Thinking outside the repository

Supporting partial curation

Greg Janée

University of California, Santa Barbara

Observation

- Curation is oriented around the concepts of deposits and repositories

- Much work on this paradigm:
  - Abstract models: OAIS
  - Development platforms: Fedora
  - Turnkey solutions: DSpace
  - Communities: Samvera
  - Certifications: CoreTrustSeal

Rest of this talk

1. Three cases that don't fit the deposit-in-repository paradigm
2. The common theme
3. What constitutes "data management"?
4. Proposed data management principles
5. Does this resonate with you?

Case 1: student group project data

- Background
  - environmental science & management school
  - master's students participate in group projects
  - projects gather, generate, refine copious data
  - groups often revisit same area, topic

- Desires
  - archive data out of principle
  - immediate: support data reuse within the school
  - longer-term: make selected data publicly available

Case 1: student group project data

2011

MASTER'S PROJECT	STUDENTS	FACULTY ADVISOR	CLIENT
Ecology and Management of Oak Woodlands on Tejon Ranch: Recommendations for Conserving a Valuable California Ecosystem	Hoagland, Serra Krieger, Andrew Moy, Shannon Shepard, Andy	Frank Davis	Tejon Ranch Conservancy

2012

Developing Fire Management Strategies in Support of Adaptive Management at Tejon Ranch, California	Sean Baumgarten Ashley Gilreath Eleanor Knecht Adam Livingston Nicole Phipps	Frank Davis	Tejon Ranch Conservancy

2014

Developing a wild pig (*Sus scrofa*) management plan for Tejon Ranch	Jocelyn Christie Emily DeMarco Elizabeth Hiroyasu	Naomi Tague	Tejon Ranch Conservancy

2018

Air-Based Land Management – Assessing Drone and Imaging Tools for the Tejon Ranch Conservancy Final Report Brief Poster TEJONDRONES Website	Cheryl Bube, Ellie Campbell, Amanda Kelley, Kalli Kilmer, Jonathan Pham	James Frew	Laura Pavliscak, Tejon Ranch Conservancy

Case 1: student group project data

- Problem
  - 20+ student groups every year
  - insufficient manpower, expertise
  - result: school is just storing whatever the students leave behind

http://www.flickr.com/photos/chiropractic/5565162849/

Case 1: student group project data

- Our approach
  - added DMP requirement to school curriculum
  - offer annual workshop on data management, individual consultations

  - students will identify, document, isolate reusable datasets
    - receive credit for doing so
  - defined metadata standard and format
    - README.txt in markdown
  - developing dataset submission form, harvest script, searchable listing page
    - built on school's Airtable instance

Case 2: museum specimen images

- Background
  - natural history museum catalogs specimens
  - uses Symbiota
  - student workers photograph specimens daily
- Desire
  - preserve full-rez, RAW images
- Problem
  - no means to do so

Case 2: museum specimen images

- Our approach
  - integrate museum's workflow environment with Library-hosted preservation storage
    - "button push" to save images
  - but identification, management, access all managed through museum's Symbiota platform

Case 3: interdisciplinary group's dataset library

- Background
  - large interdisciplinary research group
  - many datasets, many from long-running programs
  - filesystem access
  - made available via FTP and a hodgepodge of bespoke systems

- Desires
  - "preserve this"
  - immediate: facilitate data reuse within the group, by collaborators
  - longer-term: improve access mechanisms

Case 3: interdisciplinary group's dataset library

- Problems:
  - datasets are produced by decades-long, ongoing programs
  - large (TBs)
  - heterogenous
  - continuously growing
  - being actively researched
  - poorly supported

Case 3: interdisciplinary group's dataset library

```
...
2008/
2009/
2010/
2010-2011.Final_Plots/ ← for that one paper...
2011/
2012/
...
DRI_Aerosol/ ← oops, doesn't quite fit the annualized directory pattern
Level2/ ← intermediate, derived products
Programs/ ← complex relationship between data, code
README.txt ← doesn't say what you hope it would
incoming/ ← support for current workflow
```

Case 2: interdisciplinary group's dataset library

- Our approach:
  - considering archiving snapshots
  - seeking a mechanism by which datasets of sufficient maturity:
    - can be described by the researcher
    - vetted by Library curator
    - moved to readonly, preservation storage
    - with filesystem access retained

The common theme

- Library is not depositing data into a repository
  - (its own or other)
  - (not yet, anyway)

- Rather:
  - trying to improve data management within researcher's environment
  - solutions driven by researcher's workflow

The question

- What constitutes data management?
  - What is the end goal?
  - What practices are sufficient?

- Lots of best practice-type guides (DataONE, etc.)

- Looking for FAIR-like principles for data management
  - “FAIR” is catchy, captures end goals of curation
  - these principles would be precursors to FAIR

Data management principles

- Inventoried
  - inventory of datasets exists, and dataset is listed in it
- Owned
  - responsible party is known and documented
- Controlled
  - access control prevents arbitrary modifications
- Clean
  - no extraneous content
- Described
  - metadata, documentation
- Structured
  - uniform structure enforced (if applicable)
- Tracked
  - code versioning, data versions reference code versions
- Protected
  - storage redundancy

Questions for you

- Are you working with data management problems that don't fit the deposit-in-repository paradigm?

- Would a short list of data management principles be helpful?

Thanks

Tom Brittnacher, UC Santa Barbara

Natalie Meyers, University of Notre Dame
Stephen Abrams, Harvard