDAR was used to investigate the landscape response to tectonics by studying the coupled nature of hillslopes during transient adjustment to an increase in uplift rate.

Dataset Acknowledgement:
LiDAR data acquisition and processing completed by the National Center for Airborne Laser Mapping (NCALM) with funding provided by NSF’s Division of Earth Sciences, Instrumentation and Facilities program.
http://dx.doi.org/10.5066/G9NK3BZH

Dataset Keywords: California

Survey Date: 09/08/2009

Data Provider and Roles:
Funder:
- National Science Foundation
Collector:
- National Center for Airborne Laser Mapping

Total LiDAR returns: 523,561,452 pts
Area: 76.31 km²
Shot Density: 6.66 pts/m²

Coordinates System:
Horizontal: UTM Zone 10N NAD83 (CORS96) [EPSG: 26910]
Vertical: NAVD88 (GEOID03) [EPSG: 5703]


Download and Access Products:
Point Cloud Data
- openTopoID: OTLAS.052012.28910.1
- Download and Processing
- Bulk Download (Requires login)

Dataset Extent in KMZ format: Download
OpenTopography

We’re storing all metadata for semantic completeness.

DOI: 10.5066/G9NK3BZH

Data download
After 16 months of work we ...

- Redesigned and re-implemented existing technologies and processes to work with complex research data
- Helped the researchers to better understand data organization and description in order to make them shareable
- Put in place multiple ways to make data discoverable
Findings
Over-arching statements

Considering the data lifecycle holistically is key to the interaction of RCI services

Curation is not solely a technology enterprise

• Human driven
• Judgment needed
• There are social aspects to sharing data outside the research group
The Data Lifecycle

Source: DDI Structural Reform Group, DDI Alliance. “DDI Version 3.0 Conception Model.”
http://www.icpsr.umich.edu/DDI/committee-info/Concept-Model-WD.pdf
Finding 1

Researchers see the data lifecycle as a single workflow.

- All of the researchers asked some version of all of these questions:
  - Where do I put my data?
  - How do I get my data there?
  - How can I access and use my data?
  - How do I analyze and visualize my data?
  - How do I share my data with other people?
  - How do I display my data?
  - How do I reference my data?
  - Who’s going to keep my data after my grant funding ends?

- The needs articulated in the questions above cross all of the RCI service boundaries. Typically, however, these services are not offered in an integrated way.
Finding 2

It is expensive to do curation “after the fact” – i.e., at the end of the data lifecycle.

- The majority of the work involved helping the data owners organize and annotate massive amounts of historical data. Analyzing and organizing data during the collection/creation phase makes using and sharing them subsequently much more effective and efficient for the researchers.

- Researchers said to Research Data Curation Program (RDCP) staff, “I wish we had talked to you sooner.”
The Data Lifecycle

Source: DDI Structural Reform Group, DDI Alliance. “DDI Version 3.0 Conception Model.”
http://www.icpsr.umich.edu/DDI-committee-info/Concept-Model-WD.pdf
Finding 3

There is no standard definition of a dataset.

– What are the units of organization?
– What does collection mean?
– What are the boundaries of a data set?

• Data types, notions of a collection, and how data are organized are not consistent across disciplines.
We do have curation technologies and processes that work with research data:

- Digital Asset Management System (Library)
- Online Archive of California (CDL)
- Library processes for ingest, description, annotation, organization
- Storage (Cloud, project storage at SDSC, Chronopolis preservation)
- DOIs, ARKs (permanent identifiers)

The RDCP successfully used and/or adapted these tools and processes to work with all of the incoming data.
Finding 5

Researchers want tools and best practices to help them manage their data.

- Consistent with Findings 1-4, researchers said that they wanted tools to help with data collection, organization and storage. We heard this regularly from the researchers, including in a cross-pilot gathering.
Finding 6

• There is lack of clarity about:
  – which data are appropriate for long-term stewardship
  – who is responsible for this stewardship
  – who pays to store data for the long-term (10 years+)

• **Who is the steward: The researcher? The University? Who decides?**
Recommendations
Recommendation 1

The RCI Team should actively engage in a discussion of:

– RCI as an integrated suite of data lifecycle services, *and*

– how to present those services more clearly to potential users.
Needs

Create

Preserve

Store

Publish

Analyze

Your data
Solutions

- Tools and services to help you generate and save your work
- Managed, accessible, backed-up storage
- Metadata for sharing
- Best practices and community standards
- DOIs for reference
- UCSD, UC and National Repositories
- Archives for the long term
Create

Store

Analyze

Publish

Preserve

Your data

• TSCC, Colocation, Curation

• Storage, Colocation

• Curation, Storage, Colocation

• Curation

Network is everywhere
Recommendation 2

The campus should create a Data Lifecycle Advisory Council, to include campus administrators, researchers, and librarians. The Council should be tasked with advising the campus on:

- What data the university should steward for the long term.
- Who should pay at each stage of the data lifecycle.
- How intellectual property rights should be determined.
Within the curation program, we will develop and articulate a variety of services from which researchers can pick and choose as needed, to include:

- 3a Consultation
- 3b Preservation Services
- 3c Data Management Services
- 3d Research and development
Recommendation 3a – Consultation Services

• Metadata
  – How to use tools (e.g. Excel, Archivists’ Toolkit) to manage data
  – How to determine and use appropriate metadata schemas
  – How to annotate data properly for future use

• Grant support
  – Descriptions of services for proposals
  – Descriptions of tools for proposals
  – Data management plans for proposals

• Tools
  – Evaluation of existing tools
  – Development of new tools or versions of tools
  – Support for using tools in the curation process

• Getting DOIs for data reference

• “Matchmaking” with a data repository

• Determining proper storage for the life of the data
Recommendation 3b – Preservation Services

Long-term preservation of data in Chronopolis

- Production service run by UCSD
- Geographically distributed and actively managed
- An ingest node for the Digital Preservation Network (DPN)
- Well-positioned to be added as part of a data management plan
Recommendation 3c – Data Management Service

A service to help researchers deposit data into a system where it can be discovered, shared, re-used and preserved for the long term.

– *i.e. what we did with the pilots*
Recommendation 3d – Research and Development

• Continued development of the DMPTool

• Data paper development

• Data input tools:
  – for use in the field
  – for use in campus labs or on the web

• Help drive policy issues related to data

• Engage with national and international data communities
Into the future
Implementation

The UC San Diego Library created the Research Data Curation Program last year with some positions jointly funded by RCI.

We need to work on implementation of services that overlap with other RCI areas and depend on broader collaboration.
We’re hoping to open doors for production services this summer!
Thank you!

David Minor – dminor@ucsd.edu