2011-11-30 - CENDI-NFAIS Workshop - Repositories in Science and Technology

11/30/2011 - Hosted by Library of Congress

Cliff Lynch - Keynote

- Repositories ("reps") gained traction in the 90s
- Open Access movement is one origin of repositories ("author self-archiving") - Put journal articles in the institutional repository
- Disciplinary repositories - Desire to circulate pre-prints as well as open access
- In Europe, when articles were under copyright, authors put citations in the institutional repository
- Coverage at most institutiions is pretty disappointing in absence of mandates; people argue that inst. reps have been a failure
- Other view of Inst Reps: Faculty now producing many artifacts that don't fall into category of text, journal articles. Repositories should provide a safe place for this material: datasets, powerpoints, software, pre-prints, image collections, inputs to research, outputs, stuff used in teaching
- Hard to measure success with this view; very different from goals of open access movement; no one can quantify how much of this material exists; little understanding of life-cycle patterns. when is it appropriate the stuff into inst rep; short time-frame not adequate for measuring success
- Thinking that non-institutional orgs should have repositories too: NGOs, gov agencies, libraries, other non-profits; lot of room for development at this point, not much uptake
- Substantial marketplace has emerged for inst reps: commercial and open source, repositories as a service; barriers are reduced
- Many variations: consortial reps, disciplinary reps.
- We are wrestling with the question of when it makes sense to do things on a disciplinary basis vs institutional basis; real advantages to disciplinary approach (sub-disciplines, special vocabularies, etc); but inst reps are the backstop right now; many disciplines will not be represented by repositories in near future; disciplinary reps tend to want to restrict to limited number of object types;
- Challenges: What sort of metadata should be included? Conflicts between generality (apply to many cases) and specificity (allow easy discovery of specialized resources);
- Metadata demands from libraries have made it a burden for faculty to deposit materials
- Search engines tend to ignore third-party metadata;
- Aggregation of metadata across repositories is important
- Need for sorting through author identity and cleaning up names
- How should inst reps relate to storage of research data; are they appropriate for short-term storage of research data? (research in high perf computing, eg); should there be a staging area before data is preserved? we don't understand these questions very well;
- Challenges for "small data" are just as great as big data
- How do we move from a patchwork of inst and disciplinary reps into a network of repositories? Some faculty have choice of both, shouldn't have to deposit twice; has been a challenging issue to migrate, though; Functional requirements of interoperable reps: extract metadata (there is a standard); make automated deposit through a protocal interface; should be able to copy from one rep to another (different from a deposit? Replication? Object Reuse Exchange Protocol used for this but is complicated);
- Repository discovery and naming: material needs to be accessible in long run; reps should assign unique, persistent IDs; how do you find reps? You want to refer to repository at inst rather than URLs or specific hosts; we need registries of reps, lookup and discovery mechanisms
- Major issue: how do you cite data used in scholarly work, make reference to data in tables, how do we make correspondences between journal articles and data?
- Institutional stewardship is a long-term commitment; they aren't always honored; repositores sometimes go away; stewards need to accommodate that reality
- *Questions*:
- Value of DataNet? Cliff: The projects are capacity building and integrative, linking repositories together, providing tools to act on them;
- Lots of distributed efforts, where should the focus be? Cliff: It's complex: many scholarly communities are stakeholders, public uses the materials; Difficult to pull everything together while respecting specialization; OR conference helps bring people together; DataCite, ORCID are helpful efforts; CNI tries to provide a home, National Academies tries to pull together science communities; international data curation conference has been good venue; Chief Research Officers at universities don't seem to have a meeting like other university execs

Case Studies- Jerry Sheehan

- Delivering Data in Science (March) - in Paris
- #stirepos is today's hashtag

- ***Jane Greeberg - UNC Chapel Hill (Dryad)***
- Dryad is a collaborative, run by a consortium of journals
- Objectives: repository for research underlying peer-reviewed publications in basic and applied sciences
- Partnership with journals, which have a data archiving policy
- Dryad associated with DataOne
- Data built on DSpace; work with @mire, "the company that oversees the DSpace software"
- Federated searching with TreeBASE and KNB LTER

- ***Dr. Ian Bruno - Cambridge Crystallographic Data Cetner***
- study of molecules - use in drug design and development
- 140 industrial subscribers sustain their efforts
- Sustainability is still an issue; big pharma has been impacted financially; have competition with commercial apps
- Fuzziness over where ownership rests
- Value is added for subscription service; resentment that data is not open

- ***H.K. Ramapriyan, Earth Science Data and Information Systems Project, NASA Goddard Space Flight Center - EOSDIS***
- Earth observing satellites and earth science measurements
- Mission is to meet the challenges of climate and environmental change
- Data is available at no cost; EOSDIS provides data processing, management, interoperable data archives
- Satellite data is captured by flight operations, processed, sent to multiple data centers
- Other sources of data too; multiple types of instruments
- Middleware and associated clients provide search and access to data across al data centers
- Distributed data centers handle different types of data (e.g., National Snow & Ice Data Center)

- There is global directory; all datasets are discoverable; cross data center searches through REVERB
- Many data visualization and analysis tools
- 5.1 petabytes of data

- ***DSpace@MIT***

- ***PubMed Central***
- Electronic extension of NLM's print journal archive
- Free access;
- Deposit Paths: publisher sends final article in XML or author sends manuscript file, it's processed and NIH creates XML, then deposited
- NLM has formal agreements with publishers (final copy, deposits are permanent, publishers can't withdaw content
- They have non-exclusive license to use the content; they don't own it
- Author must retain rights to manuscript before signing publication agreement
- PubMed DTD now a NISO standard

- ***Library of Congress***
- LOC is a holder of large datasets that are used in research (e.g., Twitter)
- Mandatory Copyright Deposit now bringing in many new files
- So far their system is discovery and delivery; lacking many repository features

- ***DataCite and EZID***
- Creates a global citation framework for data
- Uses DLI (Digital Logic Identifier)
- Take a lifecycle approach www.cdlib.org
- UC3DCXL - open source add-in for Microsoft Excel as a data collection tool
- n2t.net/ezid (create an ID)

- ***ORCID - Brian Wilson, Thomson Reuters***
- Open Researcher and Contributor ID
- Allows reliable attribution of authors and contributors
- ORCID allows you to create a profile associated with your ID
- 282 participating orgs internationally; academic, publishers, government, societies, non-profits, etc.
- Mellon granted award to MIT, Harwaverd, Cornell to study ORCID business models
- VIVO awarded grant to ORCID for collaborative research
- Just released first code
- Will hire executive director and technical director
- Institutional seeding of profiles, delegated management
- To use OAuth (used by Google, Facebook)
- self asserted, socially validated, organizationally asserted identity = more credible assertion

- ***Chris Greer - NIST***
- Promte infrastructures as well as standards; consider themselves part of the "data community"
- Missing: a framework for the community to make decisions;
- Draws comparison to NISTs mandate to design interoperable smart grid; information requirements are similar; many different stakeholders; Smart Grid Interoperability Panel - consensus based organization; 724 members; architecture committee, testing and validation, security; stakeholders include standards bodies, regulators; participation is voluntary, but you must participate -- miss sequential mtgs or votes and you're voted off the island
- If the data community did this, NIST would be the convener, would have White House support
-