DuraSpace DTR Workshop
Meeting Results
November 9, 2011

Overview

The DTR Workshop convened to identify the highest priority needs for DuraSpace's Direct to Researcher (DTR) system development.

During the workshop, participants each presented key institutional successes, challenges and priorities in developing and implementing data management strategies for research. They then refined and prioritized those needs to identify those that are most appropriate and necessary for the DTR system.

This report summarizes the priorities. The complete electronic brainstorming and prioritization process is included in Appendix A. Appendix B lists participants and contributors.

DTR System Priorities

Participants identified five priorities for DTR:

1. Provide a short-term storage solution while research is underway. Connect the operational and archival phases of the data management lifecycle.
2. Create simple workflows across the data management lifecycle that capture meta-data and provenance.
3. Provide confidentiality, security, privacy, and predictability of data in the cloud.
4. Automate basic metadata creation and catalogue creation.
5. Create interoperability of archiving solutions with discovery systems used by specific research communities.

All of the priorities emphasize the importance of supporting the full data lifecycle and capturing the metadata that allows for effective discovery, use and cataloguing.

Priority One: Provide a short-term storage solution while research is underway. Connect the operational and archival phases of the data management lifecycle.

The issues driving this need include:

- Operational data management is not effectively linked to the data archival processes.
- There is significant risk to data during the operational phase of research. Data can be lost due to physical disaster, human error or natural turnover among graduate students working on the research project.

- The metadata and cataloguing that will be required for archival is best completed during the research process, with a reduced administrative burden.
- Effective short-term data management processes improve research efficiency in allowing researchers to more easily find data and information.

For the researcher this will assist in complying with an institutional and funding requirement. It will also reduce the administrative burden and prevent data loss or re-work when a graduate student leaves the project. It is critical to the sustainability of the research project.

The specific requirements and suggestions brainstormed by participants include, in no particular order:

1. Keep clear the distinction between technical or automated metadata and human-created metadata.
2. Adding materials into the archive and into the workflow has to be easy. It cannot require researchers to jump through pages and pages of input.
3. The data repository must be agile, active, available, sustainable, and interoperable with any preservation-focused archive.
4. Create a "share with my colleagues" button that allows me to add colleagues mail addresses to send them a link to access the data.
5. Integrate with institutional identity providers.
6. Add file/resource-level annotations.
7. Ease of setting private/public sharing and timing flags.
8. Make it easy for researchers to transform data for preservation.
9. Make the data accessible with existing tools and paradigms.
10. Create easy upload and synch of data. A simple form with 2 – 3 fields of metadata that allow a researcher to upload a batch of data files.
11. Include 'quiet' metadata capture without requiring a lot of user interaction.

Priority Two: Create simple workflows across the data management lifecycle that capture metadata and provenance.

The issues driving this need include:

- Researchers do not take the time to create provenance and metadata. Whenever possible this should be done on behalf of the researcher with the system collecting information based on the workflow of the data.
- Graduate students are often charged with creating metadata and they may move on before the research project is complete, leaving gaps that are difficult to fill after the fact.
- As data management requirements increase, largely without associated funding increases, it becomes increasingly important to streamline data

management processes and reduce administrative burdens on the researchers.

For the researcher this will prevent extra work that does not add value to the research but is required for archival purposes. It will also improve the ease with which discovery of the work can occur, and the ease with which the researcher can locate research files during the research process.

The specific requirements and suggestions brainstormed by participants include, in no particular order:

1. Capture provenance as the system completes.
2. Cannot break existing systems and processes.
3. Must address both back-end needs to make data richer and more accessible, and front-end needs to ease access and peer review.
4. Create archiving workflows: characterize and validate data, identify triggers to move from one environment to another, systemize processes, and create metadata and persistent identifiers.
5. Clarify distinctions between workflow types. Capture external workflows that occur, backend workflows that occur within DTR as part of normal operation, user specified workflows within DTR, and larger context workflows for the user in which DTR is only a subcomponent.
6. Do not overwhelm the researcher with metadata requirements.
7. Create workflows to put the data into Excel and then manipulate.
8. This is thinking long-term, but it would be nice for the workflow tool to be already thinking about how data may need to be formatted during its lifecycle.
9. Remember to use what can be divined of the Personal Space and Object Space to make the researcher's job easier
10. Define human and automated workflows. Test with user-researchers' requirements to better scope. Make no assumptions. Make sure to capture what researchers want.
11. Have simple role-based functions: researcher, data steward and other.
12. Allow workflows to capture the state of data analysis. These will be application specific, so they probably need to be agnostic to the workflows themselves.
13. Add value to my research
14. What is the researcher's role in designating data as worthy of preservation in a particular state? Do they want to keep everything?
15. Create different user roles
16. Clarify distinctions between workflow types. Capture external workflows that occur, backend workflows that occur within DTR as part of normal operation, user specified workflows within DTR, and larger context workflows for the user in which DTR is only a subcomponent.
17. Conform to policies that have been defined for each stage of research.

18. Tightly align benefits and effort to use. Don't interfere with the way I already do my work. This should only require incremental changes/cost to gain benefits. This should require no changes to just use without gaining benefits.
19. Workflow should capture the state of data analysis. These will be application specific, so probably need to be agnostic to the workflows themselves.
20. Workflow should be able to identify lifecycle/span of the resource.
21. Offer interfaces to expose versions/provenance/metadata to workflow systems
22. Allow interoperability with other tools
23. Do not break existing researcher workflow (or quasi workflow) tools.

Priority Three: Provide confidentiality, security, privacy, and predictability of data in the cloud.

The issues driving this need include:

- Requirements by the institution and funding bodies to adhere to strict practices and requirements.
- A need to address perceptions that data stored off-campus is less reliable and secure.

For the researcher this will address core concerns of control and trust. It will also allow researchers to take advantage of services such as box.net and dropbox that many are already using today in a less structured and managed way.

The specific requirements and suggestions brainstormed by participants include, in no particular order:

1. Address specifications of locale (such as requirements that the data does not leave the US, or does not enter the US). This may lends itself to creation of a public/private cloud instance such as Amazon and MIT.
2. Define your cloud-based approach. Are you handing off to a third party without the ability to physical and other audit?
3. Clarify in terms of service that institutions own research data.
4. Ensure the ability to maintain ownership relationship of data.
5. Provide terms of access and SLA's for uptime and availability. These will be critical in evaluating the service.
6. Negotiate the ability to remove data in case of a failed business, etc.
7. Publish terms of use, service, SLA's, remedies, and protection against subpoenas.
8. Publish processes for certifications and audits.
9. Clarify where extension of institutional protections applies to a 3rd party provider, and where it does not.
10. Define who has access to your content? Create assurances for data loss.

11. Define who has administrative responsibility for the different layers of the system?
12. Control and trust seem to be the underlying issue.

Participants questioned the name, “Direct to Researcher” as it implies an absence of institutional engagement and oversight that is inaccurate and not necessarily positive.

Priority Four: Automate basic metadata creation and catalogue creation.

The issues driving this need include:

- There is a distinction between the metadata that is immediately useful to the owner of the data (the researcher), and the metadata that is traditionally created by catalogers and used by archive or library systems. It is this "owner-userful” metadata that we need to encourage; and it becomes an input to the eventual archival process.
- Researchers do not take the time to create provenance and metadata. Whenever possible this should be done on behalf of the researcher with the system collecting information based on the workflow of the data.
- Graduate students are often charged with creating metadata and they may move on before the research project is complete, leaving gaps that are difficult to fill after the fact.
- As data management requirements increase, largely without associated funding increases, it becomes increasingly important to streamline data management processes and reduce administrative burdens on the researchers.

For the researcher this will improve efficiency in finding files and data, as well as allow others to find his or her work.

The specific requirements and suggestions brainstormed by participants include, in no particular order:

1. Identify and incorporate specific mandated metadata formats, which particular data areas need to have.
2. Create metadata for the uploaded asset that makes it a linkable resource (e.g. DOI, Datacite)
3. Map all domain-specific metadata records to a common schema, such as Dublin Core, so the entire repository can be searched (e.g. by title, keyword and creator)
4. Provide the ability to moderate or refine the thesaurus from captured (and user-provided) metadata. If you go one step beyond captured metadata, the ability to collapse terms, correct typos, etc. will be very helpful. This task might be the responsibility of a librarian, grad student, post-doc, or faculty.

Some sense-making or cross-walks of captured metadata may also be necessary.

5. Be aware of discovery metadata vs. meta-data for data re-use
6. Rights regarding metadata
7. Allow researcher to select from a human readable list of optional meta-data actions (e.g. grab info from PubMed), or maybe more generally, allow linking to publications.
8. Imagine that meta-data will increasingly tie to research compliance Hypothesis: Institutional comfort level with use of DuraCloud by researchers will be influenced by confidence levels in the integrity of metadata capture and design.
9. Bitstream & semantic fixity
10. Extract information that permits correlation between datasets
11. Pull from the researcher profile to automatically generate, but allow flexibility/changes. There will be cases, such as when a tech is managing data on behalf of PI, where the owner will need to change to the PI and not the logged in tech.
12. Use standard meta-data schemas to increase richness and interoperability (e.g. Darwin Core, EMBL, mzXML).
13. Create facilities to content models, which allow automated generation of structural meta-data.
14. Include the ability to pull metadata information from other tools or systems
15. We may want to have at least one required like identifier or title. The rest should be automated or added as needed. Meta-data should be linkable and have persistent identifiers.
16. Software applications used to create derived data set
17. Know who the user is, what their context is, and use that to create metadata about their relationship to the data.
18. Ability to take advantage of 'linked' data.
19. Tagging/annotation support on files & containers
20. Provenance metadata (creator, version, copy-of, & c.)
21. Ability to reuse metadata that has already been created/contributed - continually lowers overhead for subsequent metadata creation.
22. On upload, extract metadata and send to external systems to extract more metadata and resources.
23. Ability to re-create the data if necessary
24. Automatic file format identification & characterization for scientific formats.

Priority Five: Create interoperability of archiving solutions with discovery systems used by specific research communities.

The issues driving this need include:

- The importance of a cohesive lifecycle that creates a smooth transition from the operational environment into the archival environment
- The need to have research data and information discoverable from the archive with complete metadata.

For the researcher this will improve discoverability of work while reducing administrative burden.

The specific requirements and suggestions brainstormed by participants include, in no particular order:

1. Must include the concept of published and unpublished data.
2. Consider choosing one or two norms or standards around discoverability metadata.
3. Business question: with which groups do you partner? Many players are developing their own tools.
4. Registry infrastructure for registering tools. Other archives could grab that. Feature extraction plugins that can operate off a type of data.
5. Create a schema-to-schema engine. Map to a DTR standard and then transfer it into any other schema. Map all to Darwin Core to create cross-repository search, or create the index. They come from the community, and DTR can accept them.
6. ResearcherID, Orchid, OpenID - common or core - general metadata that becomes associated with everything the researcher brings in. It becomes a value-add of bringing data into DTR.
7. Should DTR support DataCite? This could allow you to follow citations of data. As soon as the data is in for the first time, get the marker on it. It should support data citation in general.
8. Interoperability of archival solutions with discovery systems used by specific research communities.
9. Don't rule out being a shared archiving solution at this stage of the game. Don't assume that everything will be done at the institutional level. It creates duplication and not all will have the ability to provide one.
10. How do we manage unique identifiers? Give people the ability to say, 'I have already cited this,' otherwise we stick an identifier on it.

APPENDIX A

DTR WORKSHOP ISSUE PRIORITIZATION

Research Data Management Challenges	DTR Priority on 3.0 Scale
Priority IA: Need for a nimble short-term storage solution while research is underway. Allow the researcher to upload their data aka DropBox and provide some basic processing to make it more useful Priority IB: Creating balance of local storage for data management and central archive for long-term access and preservation as well as the interplay of data between the two.	2.6
Priority II: Creation of workflow processes across the data management life cycle. Provenance tracking: What has happened to the data over its lifetime, or at least since it has come under management? Use cases of researchers. Keep it simple Flexible and researcher-centric (rather than library- or archive-centric) - e.g. permit different choices for how to handle multiple versions of data sets, corrections, etc.	2.6
Priority III: Can confidentiality of data be maintained using cloud storage? Lawyer from the U of VA gave a very good response to this question. I think there are a number of ways to do this e.g. maybe store some of the data on a local system with the metadata and other bits in the cloud for highly sensitive data. Data mashups in the	2.3

truest sense.	
Priority IV: Automation of *basic* metadata creation and catalogue creation. Need to create a shared understanding of what "basic" metadata means. Our conversation clarified that there is a distinction between the metadata that is immediately useful to the owner of the data (the researcher), and the metadata that is traditionally created by catalogers and used by archive or library systems. It is this "owner-userful" metadata that we need to encourage; and it becomes an input to the eventual archiving process.	2.3
Priority V: Interoperability of archiving solutions with discovery systems used by specific research communities. Add domain specific tools and services as well. Design for change, evolution, and diversity in all ways will be a key requirement.	2.3
Additional Requests	
Management of research data created beyond the institution or across institutions.	2.2
Providing controlled sharing of files (such as box.net).	2.2
Full life cycle view of storage, backup, replication and archival.	2.2
Need to manage and archive very small data sets.	2.2
Build DTR in phases. Problem too complex	2.2
Creation of work spaces and collaboration spaces.	2.1
Need to manage data access and sharing with sufficient granularity to manage both practices and preferences.	2.1
Early involvement in review of data management process is critical and unusual.	2.0

Interoperability between local, inter-institutional, disciplinary, and interdisciplinary solutions	2.0
What assurances do cloud storage providers in the case of subpoena of research data? Relates both to confidentiality and intellectual property protection	1.8
Management of data that currently resides in lab notebooks. More generally, issue of both discipline specific formats and workflow systems	1.8
How do we make the data useful - format and access - to the consumer of the data?	1.8
Provide opportunities for library/institution to add value: e.g., citation metrics, alt-metrics, compliance, cost-sharing/discounts, curation, preservation	1.8
Dspace is not a data repository. Need a solution that addresses research data use cases. Fedora is ideal for this - a lot of value will also come from projects sharing Fedora Content Models for specific domain models.	1.7
Creating the value proposition and approach for researchers (sustaining data vs preservation).	1.7
Make licensing simple (not just open) One concern our faculty has expressed is researchers licensing data in ways that makes it incapable for them to use it because -- it's not licensed for commercial use, or uses an incompatible license. Data Management planning is helping this, but it would be nice to provide researchers a simplistic licensing system that allows the research data to not just be open, but actually available for other researchers to use.	1.7
What is the economic model that allows for sustained change? Is a shared solution demanded by the economics?	1.6
The graduate student is often key to metadata, or all data.	1.6
Address stakeholder requirements for data management & access: data publishers; journal publishers; university; researcher; library & research support; funders	1.6
Creating citablility in data management processes. Could likely use a system like Citation Style Language (http://citationstyles.org/) to make this easy and flexible.	1.5
Preservation of the research process where the process itself will become historically relevant.	1.5
Data stewards can touch a relatively small number of	1.5

researchers. Researchers share data management plans with each other & take language from websites for the DMP (e.g., copy repository policies into DMP)	
Support, collaboration, sharing for the newly formed research data managementt services function at the university.	1.5
Management of raw data that will not be part of the published results.	1.5
The data must be archived where it is housed. The interdependencies and context are critical.	1.4
Need to safeguard the archive, but create a full ecosystem that allows capability needed outside the archive. Data archives often have responsibilities that are distinct from those of data access environments. They need to fulfill those roles before branching out into others.	1.4
There is very high variability in data management processes both within and across domains and departments. Given our experience we say that "70-80%" of the requirements across all domains are the same, so we should be able to provide at least 80% of the solution.	1.3
Storage management. Current vendor solutions don't meet complexity of needs. Gluster is an interesting storage management solution.	1.3
Need for common export and standards for electronic lab notebooks - both specific and metaphorical notebooks. This is not so much an open standards process, though that helps interoperability, but discovering a good way to handle diversity and change.	1.3
Clear policy on data management If the process of data managements ideally starts upfront when the researcher begins to collect data, then there is a need to work through and develop clear policy and also develop tools that will reflect this.	1.3
Absence of archival of research data. This may be because there is no incentive, nor support to assist the researcher in the proper archiving of data.	1.2
What are the right roles and responsibilities for planning and management throughout life cycle.	1.2
Competitiveness of data management plana will increase in review of proposals. Must get ahead of this.	1.2

Should be able to check compliance issues related to the research data. Use records and compliance systems concepts. No need to reinvent the wheel	1.2
Local management prevents sharing and reuse.	1.1
Need for a policy framework for data management plans.	1.1
Ability to access and share proven negatives and rejected papers.	1.1
How do we draw faculty into engagement with us given value of faculty freedoms?	1.0
What is the records management process for the data mgmt plan?	0.9
How can old data be identified and detected before it is lost? How do you facilitiate researcher curation of their data -- especially when the research may have moved on to other problems. Building durable URLs that are not necessarily assumed to be "forever", but come attached with a lifespan.	0.9
Size, type of updates, and range of data types growing exponentially, especially in social sciences. (Can I get a copy of all of Facebook for analysis?)	0.9
'Just in time' tutorials for researchers on various aspects of the research data lifecycle. I believe most researchers understand the data life cycle. What is missing is the incentive to think more clearly upfront about the value of data management.	0.6

APPENDIX B

WORKSHOP PARTICIPANTS

The following participants attended the workshop and contributed to this report.

Advising Contributors	
Micah Altman	Harvard University
Mark Leggott	University of Prince Edward Island
Thorny Staples	Smithsonian Institute
Madelyn Wessel	University of Virginia
Institutional Participants	
Karim Boughida	George Washington University
Tim Dilauro	The Johns Hopkins University
Steve Gass	MIT
Geneva Henry	Rice University
Mary McEniry	ICPSR
Susan Wells Parnham	Georgia Tech
Terry Reese	Oregon State University
Gail Steinhart	Cornell University
Brian Westra	University of Oregon
Mike Wright	NCAR
Project Team Members	
James Yoon	Fluid Project
Bill Brannan	DuraSpace
Dan Davis	DuraSpace
Jonathan Markow	DuraSpace
Brad McClean	DuraSpace
Andrew Woods	DuraSpace