2011-11-09 - DTR Kickoff Meeting

Set I From Bill Branan:

Attendees

Madelyn Wessel (General Counsel for UVa)
Micah Altman (Harvard)
Gail Steinhart (Cornell)
Tim DiLauro (Johns Hopkins)
Brad McLean (DuraSpace)
Thorny Staples (Smithsonian)
Mark Leggot (Library director at UPEI, PI on Islandora, President on Discovery Garden)
Terry Reese (Oregon State - preservation of research data)
Brian Westra (Univ of Oregon)
Bill Branan (DuraSpace)
Steve Gass (MIT) - developing services for supporting
Susan Parham (Ga Tech) - IT goverance, repository
James Yoon (Fluid, Toronto)
Dan Davis (DuraSpace)
Andrew Woods (DuraSpace)
Mary McEniry (Researcher at Univ of Michigan, ICPSR)
Geneva Henry (Rice)
Mike Wright (NCAR in Boulder, Co)
Karim Boughida (George Washington Univ)
Kim Thanos - (Thanos) Facilitating
Jonathan Markow (DuraSpace)

Needs, Challenges

Rice (Geneva Henry)

- Only archived one dataset, in the institutional repository, because of requirement by PLoS for publishing
- Most research is stored locally on hard drives and flash drives, long term availability and use is not really a concern
- Hard to know when it's important to put data in long term storage, don't want to be hindered while working with data
- Researchers not thinking about cite-ability
- Researchers mostly care about local access (backup an afterthought)

Johns Hopkins (Tim DiLauro)

- Seeing a huge range of behaviors across researchers and across disciplines
- Data management plan is a requirement of NSF grant, so that's a good opportunity to discuss
- What is the right data archive for data - want to use a natural home where it makes sense
- 2 full time data management consultants, high engagement
- Doing a lot of outreach to encourage faculty/staff to get in touch for help with their data management plan
- JHU has a high level of control
- Have found that storage management is difficult
- Some researchers want to have their own equipment for data management, trying to convince these folks to push their content into the archive sooner rather than later

Univ of Michigan (Mary McEniry - Research Scientist)

- Very concerned about data being secure and confidential
- No one has ever asked her for a data management plan, but wants to make sure data is preserved
- Big question mark about storing data in a cloud (concerned about confidentiality, security, ensuring data persists)
- Doesn't want to be worried/concerned about how data is managed - wants to just trust whoever is doing the archiving
- Concerned about the possibility of cloud provider being able to provide access to data if served a subpena
- Doesn't think about archiving of data until the end of the project

MIT (Steve Gass)

- DSpace is very limited in its capabilities for providing for the needs of real data management
- Challenge: figuring out the roles and responsibilities in the institution for data management
- There is a healthy scepticism among faculty regarding the capabilities of central IT
- Need a staging repository for in-work data that can be easily transitioned into the archive
- How to subsidize metadata creation
- Need to build a system which solves problems for researchers rather than just taking more of their time
- Difficulty is the funding environment, can't guarantee funding will contine to be available regardless of what the

Univ of Oregon (Brian Westra)

- Policies are useless without the services/capabilities to actually comply
- Need to be able to organize and categorize data on creation
- Challenge in adoption rate
- Metadata need to be captured up front, and tools need to make that easy
- How to know how long to keep data

Oregon State Univ (Terry Reese)

- One challenge is that much of the research done is not exclusive to OSU, it's a multi-institution effort, so who has the responsibility to archive the research products
- Library doesn't really have the funding to provide a general archive for research data

NCAR/UCAR (Mike Wright)

- NCAR is NSF-funded institution for atmospheric science
- Tends to deal well with large data sets, but doesn't handle small data sets very well
- People expecting to be able to have tools to work with data directly from the archive (rather than just downloading files)
- People want to be able to collaborate with others using the data in the archive

Cornell University (Gail Steinhart)

- Building research data management data group to help define services for research data preservation
- Using repository system for archiving, works well for some data sets, not so well for others
- IR not as good for access and re-use of research data
- Need help to support collaboration, about to sign a contract with box.net
- data curation profiles.org out of Purdue would be a good resource
- researchers just don't know what they need to do with data
- grad students do a lot of good work, but they come and go, so their data needs to be captured
- what the researcher expects as far as data outputs are not necessarily what they end up with
- researchers want to share, but they generally want to be able to have a conversation with so that people understand the dataset

George Washington Univ (Karim Boughida)

- GW is moving from teaching/learning university to research university
- Researchers concerned about data privacy
- Don't have the resources to handle large data sets (example, 1 PhD student with 30TB of data)
- Important to consider DTR from the beginning of the workflow, capture data, document all changes

Comments from others

- No institution has yet defined a comprehensive set of data manangement pratices
- Harvard: There are two different systems for generating/manipulating data and preserving data, and there is a level of effort to move from one to the other
  - How can we leverage existing systems?
  - The size, scope, characteristics of data being gathered/captured is changing rapidly
- Thorny: Need to get away from traditional library practices
  - Need to talk about sustaining data not archiving data (the idea of archiving data is getting in the way)
  - Researcher should have ability to control access until they hand it over to the institution
  - You can't take an artifact out of context and say that it's preserved
- Mark Leggot
  - Started by thinking about the "research data lifecycle" and handling each part
  - Need to steward the data all the way through, not just take it after being thrown over the wall
- Ga Tech
  - Faculty members have said that metadata (and even data) is a grad student concern
  - Where to responsibilities overlap between institutions, organizations
  - Researchers are thinking in terms of "archiving data once the research is done"
- James (Fluid)
  - Would be interested to hear the needs of the consumers of archived research data
  - Tim: it's a hard problem to know how to best provide access to data in the most useful ways
    - It's useful to start by asking researchers who else may want to use their data, but that is a very limited viewpoint
    - It's hard to get failed projects published and available
- It's dangerous to try to cram too much into the IR/archive
- Question is raised: Who owns the data, the researcher or the institution (department)
- One researcher noted: I want to have my data on my laptop

Notes:

- Researchers need immediate access to their data while research is ongoing, we can't slow them down
- Ideally, the archive would be able to serve the operational needs of the researcher
- There may be a need to annonymize data before it can be made publicly available
- Becoming clear that encryption is necessary
- Appears that the idea of creating a data management plan for research data is really catching on
- Institutions need a way to help researchers actually meet their plan
- Researchers tend to want to share data management plans
- Researchers need to see the direct benefits up front
  - Being able to simply organize data in a useful way is a good start
- There is a big distinction between the metadata that the researcher needs to do their work and the metadata that the cataloger generates after the fact
- Should be able to pull together tools that already exist
- What is the greater need: Handling existing data from research that is already done vs Setting up for the next round of research?

Priority items:

1. Need for a nimble, short-term, durable storage solution while research is underway. Create a balance of local storage for data management and central archive for long-term access and preservation

- Connect the operational and archival phases of the data life cycle

2. Creation of workflow processes across the data management life cycle -> Should allow the researcher to follow either the same or a simpler workflow but at the same time captures the information needed to steward the data in the long term

- Example: User creates a file using an existing tool, our system notices that the file was created and captures the file along with all available information (the checksum, the user, the file size, the creation date, the file name, the directory in which it was stored, etc)
- Fronend workflow: human driven, peer review, access control
- Backend workflow: making data richer, more accessable, more useful
- workflow for the user should be as similar as possible to current processes

3. Confidentiality of data in the cloud (security, data privacy, predictability)

- Auditability is important
- Specification of locale
- Need to ensure that it's clear that ownership does not change hands
- Concerns about access
- If system failed, would it be possible to get the data out
- No gaurantee from storage providers that they won't lose their data or access your data
- It would be very helpful to researchers if we were to define/describe what the cloud is

4. Automation of basic metadata creation and catalogue creation

- Researcher benefits:
  - helps them to find their data
  - not required to provide much if any information
- Tie level of capability/service to a greater amount of user-supplied metadata

5. Interoperability of archiving solutions with discovery systems used by specific research communities

- Tim: should include the concept of published or not published
- Would be useful to have a registry for tools that can be plugged into DTR, particularly for feature extraction tools (this is in-work at JH, Data Conservancy)
- Should consider including researcher ID (ORCID) as a part of the metadata captured (or provided by the user on signup)
- Should provide a way to cite information stored via DTR

Need to make sure the system provides a clear value-add for each stakeholder group

- Key Stakeholders:

Researchers should want this because:

- It's helping to relieve administrative burden
- It helps them comply with institutional and funding requirements
- It protects against data loss

Should perhaps think about data as a link-able resource rather than just a file

Will be important how we market this

- People want to know this is a professional service
- Steve: avoid the term "records management" - researchers will see it as red tape

Should expect the least amount of work possible from the researcher for setting up and configuring this system

Tim: DTR is likely an operational environment, that helps to facilitate the transfer of data into an archival system
Steve: Don't want to rule out the idea of DTR being a shared archive. Many communities don't have shared solutions and could benefit from a commonly available solution

Next Steps - How to move forward

- Looking to partner with 3-4 institutions, for piloting, in the first iteration
- Likely to push through 3 prototypes, with production deliverable at the end of 2012
- Challenge to get any time from researchers, especially a third party
- Hard to line up time frames that fit for both our team and research teams
- Could provide an incentive to researchers (we will store your data) to encourage their involvement
- We may get some political bad feelings from IT centers
- May help to have DuraCloud running locally at the institution
- Need to have a package that is more attractive than Amazon
- Lots of use of dropbox, box.net in research teams now
- We could potentially get a set of researchers who aren't doing active research, but have experience and time
- We should look to smaller institutions or humanities departments that don't have enough IT support
- Thorny: Look to archeologists or other similar
- Consider looking to researchers who consider digital preservation their research area or topic
- Consider working with grad students

- Steve: Try to articulate expectations for timeline, involvement, commitment and what is being offered: so that those in the room can point their potentially interested researchers at it

Set II From Brad McLean:

DuraCloud Direct To Researcher Meeting
Roslyn, VA
Nov 9, 2011

Introductions

- Jonathan Markow - DuraSpace (Is DTR the right name? Not making assumptions yet)
- Madelyn Wessel - UVA general cousel
- Micah Altman - Harvard qualitative data sciences - preservation (dataverse)
- Gail Steinhart - Cornell librarian - (NSF data managment popular)
- Tim Dilauro - libraries (data conservancy), dig research and curation
- Bradley McLean - Coordination across Duraspace projects
- Thorny Staples - Smithsonian, offices of research data services
  (small science data currently all under desks)
- Mark Leggott - UPEI lib director; PI Islandora; DG CEO
- Terry Reese - Oregon State; (chair) tech/lib intersection data research pres.
- Brian Westra - Oregon state; librarian, working w/ faculty.
- Bill Branan - Duracloud developer
- Steve Gass - MIT; assoc dir research and inst services; research data management team - nascent services for MIT
- Susan Parham - GATech - data curation program. Faculty and IT
  interface, repository
- James Yoon - Okage U canada; Fluid project; UIs for web services
- Dan Davis - Data conservancy and Duracloud
- Andrew Woods - Duracloud developer
- Mary McEniry - researcher, demography archive at ICPSR
- Geneva Henry - Rice Univ; dir center for dig scholarship; working w/ faculty around NSF management mandate, and office of sponsored research.
- Mike Wright - Natl's center for atmos research; director; large data center, issues w/ small science - work closely w/ universities, NCAR provides services to them.
- Kim Thanos - Thanos partners - mediation / facilitation.
- Karim Bhougida - Geo Wash U - univ librarian.

Kim- Morning about sharing needs and challenges around current state.
Slides from groups.

Current State: Progress and Needs

Slides given:

- Geneva (Rice)

-
    - Only one dataset - driven by PLoS requirements
    - Talking to high energy physics and astrophysics people - model "ideal" because everything handled by CERN - works for Big Data; however lab notebooks not being managed. CERN now managing digital lab notebooks.
    - Nanotech: Currently small data is stored locally on drives; flash drives; long term availability and reuse isn't considered (yet)
    - Bioeng / biosciences similar: lot of data sets being generated; hard to figure out when to capture the data sets from the equipment; movement of the data to/from HPC is important. Compliance (patient data) is a key concern in life sciences. Researchers not yet thinking about citations or data ownership.
    - Computational & applied mathematics (CAAM) similar.

- Tim Dilauro (JHU)

-
    - Engaging specifically on NSF proposals - 50% response rate on new proposals; little consistency even within disciplines - range of behaviors.
    - Dealing with multiple disciplines; services offered at proposal stage to write the DMP (what data, lifetime, privacy, security, embargo; requirements on sharing), also focus on identifying the correct archive for the data - get it to it's natural home as well as the JHU archive.
    - Services also provided during the award phase from the JHU data archive - deeper engagement on the DMP (beyond the two page limit); funded from the research project.
    - Two full time data management consultants - Tim is a backup to them.
    - DC component based design; well defined APIs; two implementations of archival storage components (ELM - file based; Fedora based).
    - Challenges - not done with system; incremental implementation; working on user support (users placing own content in and setting policies)
    - Working to figure out which properties are common across domains, which vary, which unique.
    - Some issues with HSM system. Looking into how to deal with storage vendors.
    - Focus on engagement - the process has been key.
    - Trying to get researchers to push data sooner, still keep own
      operational environments.
    - "Sideband data access" - pointers to external sources that can allow authorized users to get to items. Examples: clips from large videos; running queries on archived databases.

- Mary McEniry (U Mich)

-
    - Research across 20 countries and 144k adults
    - Cloud Computing - how do you handle confidentiality? Can't physically verify the data.
    - ICPSR testing clouds for backups.
        - NB. NYT Thomas Friedman articles on cloud computing
        - Question of how much protection you would get in the cloud against a subpoena?
        - (Discussion of encryption, key escrow)
        - Dan: What is an operational system, what's an archive, what's in between? Discussion
- Steve Gass (MIT)
    - Control of research data decentralized; sits with PIs and their research groups. Managed locally in each project.
    - Office of sponsored programs assists with the grant process.
    - Have a Research Data Management Team; slowly reaching out to researchers
    - DSpace not a data repository, so few use cases are met.
    - Challenge is figuring out rights and responsibilities across data lifecycle.
    - Challenges: Rocky relationship with IT; skepticism from faculty; need to be more involved in the grant application process. How to subsidize the consultation with faculty and the metadata creation.
    - Discussion: Data management plans are forcing the conversation - "tilt the floor to herd cats" - note the "DMP creation tool".
    - Difficult to understand the correct interpretation of the NSF policies
    - Valuable to have the institution set its own policies
    - Need to have resources match the policy (keep for 3 years = infrastructure to do it)
- Brian Westra (U Oregon)
    - More "fragmented" than "decentralized" - not organized.
    - Need to manage data at the point of collection - need better organizational tools.
    - Working on ELNs w/ chem faculty. Interpretation of NSF requirements might be publication of entire set of notebooks.
    - Using OMERO (open microscopy project)
    - Trust relationship is key; needs assessment process helped with this.
    - Faculty don't know what metadata is, and they don't have tooling!
- Terry Reese (OSU)
    - 3 years started talking to faculty and deans on campus about data mgmt. When library did IR, talked to researchers.
        - Central IT is much smaller than other departments (8th or 9th in size). Departments already thought through their own plans;
        - College of oceanography has great system from Microsoft partnership; departments not collaborating with each other.
    - College for sponsored research took lead on NSF, not the library.
        - Library does ad hoc consultation. Most work within departments.
        - Library trying to figure out what role to play. Easy relationship with the grad school (thesis).
        - Departments interested in having library pick up the materials after five years of own curation.
    - Much of the data created isn't just at OSU; not clear what OSU's role or responsibility is in maintaining access to the research. Library probably won't be responsible for data, will likely concentrate on the interconnections.
- Mike Wright (NCAR)
    - Fed Funded Research Data Center
    - Data required to be sent up to national data centers - not always happening.
    - Data management policies and formats are a challenge
    - Large data sets working well, but small scale/small data is not well served.
    - Working on data citation and recognition of data managers as a role.
    - Budget question about balance between preservation and new research.
    - Have a group that actively looks for data about to be abandoned to bring in and archive.
    - Looking to have tools and services that work with the data - beyond a simple download - apply an analysis tool to the data in situ.
    - Last week workshop NSF / OCI on earthcube - anticipated future massive project.
- Gail Steinhart (Cornell)
    - Forming a Research Data Management Service Group
    - Reached 250 researchers in sessions, 40 consultations so far.
    - Help researchers locate appropriate repositories (both internal and external).
    - Internal IR not well matched to how researchers generate data.
    - Examples, annual data set updates where visibility of old data needs to be controlled.
    - Developing archival repo for larger data sets, moving from planning as dark to public access.
    - Launched redcloud - on demand data services.
    - Contract w/ box.net coming.
    - looking to expand viva? vivo? application
    - Datacurationprofiles.org
    - NB: Lots of sharing of DMPs across researchers / attempts to write it once for life.
    - Looking for ways to charge up front.
    - Env science want to share the data, but want control / correct understanding of the data set - looking for conversations before sharing
- Karim Bhougida (GWU)
    - Moving from teaching/learning to research univ.
    - In pilot projects, Researchers didn't see the value.
    - Concerns: access control; turf
    - Three areas working:
    - Crash test data (terabytes of data)
    - Anthropology field notes (scanning)
    - General
        - Insufficient resources for sizes of data sets that would like to be stored.
        - Challenges in being involved in DMP for researchers that are using external repos.
    - NB. DMP for certain disciplines is "I don't produce data".
    - Workflow management

(List of discussion points captured from the presentations to drive conversation - please add any others)

Heard: Everyone is trying to lead change at their institution; will try to provide feedback for institutions as well is to DuraSpace.

- Madelyn Wessel (UVA)
    - Glad to hear about compliance concerns - and this is not just researcher but also institutional responsibility.
    - Not clear that there is a comprehensive, clear inst policy on data management policy.
    - Grateful for funders that impose open access/data mandates. Funder push towards sharing is tilting the floor correctly.
- Micah Altman (Harvard)
    - Proactive w/ dataverse network. Researchers manage own virtual data archives.
    - Gaps - research support making researchers move data in and out of HPC systems, etc.
    - Themes:
        - Consistent policy framework for data management plans - what do we mean
        - Managing to make research success
        - Managing to contraints like privacy.
        - Short term access in Disclipline
        - Long term access for broader or interdiscipliary use.
    - Roles and stakeholders:
        - Libraries
        - Discp archives
        - Funders
        - IRs
        - Journals "supplementary material"
        - ??
    - Metrics to show usage for management
    - Data management service is a key value in itself.
    - Researcher perspective
        - I have to write DMP
        - Size, frequency, updates; changing rapidly in social sciences
        - Confidential / private information in social sciences under dozens of legal regimes; could be getting worse.
    - HIPAA would like to reach out farther.
- Thorny Staples (Smithsonian)
    - Large scale projects since 1992
    - Problem: Need to get away from library ideas of collections and artifacts.
    - Should talk to the researchers at the start of the project, and have them capture the metadata that is useful to the researcher's organization, not to the library.
    - Need to build systems that support this.
    - We are building networks of information, each institution should have it's own corner of the network, but not try to own collections.
    - Can't take artifact out of context and think you've captured the research. You have to keep the scholarship links to have the value.
- Mark Leggott (UPEI)
    - Started stewarding research data 5 years ago as a sustainability action, not a mandate.
    - Started with the research data lifecycle; manage everything created all the way through.
    - Initializing (grant)
    - Analyze and collect (get data)
    - Reporting (posters, etc)
    - Formalizing (papers)
    - Popularizing
    - Visualizations should be as easy as trip-it works.
- Susan Parham (GATech)
    - Lots of interviews with researchers and assessment
    - Don't want to know about metadata - Grad students do that, and it would be micromanaging.
    - **"Everything related to the data - go talk to the grad students"**
    - Not clear reponsibility boundaries between institution / government agency / department / researchers.
    - Urban legends about how long to keep data, what the library will provide, what the repository can do.
- James Yoon (Fluid)
    - Systems built for the end users.
    - Heard domain goals: Want to save data for the future - what are the needs of the user on the receiving end.
        - (Tim / Rumsfeld) "known knowns, known unknowns, unknown unknowns"
    - Storage
    - Archives (replication, services, etc)
    - Preservation (migrate into new environments)
    - Curation (make content usable to community over time) OAIS has process around this.
- Discussion of Publishing Failures
    - Can do this
    - Should do this via data
- UVA is starting a section for rejected papers.
    - Mathematica rejected papers.
    - Important to colleagues.
    - Convince people not to distinguish between published and unpublished data
- Thorny: Need to help the organization of the data as captured.
    - Capture their SPSS or R settings as metadata.
    - Mark: User's profile is critical to metadata about the objects.
    - Thorny: distinction between broad general metadata done by catalogers vs. researcher's immediate needs.
    - Brian: Looking to ELNs to help capture, but metadata varies.
    - Mark: VRE example: Started with chem models; as ingested, pulls in related papers. Adds huge benefit.
    - Mark: Key is in pulling the existing tools together, but creating them; each discipline already has the tools.
- Brad
    - Feedback loop to researcher is critical
    - Bad data mandate
    - Electronic lab notebooks - standards for exporting.
    - Mike ELN is a metaphor - need tooling to support this.

- Brian - computational workflow - can we record it? Save the steps, not the output; will "Galaxy" do this.
- Tim - lots of workflow systems; challenge is integration and capture of all the systems. Big challenges for archives -complete or visible failure is not an option; building too much into the archive is risky. Need to cast this as an ecosystem with the archive as part of it.
- Bill
    - Distinction between research completed vs yet to be completed.

Discussion

- Discussion about mandates to place data or papers within one or two years. Challenges of when to have transparency.

- Madelyn: Need to figure out what we think data is, and what the norms are within each discipline.

- Geneva: Must have discussion about the issues up front. Faculty ownership is key.

- Brian: Research Lab notebook system must synch to his laptop all the time.

Prioritizing Needs and Requirements

- Researcher data to the archive discussion

- Workflow processes discussion

- Are workflows those of DTR? Of external tools, captured by DTR? Driven by DTR?

- On connecting DTR to discoverability:
    - schema to schema engine. Map to DTR standard and spit out in many schemas. (XML/RDF schemas).
    - N.B. Standard scholar metadata (orcids, researcher ids)
- Should DTR support datacite or doi? Should be agnostic, support data citation in general.
- Should DTR work with datacite to get a marker on the data as soon as it gets to DTR. Or create on demand "I need to cite this data".

Working Agreements / Next Steps

Jonathan - How are we (duraspace) going to engage with you (institutions) during the pilot phases of this project. Would like thoughts on what you need to know to decide to commit?

DuraSpace will take the output of this meeting and try to scope based on the features and ideas from the session, including input from our advisory team member. We'll report back on our immediate path, and looks for additional feedback. We are looking to partner with 3-4 institutions. Timespan is to complete by end of 2012.

Researcher time issue? Hard to engage researchers in prolonged prototype.

We anticipate a series of short engagements at three month intervals.
(Tim) difficult to get the time and the ongoing engagement from the researchers.
(Steve) time alignment around availability of researchers are likely to be problematic.
(Susan) Provided data storage at no cost as an incentive to have the researchers participate.
(Geneva) Hard to get time at tier 1 inst; need to have trust established first.
(Brian) Will need to have someone (library) on campus to mediate the relationship, with (grad students, not researchers). Can't get the researchers to come to demos of ELN.
(Mark) DuraCloud private vs public: use the private downloaded version as a way to get DTR in the door. (Local storage plus preservation copy)
(All) Many users using dropbox and box.net. Note, in use not for all data, but for some that they are sharing.
(Karim) Find past researchers who might be interested in giving feedback.
(Tim) Have an end-of-career researcher who is approaching the DC - a possible early user.
(Thorny) Smaller institutions interested in providing services might be a good place to find researchers. (William and Mary, for example).
Chase the humanities. Go for well known small liberal arts colleges.
(Gail) Ensure there is an exit strategy at the end
(Tim) Thinking about zythos; if you have dropbox front-end, and transparent deposit.
(Brian) Regional univs w/ masters research programs
(Thorny) Try the archaeologists.
(Terry) Also, look at satellite campuses.
(Susan) Dig pres researchers.
(Thorny) Fields with complexity problems rather than scale problems.
(Karim) Public Policy as a target area.
(Mark) Target the adjunct faculty.
(James) Try to bring the participants in as co-designers as well as users.

(Mark) Adjuncts and grad students a good target as they don't have existing resources to push data to.

(Steve) Help for data that supports dissertations; currently data isn't part of the record in DSpace. Strategy that works with today's grad students is a good idea.

(Steve) Advice: Articulate clearly and in detail what duraspace expects in terms of involvement, committment and timeline; and what is the offering E.G. Help them make the pitch to their faculty.