Data Services: Making It Happen

Heather Coates, Stacy Konkiel, Michael Witt
ACRL 2013: April 12, 2013
Agenda

- Panel Introductions
- Back to the 80s
- The data challenge
- Data Services: In 6 Acts
- Discussion
The data challenge

• The data deluge: Volume, Velocity, Variety
• Privacy & security concerns
• Digital data vs. analog data
• Making sense of it: discovery, access, & reuse
• The case for Open Data
• The Academy responds
Why librarians should get involved

• Trust
• Interdisciplinary/Collaborative
• Existing Infrastructure
  – Preservation
  – Digital Content
It’s about supporting research, at all phases of the life cycle
Data Services: In 6 Parts

- Consultations [Heather]
- Training [Heather]
- Metadata & Documentation [Stacy]
- Preservation [Stacy]
- Institutional Repositories [Mike]
- Data Citation [Mike]
Data Services Routes

Mario Circuit 1
Ghost Valley 2
Rainbow Road
How adventurous do you feel?

Mario Circuit 1: Well-tread, safe, smooth

Ghost Valley 2: Bumpy, guard rails

Rainbow Road: Sharp turns, no rails, hard to see
Data management plans are great!

Ugh, more administrative overhead...
Consulting - Be like Yoshi

• Help overcome obstacles
• Navigate unfamiliar & difficult territory
• Be a good sidekick...but don’t let Mario sacrifice you to win!
What happens during a consult?

• Meeting the funding agency requirement
  – Gather info for draft or review and refine DMP

• Culture change
  – *Nudge* towards institutional cyberinfrastructure
  – *Suggest* better data management practices
  – *Encourage* use of IR for dissemination
  – Navigating the system - referral to other units
Data management plans

- Describing data
- Defining standards
- Intellectual property and rights management
- Licensing
- Archiving and preservation
Strategies for success

A good reputation is more valuable than money.
1522 25 30 32, 12
What path to take?

- Mario Circuit 1: The beaten path
- Ghost Valley 2: Moderately rough terrain
- Rainbow Road: Uncharted territory
Data Management Training
Training: Skills boost

- Identify skills gaps
- Build on existing efforts
- Collaborate
- Relate it back to the mission
- Evaluate & assess
- Revise and refine
What level of training do you offer?

Data management plans & planning

Basic

Moderate

Advanced
What level of training do you offer?

File organization & naming

- Basic
- Moderate
- Advanced
What level of training do you offer?

Backup & storage

Basic

Moderate

Advanced
What level of training do you offer?

Documentation & metadata

- Basic
- Moderate
- Advanced
Training: Various contexts

- Scholarly communication
- Preservation & curation
- Responsible Conduct of Research (RCR)
Training: Strategies

• Make it relevant to expressed/identified needs
• Focus on the practical & concrete
• Demonstrate strategies
• Provide opportunities to practice on examples
• Use real-world datasets
• Target to discipline or research approach
METADATA & DOCUMENTATION: AVOIDING PITFALLS
Things I would rather do than think about scientific metadata:

• Fight a pit of crocodiles

• I’m neutral.

• Nothing! It’s very rewarding.
Metadata & Documentation: Related?

Produced in the course of research

Necessary to understand data, yet...

Documentation is often...

Incomplete or not well understood by others

Not easily associated with data
From Documentation to Metadata

Metadata can either be added...

Manually

Automatically

Most common
Metadata: Questions

Level 1

What is produced automatically?

Level 2

What is easily transformed into relevant standards?

Level 3

What information is needed to open, understand, & work with the data?
Metadata: Pitfalls

- Rejecting researcher metadata
- Requiring too much of researchers
- Not leveraging existing documentation & expertise
- No easy to understand documentation
Find (and use!) the low-hanging fruit

- Data Curation Profiles
- Documentation

Leverage (human) resources

- Metadata and cataloging librarians
- Graduate students
- Subject Librarians

Define a core set of metadata for

- ...your library
- ...widely-used disciplinary metadata standards

Data Curation

- “Full service” option for researchers

Level 1

Level 2

Level 3
Metadata : Requirements

• Data creator name(s)
• Data set title
• File information
  – Format
  – Software required
  – Other technical information
• Methodology
  – How the data was created or collected
Vortex II Forecast Data - forecast_20100615140000Z_run001
Pinfal, Beth; Brewster, Keith; Mattocks, Craig; Bhangale, Ashish; Withana, Eran C.;
Hera, Chathura; Terkorn, Felix; Chandrasekar, Kavitha

URI: http://dx.doi.org/10.5967/M0D21YHV
http://hdl.handle.net/2022/15157
Date: 2010-07-28
Date(s) Covered: 2010-05-15
14:00:00 hours
Mike Rd Claude TX 79019 USA
Methodology: The input data for this forecast includes the following: Rapid Update Cycle (RUC) data downloaded from NOAA with a 13km resolution for forecast date 20100502 at 06Z with data for hourly offsets from 08 to 22. The file format for this input data is grb. The forecast is initialized based on ARPS Data Analysis System (ADAS) Real-time meteorological data assimilation netcdf files with CONUS coverage at 10km resolution produced hourly by CAPS at Oklahoma University that uses the netCDF file format. The data is for 20100502 at 13Z.
File Information: This particular collection contains namelist.input, cape.zip, radar.zip, precip.zip, surface.zip, updraft_helicity.zip, vorticity.zip, xsec.zip, and wfrout_d01_2010-05-02_13_00_00.nc. namelist is configuration file of WRF. cape is short for Convective Available Potential Energy, a measure of the instability in an air mass. cape.zip is the visualization of cape and contains 24 png files. radar is Mix of radar minimum and radar maximum visualizations. radar.zip represents the mixed results of putting those two radar types together. radar.zip is the visualization of vorticity and contains 28 png files. precip is short for Precipitation, the sum of the rain, snow and hail in given in liquid equivalent depth. precip.zip is the visualization of precip and contains 4 png files. surface is meteorological parameters on the earth's surface, or in a model on the first level above the ground. surface.zip is the visualization of surface and contains 16 png files. updraft_helicity is the dot product of the vertical velocity and the vertical vorticity. It is presented as a summation over a 3-km depth. updraft_helicity.zip is the visualization of updraft_helicity and contains 16 png files. vorticity is the localized rotation of the air. In model plots it is often the vertical component of vorticity, the rotation of the horizontal winds. vorticity.zip is the data for each of the three vorticity values.
DATA PRESERVATION: THE GOLDEN RING
Data Preservation & Library Roles

- Dedicated data repository
- Sufficient staffing
- Data curation services

- Leveraged IR & staff
- Consultations & education

- No IR
- Limited/leveraged staff
- Use of 3rd party resources
Our Library Takes The Role of ________ in Data Preservation

- Lead hedgehog!
- Sidekick
- Elusive genius
Data Preservation

• How long?
  – Standards vary from discipline to discipline (and grant-to-grant)

• How much?

• What formats?

• At what cost?
Data Preservation Challenges

• Was upstream stewardship sufficient?
• Custom and proprietary formats
• Preserving data for emeritus faculty
• Digital data more fragile than analog data in most cases
Data Preservation

Level 1
- Refreshing

Level 2
- Replication

Level 3
- Migration

- Emulation

Metadata attachment
Data Preservation Options

- Subject Repos
- Data IRs
- Institutional Repositories

OF QUALITY
Data Preservation: Other Opportunities

- Partnerships with Archives
- Secured physical storage
- Replicating your institutions digital data
1. Refer patrons to data repositories

- Find data repositories in **Databib**
- Many different flavors: publisher (**Dryad**), consortium (**3TU.Datacentrum**), instrument (**CHANDRA**), sub/disciplinary (**RKMP**), country (**Research Data Australia**), institutional (**PURR**), general purpose (**FigShare**), etc.
Have you ever gotten a reference question that involved a data repository?

YES  NO
2. Refactor your existing institutional repository

- Extending your digital document repository to include dataset submissions
- For example, IDEALS at the University of Illinois
3. Deploy a dedicated data repository

- Design and implement a stand-alone institutional data repository
- For example, the Purdue University Research Repository (PURR)
Do you have a data repository at your institution now?

YES  NO
If not, are you planning to implement one?

YES

NO
1. Instruction

Information literacy outreach

libguides
Do you include citing data in your information literacy outreach or libguides?

YES  NO
2. Outreach and Advocacy

- Brochure
- Writing Lab
- Press and publishers
3. Practice what you preach

- Cite your own data
- Suggest citation for data in your libraries
- Include in supporting documentation
- COiNS, embedded RDF, BibTex, etc.
- Show publications that cite datasets
- DataCite: minting DOIs for datasets

www.DataCite.org
Resources


