Texas Digital Library

TDL Data Management Working Group Report

Published August 28, 2015

Table of Contents

Introduction

The need for Data Management services is one of two large-scale needs consistently expressed by Texas Digital Library (TDL) members, a need driven in part by the February 2013 mandate from the White House's Office of Science and Technology Policy to make the results of federally funded research publicly accessible.[1] For more information on how federal agencies plan to implement this policy, please see Appendix D.

The TDL Data Management Working Group convened in Fall 2013 to begin to address this gap, with a particular focus on finding solutions for making research data accessible and reusable.

The charge of the group was to help the Texas Digital Library determine what kinds of data management services it could provide at a consortial level.

Its objectives included:

- Articulating criteria for selecting pilot projects
- Evaluating proposed projects based on that criteria
- Selecting no more than three projects to implement
- Investigating issues related to storage and accessibility of data sets
- Documenting findings and recommendations for services

[1] The February 2013 OSTP directive, entitled "“Increasing Access to the Results of Federally Funded Research” mandated that, each Federal agency with over $100 million in annual research and development expenditures develop a plan to support increased public access to the results of research.

In an effort to envision services that address a wide range of research needs, the TDL enlisted representatives from libraries of varying sizes, public and private, and research foci to participate. Members of the group included: Bruce Herbert, PhD (Chair), Texas A&M Libraries; Martha Buckbee, UT Southwestern Medical Library; Jeremy Donald, Trinity University; Maria Esteva, PhD, Texas Advanced Computing Center; Colleen Lyon, UT Austin; Christie Peters, University of Houston Libraries; Kristi Park, Texas Digital Library; Ryan Steans, Texas Digital Library; and Santi Thompson, University of Houston Libraries.

Methodology

Following an initial period of review and self-education about a number of data management platforms (including DSpace, Dataverse, Hubzero, and Figshare), the TDL Data Management Working Group took several steps to undertake an effective evaluation of possible consortial data management services. These included:

Development of Researcher Use Cases. These use cases define the scope, workflows, technology requirements, and necessary policies to fulfill three common use cases in which a researcher might seek out a data repository to manage and/or publish research data. (See Appendix A)

Development of Evaluation Matrix. Next, the group translated the needs outlined in the Researcher Use Cases into discrete, measurable requirements for use in evaluating potential technology platforms. This work resulted in an Evaluation Matrix. (See Appendix B)

Testing. The committee performed testing in two phases: a first phase to determine what was possible in each system and to ensure that the systems were optimally configured for the testing criteria, and a second final evaluation phase to rate each system using the Evaluation Matrix.

As it moved into the testing phase, the working group decided to limit further review to Dataverse and Hubzero. It was determined that further evaluation of DSpace, the capabilities and limitations of which TDL has deep knowledge and understanding, was unnecessary. Figshare, a commercial product, was removed from consideration because of budgetary considerations.

Consequently, Texas Digital Library staff implemented testing installations of both Hubzero and Dataverse. One team performed an initial evaluation of Hubzero and a second team evaluated Dataverse, according to the “Phase 1” parameters. Following a discussion of Phase 1 testing results, the group decided to continue with Phase 2 testing of Dataverse only. This decision was based on two factors: (1) the perceived complexity and resource-intensiveness of Hubzero, both from a technology and organizational perspective and (2) the conclusion that Dataverse, at least after Phase 1 testing was completed, seemed to meet many of the group’s requirements for a successful data repository service.

For the second phase of testing, the working group held an in-person meeting to conduct a thorough group evaluation of Dataverse using the Evaluation Matrix. The group rated each criterion on both its importance and on how well the system performed. The results of these ratings were compiled and analyzed to produce an evaluation of Dataverse's strengths and weaknesses as a consortial data repository platform.

Evaluation of Dataverse

To review the full results of the Dataverse testing exercise, please see Appendix C. Importance levels listed are on a scale of 0-3, with 3 being most important. The scores listed are the average of the scores given by working group members.

Strengths

The in-person evaluation generated numerous advantages for a researcher looking to deposit or access data. Dataverse has the flexibility to work with a variety of file formats, offers a user-friendly interface for ingesting or downloading content, generates DOIs and links to other identifiers to encourage long-term access and interoperability among systems, and incorporates a robust metadata platform. The full list of strengths include:

- File formats--the system ingests a variety of file formats (PDF, Excel, .hdf, .tar, .img, .tif, .gdb, .zip.) It is unclear if the system notifies you when a file is out of compliance (or if there is an option for non-compliance). **Importance: 3**

- Platform offers user the ability to drag and drop files from their desktop. Unclear how it would interact with other file destinations (including Dropbox). **Importance: 3**

- The system provides licensing information for data ingested into the repository. The default value is CC0. **Importance: 3**

- The system accounts for version changes to datasets and routinely asks users what type of version change (major or minor) they are making. **Importance: 3**

- The system generates DOIs for each dataset. **Importance: 3**

- User Authentication -- the system has the ability to approve or deny access to the system and its functionality based on specified groups or affiliations. **Importance: 3**

- The system links to a wide number of associated documents and author credentials, including ORCID, object DOIs, grant numbers, ETD handles. Does not link to author profile pages, ETD handles, or Data Management Plans. **Importance: 3**

- Data consumers are provided with both rich metadata as well as operational reuse information, e.g., a README file, at the discretion of the inputter. **Importance: 3**

- Access levels--differing access levels can be customized for users, with a variety of access levels (e.g., to specific datasets) to choose from. **Importance: 2.8**

- The system has a default CC0 license for ingested content. It may be configured to offer other options to the inputter. **Importance: 2.8**

- Database File Formats: Dataverse successfully ingests db, GIS db, SQL db files.. **Importance: 2.6**

- The system can accommodate GIS data. **Importance: 2.6**

- The system allows data ingester to share unpublished data with others. **Importance: 2.6**

- The system has the ability to require users to enter certain metadata fields before successful ingest, and it offers many helpful options for users to select and populate metadata fields. **Importance: 2**

- The system articulates appropriate uses of the data; it absolves the data creator from responsibility if data is used illegally or inappropriately. **Importance: 1**

Weaknesses:

The in-person evaluation also identified several weaknesses with Dataverse, particularly around limitations for embargoing data and exporting it from the system. The working group is aware of the weaknesses, however, and does not view them as barriers to acquiring data sets and making them accessible. The full list of weaknesses include:

- The system does not alert user of copyright issues or policies, or require the user to agree to statements regarding the copyright of the material they seek to upload prior to the ingest of content. **Importance: 3**

- Data Export -- the system does not allow for the batch export of data. **Importance: 3**

- Data reuse information--the system contains some metadata fields that would provide reuse information to the user (e.g., software version). These fields must be completed once content is uploaded to repository. **Importance: 2.8**

- The system offers no direct embargo option upon ingest. Users can restrict access to data or choose to unpublish it within the repository. **Importance: 2.8**

- Metadata export--the system has only a limited number of metadata fields, associated with the citation information of an object, that can be exported. **Importance: 2.6**

- Embargoes for Dataverses and files are not announced to the user either in the search results or 'record' views for items. Users would be unable to determine when a locked dataset or file would become available for download. **Importance: 2.4**

- Controlled vocabulary terms are offered only as a broad list of subjects (e.g., 'Social Sciences"). **Importance: 2.33**

- The system does not notify user on institutional preservation policy regarding storage. **Importance: 2.2**

- Preservation normalization -- the system does not normalize objects to preferred file formats to support long-term preservation. **Importance: 1.8**

Recommendation

After testing the list of requirements against the system, the TDL Data Management Working Group agreed that Dataverse provides the best combination of system performance and robustness, user ease, platform scalability, and an active open source community that responds to the evolving needs of the user community. The group recommends that TDL, through its membership, adopt Dataverse to facilitate the discovery of research data and its associated metadata.

Next Steps

TDL should convene a Dataverse Implementation Working Group to establish a statewide repository for storing and providing access to research data. Over a 12 month period, members of the working group should focus on key areas to ensure the system offers services that meet the needs of researchers, including:

- Costs and possible funding models
- Technical configuration
- Outreach, workflows and training
- Policy and governance
- Metadata

The implementation working group should approach this work in three key phases:

1. Conduct a pilot implementation of the repository, to be completed by March 2016
2. Assess and refine the pilot implementation, to be completed by June 2016
3. Launch a full implementation, to be completed by September 2016

Upon completion, the Dataverse repository will offer researchers in TDL member schools the ability to meet federal and grant required mandates, to share their work in more transparent and accessible ways, and to allow others to reuse their work to leverage and expand research.

Appendices

Appendix A: Researcher Use Cases

Title: Researcher needs to make their research data publicly available

Primary Actor

Primary actors may include PIs of federally funded research, researchers working on unfunded research or funded research with no retention requirements, and graduate students working on theses, dissertations, or other data-generating projects.

Scope

1. Funded research

The emerging federal mandates to require researchers to make the federally-funded research publicly accessible after the research project has ended. Federal agencies (will) require research data to be publically accessible six months to a year after the project ends. A few foundations also require public access to funded research results.

Under this scenario, it is common that the researcher might not really care if the data is usable. The PIs are mostly focused on ease of ingestion and controlling access until required to release data in order to meet compliance with the federal mandates with as little work as possible.

In addition, TAMU compliance folks are very interested in having easy methods of monitoring research data sets that have been curated in a repository and made publically accessible. They would likely want to query a repository database through an API on a routine basis so they can monitor all research projects at an institution and develop compliance reports.

2. Unfunded research

The number of journals that allow or even require sharing the data used in a paper are growing. This requires specifying that data are deposited publicly and list the name(s) of repositories along with digital object identifiers or accession numbers for the relevant datasets.

Example, PLOS One data sharing policies: http://www.plosone.org/static/policies#sharing

3. Graduate Student Research that is associated with the student's thesis or dissertation

Archiving research data along with a thesis and dissertation may protect against loss of the data as the student moves away from the university and enhance the usefulness of our ETD collections.

Workflow

A. Funded research
   - Upload the data
   - Describe the data
   - Obtain permissions for inclusion of copyrighted data used in research (if necessary)
   - Submit uploaded and described data set for review (if necessary)
   - Publish the data
   - The data repository generates a citation for the data set using system-generated identifier, such as a digital object identifier

B. Unfunded research
   - Same as A

C. Graduate student research that is associated with the student's thesis or dissertation
   - Same as A, but upload data with thesis or dissertation

Repository Policies

Terms of Deposit

- Requirements for researchers to curate data in repository
    - Appropriate data by author, type of data, and significance of data.
    - Required metadata

Copyright Infringement Notification

Accessibility Statement

Collection Policy

- Designated community

Terms of Use

- Modification of the agreement. The next TDL working group will need to determine whether researchers can modify the terms of the agreement a la carte.
- Eligibility, registration, access and termination
    - Embargoes
    - Data controlled by Institutional Review Board. As an example, see Restrictions on the Policies of Harvard's Dataverse site (http://best-practices.dataverse.org/harvard-policies/harvard-terms-of-use.html).
- Privacy and confidentiality
- Conduct
    - Intellectual property rights (faculty, school, student) . This could be covered by selecting licensing options at ingest (I.9) or confirming copyright (I.3) and displaying licensing information for the user (A.9).
    - Licensing agreement for copyrighted or CC data sets
- Disclaimers

Digital Preservation

- Preservation Policy
- Preservation strategies and workflows
- File format recommendations

Repository Features

Simple ingestion

Controlled vocabulary

Provide DOI so data has a citation

Allow grant number, ORCID, DOI, and/or ETD handle to be linked to data sets.

Repository has an API allowing compliance officers to generate reports of data availability

For ETDs, update Vireo to manage both ETD and data ingestion in different repositories.

Ability to ingest large data sets or data sets of specific formats.

Licensing options for data (both CC and other types of licenses)

Link to data management plan/compliance plans

Flexibility with metadata schema

--metadata templates for various data formats

Special types/formats of data can be accommodated, specifically:

- Databases (static vs. actively manipulated)
- GIS data

Title: Researcher needs a virtual research environment to share active data, which may or may not be publicly accessible, within a prescribed collaborative network

Primary Actor

Researchers involved in collaborative networks.

Scope

All researchers operate in collaborative networks. Sometimes these networks are formal, such as an advisor and her students or a team of researchers working on a grant funded project, or they can be more loosely coupled groups of researchers conducting similar work. Data sharing may provide these loosely coupled groups a competitive advantage. A collaborative network allows researchers to share data while projects are being conducted, schedule data ingestion while projects are ongoing, and set a public release date that complies with various retention policies.

Workflow

- Log in to a shared environment
- Create a project
    - Invite collaborators
    - Create or upload data to be shared with collaborators
- Work on collaborative project (using features listed below)
- Conclude project
- Select data for long-term preservation and archiving
    - Data is curated in the system or exported to another repository
        - Cite copyrighted data used in research (if necessary)
    - Publish to institutional repository
        - Describe the data
        - Cite copyrighted data used in research (if necessary)
        - Submit uploaded and described data set for review (if necessary)
        - Publish the data
        - Cite data set using system-generated identifier

Repository Policies (https://purr.purdue.edu/)

Terms of Deposit

- Requirements for researchers to curate data in repository
    - Appropriate data by author, type of data, and significance of data.
    - Required metadata

Copyright Infringement Notification

Accessibility Statement

Collection Policy

- Designated community

Terms of Use

- Modification of the agreement. The next TDL working group will need to determine whether researchers can modify the terms of the agreement a la carte.
- Eligibility, registration, access and termination

    - Embargoes
    - Data controlled by Institutional Review Board. As an example, see Restrictions on the Policies of Harvard's Dataverse site (http://best-practices.dataverse.org/harvard-policies/harvard-terms-of-use.html).
- Privacy and confidentiality
- Conduct
    - Intellectual property rights (faculty, school, student) . This could be covered by selecting licensing options at ingest (I.9) or confirming copyright (I.3) and displaying licensing information for the user (A.9).
    - Licensing agreement for copyrighted or CC data sets
- Disclaimers

Digital Preservation

- Preservation Policy
- Preservation strategies and workflows
- File format recommendations

Repository Features

Updates and microblogging

To-do lists

Project notes

- Notification of updates

Project team

- Allow data to be selectively shared with collaborators.

File management

- Version control tracking
- Embed repository ingestion process in a collaborative platform like Hubzero.

Publishing

- DOI

Customize access levels to lalow data to be selectively shared with collaborators"?

HiPAA certified

Support encrypted data transmission/transfer

Licensing options for data (both CC and other types of licenses)

Title: Researcher seeks data to (re)use

Primary Actor

Researcher is interested in conducting a meta study reusing data developed in earlier studies.

Scope

The whole promise of open data is that the data will be reused and remixed to support new scholarship. Discoverability and usability of the data is of major importance to support reuse. It is really hard to use someone else's data.

Workflow Steps

Locate appropriate repository

- Search Engine
- Social Media

Determine access rights

- Is there an embargo?
- What are terms for reuse per license?
- Is authentication needed?

Determine what software is necessary to interact with the data (internal to repository vs. external)

- If external, download data and metadata files and use data
- If internal, use data

Cite data using DOI

Specific Policies

Access

- Embargoes
- Data controlled by Institutional Review Board. As an example, see Restrictions on the Policies of Harvard's Dataverse site (http://best-practices.dataverse.org/harvard-policies/harvard-terms-of-use.html).

Intellectual property rights: CC

Usability requirements

- data format – nonproprietary formats
- Metadata
- Contextual information that supports reuse by others

Preservation promises

Requirements for researchers to curate data in repository

- Appropriate data by author, type of data, and significance of data.
- Required metadata

Specific Repository Features

Discovery Interface

- Ensure sufficient metadata for re-use
- Extensibility
    - OAI
    - API

Download mechanism

- Data
- Metadata

DOI

Licensing options for data (both CC and other types of licenses)

Appendix B: Evaluation Matrix

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
I.1	Ingest	1	Upload -- the system offers a simple ingest option for user			
I.2	Ingest	1	Controlled vocabulary -- the system provides users with standardized lists of terms to describe their data (using drop down menus or other interfaces)			
I.3	Ingest	1	Copyright Permissions Verification/Notification -- the system requires the user to agree to a series of statements regarding copyright before successful ingest			
I.4	Ingest	1	File Size -- the system has the ability to ingest large data sets (2 MB, 2 GB, 1 TB)			
I.5	Ingest	1	Metadata Schema -- the system allows users to select from multiple metadata schema/templates to describe their data. Metadata provides ability to search for individual variables.			
I.6	Ingest	2	Embedded Ingest -- the system embeds allowing active data to be ingested for public access and/or long-term preservation.			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
I.7	Ingest	3	Data Reuse Information -- the system requests information on the data set from the user in order to make the data reusable to other researchers			
I.8	Ingest	1; 2	Accessibility -- the system's ingest functionality complies with ADA regulations?			
I.9	Ingest	1; 2	Licensing -- the system allows users to select licensing terms (Creative Commons and/or others)			
I.10	Ingest	1; 2; 3	Required metadata -- the sytem has the ability to require users to enter certain metadata fields before successful ingest			
I.11	Ingest	1; 2; 3	Embargoes -- the system allows users to select a period of time to hold their data from release to the public			
I.12	Ingest	1; 2; 3	Preservation Storage Notification -- the system informs the user of the institution's policy on how long the data will be preserved and who makes long-term retention decisions.			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
I.12	Ingest		File Formats -- the system supports ingest of various file formats: images [uncompressed and compressed], video [uncompressed and compressed], text file, R file, excel and notifies user when they are out of compliance			
P.1	Processing	1	Interoperate with Vireo -- the sytem can receive data sets from Vireo and connect this data with documents in DSpace			
P.2	Processing	1	Static Databases -- the system can accommodate completed/unchanged database content			
P.3	Processing	1; 2	Actively Manipulated Databases -- the system can accommodate actively changing database content (and its virtual environment needed to remain renderable)			
P.4	Processing	1	GIS Data -- the system can accommodate GIS data			
P.5	Processing	2	System Notifications -- the system sends alerts to all administrators, team members, and/or content owners when data or metadata have been edited			
P.6	Processing	2	Version Control -- the system tracks changes to active data and makes these versions available to team members with granted access			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
P.7	Processing	1; 2; 3	Database Hosting and Maintenance -- the system visualizes the data based on existing database rules. This data will be the "final" version of the data, not the active data			
P.8	Processing	2	HiPAA Compliance -- the system complies with HiPAA regulations on data security, sharing, and access?			
P.9	Processing	2	Encrypted Data Transfer -- the system supports data encryption for transmission/transfer			
P.10	Processing	1; 2	Digital Object Identifiers -- the system offers the ability to generate a DOIs for object ingested into the repository			
P.11	Processing	1; 2; 3	Embargoes -- the system tracks embargoes and releases information to the public upon conclusion of embargo			
P.12	Processing	1; 2; 3	Preservation normalization -- the system normalizes objects to preferred file formats to support long-term preservation			
P.13	Processing	1; 2; 3	File Format Specifications -- the system alerts administrators when ingested content does not conform to a pre-set list of preferred file formats			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
P.14	Processing	1; 2; 3	User Authentication -- the system has the ability to approve or deny access to the system and its functionality based on specified groups or affiliations			
P.15	Processing	1; 2; 3	Access Levels for All Users -- the system allows for differing access levels among any collection			
P16	Processing	1; 2	Administrative Metadata -- the system automatically captures technical, structural, rights, and preservation metadata after ingest.			
C.1	Curating	2	Microblogging -- the sytem allows users to communicate with team members using a micro-blog interface			
C.2	Curating	2	To-Do Lists -- the sytem allows users to generate to-do lists for team work on active data projects			
C.3	Curating	2	Collaborative Working Spaces -- the system allows team leaders to share active data with defined team members so it can be revised by team			
C.4	Curating	2	Access Levels for Teams-- the system allows team leaders to establish differing access levels for each team member			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
A.1	Access	1	Linking from data set -- the system links data set to grant number, ORCID, DOIs of objects and related content (including publications produced from the data), data management plan, compliance plan, researcher profile page, ETD handle and/or others as we think about them			
A.2	Access	3	Metadata Reuse -- the system allows for the export of data set metadata in various outputs for reuse			
A.3	Access	3	Data Reuse Information -- the systems provides user with contextual information needed to reuse data. Metadata provides ability to search for individual variables.			
A.4	Access	3	OAI Harvest -- the system exposes metadata to OAI harvestors for aggregation			
A.5a	Access	3	Data Reuse -- the system allows for the export of data in various outputs for reuse			
A.5b	Access		Data Export -- the system allows for the batch export of data			
A.6	Access	3	Search Engine Results -- the system exposes metadata with popular search engines (Google, Yahoo, MSN) to increase discoverability			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
A.7	Access	3	Social Media -- the system interoperates with popular social media outlets (Facebook, Twitter, Research Gate?) to increase discoverability			
A.8	Access	1; 2	Accessibility -- the system's access interface complies with ADA regulations			
A.9	Access	1; 2; 3	Licensing Terms -- the system articulates the licensing terms associated with data sets			
A.10	Access	1; 3	API -- The repository has an open API allowing stakeholders to create new interfaces or reports with metadata i.e. Compliance officers generate reports of data availability			

ID	Function	Use Case #	Evaluation Factor	Notes/Results	How well does the system perform this function (0-3)?	How important is this feature (0-3)?
A.11	Access	1; 2; 3	Disclaimer -- the system articulates appropriate uses of the data; it absolves the data creator from responsibility if data is used illegally or inappropriately. Example: "In no event shall City of Redmond become liable to users of these data, or any other party, for any loss or damages, consequential or otherwise, including but not limited to time, money, or goodwill, arising from the use, operation or modification of the data. In using these data, users further agree to indemnify, defend, and hold harmless City of Redmond for any and all liability of any nature arising out of or resulting from the lack of accuracy or correctness of the data, or the use of the data.			
A.12	Access		Embargo -- For data sets indended to be released to the public, the system should stipulate embargo information to the user.			

Appendix C: Final Testing Results

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
I.1	Ingest	1	Upload -- the system offers a simple ingest option for user	3	2.5	Platform offers user the ability to drag and drop files from their desktop. Unclear how it would interact with other file destinations (including drop box).
I.2	Ingest	1	Controlled vocabulary -- the system provides users with standardized lists of terms to describe their data (using drop down menus or other interfaces)	2.33	1.4	Controlled vocabulary terms are offered only as a broad list at the subject level.
I.3	Ingest	1	Copyright Permissions Verification/Notification -- the system requires the user to agree to a series of statements regarding copyright before successful ingest	3	0	The system does not alert user of copyright issues or policies prior to the ingest of content.
I.4	Ingest	1	File Size -- the system has the ability to ingest large data sets (2 MB, 2 GB, 1 TB)	3	2.16	File sizes of 1.5GB or higher slowed or stopped the upload process for the TDL instance. Harvard instance seemed to do better. System would not be expected to handle more than 1TB at a time.
I.5	Ingest	1	Metadata Schema -- the system allows users to select from multiple metadata schema/templates to describe their data. Metadata provides ability to search for individual variables.	2.6	2	The system provides metadata schema for broad disciplines. The user must have these metadata schema configured when setting up the Dataverse folder.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
I.6	Ingest	2	Embedded Ingest -- the system embeds allowing active data to be ingested for public access and/or long-term preservation.			This criterion was eliminated during final testing due to overlap with other criteria.
I.7	Ingest	3	Data Reuse Information -- the system requests information on the data set from the user in order to make the data reusable to other researchers	2.8	1.8	The system contains some metadata fields that would provide reuse information to the user. These fields must be completed once content in uploaded to repository.
I.8	Ingest	1; 2	Accessibility -- the system's ingest functionality complies with ADA regulations	2	1	Fireyes found problems in header and footer common to all pages.
I.9	Ingest	1; 2	Licensing -- the system allows users to select licensing terms (Creative Commons and/or others)	2.8	2.8	The system has a default CC0 license. It may be configured to offer other options to the user.
I.10	Ingest	1; 2; 3	Required metadata -- the sytem has the ability to require users to enter certain metadata fields before successful ingest	2	3	Offers many helpful options for users to select and populate metadata fields.
I.11	Ingest	1; 2; 3	Embargoes -- the system allows users to select a period of time to hold their data from release to the public	2.8	0.4	The system offers no direct embargo option. Users can restrict access to data or choose to unpublish it within the repository.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
I.12	Ingest	1; 2; 3	Preservation Storage Notification -- the system informs the user of the institution's policy on how long the data will be preserved and who makes long-term retention decisions.	2.2	0	Does not notify user on insitutional preservation policy regarding storage.
I.13	Ingest		File Formats -- the system supports ingest of various file formats: images [uncompressed and compressed], video [uncompressed and compressed], text file, R file, excel and notifies user when they are out of compliance	3	3	The system successfully ingested a variety of file formats (PDF, Excel, .hdf, .tar, .img, .tif, .gdb, .zip. Only the .tgz file failed on ingest, but we believe this was a size issue, not a format issue. It is unclear if the system notifies you when a file is out of compliance (or if there is an option for non-compliance).
P.1	Processing	1	Interoperate with Vireo -- the sytem can receive data sets from Vireo and connect this data with documents in DSpace	2.4	1	While not currently a function of Vireo, this kind of interoperability is possible in future development.
P.2	Processing	1	Static Databases -- the system can accommodate completed/unchanged database content	2.6	3	Testing successfully accommodated ingest of db, GIS db, SQL db to the system.
P.3	Processing	1; 2	Actively Manipulated Databases -- the system can accommodate actively changing database content (and its virtual environment needed to remain renderable)	0.6	0	Not intended for active datasets.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
P.4	Processing	1	GIS Data -- the system can accommodate GIS data	2.6	3	The system can accommodate GIS data.
P.5	Processing	2	System Notifications -- the system sends alerts to all administrators, team members, and/or content owners when data or metadata have been edited	2.8	1	System notification functionality is possible but not working in the instance of Dataverse used for testing.
P.6	Processing	2	Version Control -- the system tracks changes to active data and makes these versions available to team members with granted access	3	3	The system accounts for version changes and routinely asks user what type of version change (before the decimal point or after) they are making.
P.7	Processing	1; 2; 3	Database Hosting and Maintenance -- the system visualizes the data based on existing database rules. This data will be the "final" version of the data, not the active data	1.4	2	Integration with Two Ravens will provide some ability to explore and manipulate tabular data in recognized formats (Stata, SPSS, RData, Excel, CSV).
P.8	Processing	2	HiPAA Compliance -- the system complies with HiPAA regulations on data security, sharing, and access.	2.4	0	Lack of HIPAA Compliance. Future Dataverse development plans include data privacy tools.
P.9	Processing	2	Encrypted Data Transfer -- the system supports data encryption for transmission/transfer	2.4	3	Supports https for all URLs.
P.10	Processing	1; 2	Digital Object Identifiers -- the system offers the ability to generate a DOIs for object ingested into the repository	3	3	The system generates DOIs.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
P.11	Processing	1; 2; 3	Embargoes -- the system tracks embargoes and releases information to the public upon conclusion of embargo	2.4	0	Embargoes for Dataverses and files are not announced to the user either in the search results or 'record' views for items. Users would be unable to determine when a locked dataset or file would become available for download.
P.12	Processing	1; 2; 3	Preservation normalization -- the system normalizes objects to preferred file formats to support long-term preservation	1.8	0	The system does not support this functionality.
P.13	Processing	1; 2; 3	File Format Specifications -- the system alerts administrators when ingested content does not conform to a pre-set list of preferred file formats	1.75	0	All file types appear to be accepted for upload.
P.14	Processing	1; 2; 3	User Authentication -- the system has the ability to approve or deny access to the system and its functionality based on specified groups or affiliations	3	3	User authentication system provides ability to approve or deny access.
P.15	Processing	1; 2; 3	Access Levels for All Users -- the system allows for differing access levels among any collection	2.8	3	Access levels can be customized at the user level, with a variety of access levels to choose from.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
P16	Processing	1; 2	Administrative Metadata -- the system automatically captures technical, structural, rights, and preservation metadata after ingest.	3	1.75	The system captures some technical metadata, including file format, file size, date of deposit, and the md5 checksum.
C.1	Curating	2	Microblogging -- the sytem allows users to communicate with team members using a micro-blog interface	1	0	Microblogging between team members is not possible
C.2	Curating	2	To-Do Lists -- the sytem allows users to generate to-do lists for team work on active data projects	0.8	0	The system does not support this feature.
C.3	Curating	2	Collaborative Working Spaces -- the system allows team leaders to share active data with defined team members so it can be revised by team	0.8	0	Realtime edits to files and datasets by members of a working group are not possible. Only creating new versions of files.
C.4	Curating	2	Access Levels for Teams-- the system allows team leaders to establish differing access levels for each team member	2.6	3	The system allows data ingester to share unpublished data with others.
A.1	Access	1	Linking from data set -- the system links data set to grant number, ORCID, DOIs of objects and related content (including publications produced from the data), data management plan, compliance plan, researcher profile page, ETD handle and/or others as we think about them	3	2.5	The system links to a wide number of associated documents and author credentials, including ORCID, object DOIs, grant numbers, ETD handles. Does not link to author profile pages, ETD handles, or Data Managment Plans.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
A.2	Access	3	Metadata Reuse -- the system allows for the export of data set metadata in various outputs for reuse	2.6	0.5	The system has only a limited number of metadata fields, associated with the citation information of an object, that can be exported.
A.3	Access	3	Data Reuse Information -- the systems provides user with contextual information needed to reuse data. Metadata provides ability to search for individual variables.	3	3	Data consumers are provided with both rich metadata as well as operational reuse information, e.g., a README file, at the discretion of the inputter.
A.4	Access	3	OAI Harvest -- the system exposes metadata to OAI harvestors for aggregation	2.3	2	OAI-PMH interface exists.
A.5a	Access	3	Data Reuse -- the system allows for the export of data in various outputs for reuse (e.g. excel could be exported to csv or vice versa)	1	0	The system does not allow end users to select a file format of choice when downloading data.
A.5b			Data Export -- the system allows for the batch export of data	3	0	The system does not support this functionality.
A.6	Access	3	Search Engine Results -- the system exposes metadata with popular search engines (Google, Yahoo, MSN) to increase discoverability	3		Unknown in this round of testing.

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
A.7	Access	3	Social Media -- the system interoperates with popular social media outlets (Facebook, Twitter, Research Gate?) to increase discoverability	1.2	2.6	The system integrates with Facebook, Google Plus, and Twitter.
A.8	Access	1; 2	Accessibility -- the system's access interface complies with ADA regulations.	2.8	1	Fireyes found problems in header and footer common to all pages.
A.9	Access	1; 2; 3	Licensing Terms -- the system articulates the licensing terms associated with data sets	3	3	The system provides licensing information for data ingested into the repository. The default value is CC0.
A.10	Access	1; 3	API -- The repository has an open API allowing stakeholders to create new interfaces or reports with metadata.	2.75	3	The repository has an open API allowing stakeholders to create new interfaces or reports with metadata

ID	Function	Use Case #	Evaluation Factor	How important is this feature? (Average Score 0-3)	How well does the system perform this function? (Average Score 0-3)	Summary Results
A.11	Access	1; 2; 3	Disclaimer -- the system articulates appropriate uses of the data; it absolves the data creator from responsibility if data is used illegally or inappropriately. Example: "In no event shall City of Redmond become liable to users of these data, or any other party, for any loss or damages, consequential or otherwise, including but not limited to time, money, or goodwill, arising from the use, operation or modification of the data. In using these data, users further agree to indemnify, defend, and hold harmless City of Redmond for any and all liability of any nature arising out of or resulting from the lack of accuracy or correctness of the data, or the use of the data."	1	3	Ability to include this information exists.
A.12	Access		Embargo -- For data sets indended to be released to the public, the system should stipulate embargo information to the user.	2.5	0	Does not stipulate embargo information to end users.

Appendix D: Public Access Mandates of U.S. Federal Agencies

In a memo[1] released by the Office of Science and Technology Policy (OSTP) on February 22, 2013, each Federal agency with over $100 million in annual conduct of research and development expenditures was directed to develop a plan to support increased public access to the results of research funded by the Federal Government. This included any results published in peer-reviewed scholarly publications that are based on research that directly arises from Federal funds, as defined in relevant OMB circulars (e.g., A-21and A-11).

A major impetus behind the TDL data repository working group was to design a community tool that could support TDL member institutions responding to the needs of their researchers for access to an institutional data repository.

Summary Table of Agencies' Public Access Policies[2]

Agency	Article Solution (A)	Maximum Embargo Period	Data Solution (D)
AHRQ	PubMed Central (PMC)	12 months	Commercial repository, yet to be named[3]
ASPR[4]	PMC	12 months	Scientific data repositories, data.gov data registry[2]
CDC[3]	CDC Stacks, using NIHMS submission system	12 months	Multiple solutions + data registry
DOD	Defense Technical Information Center (DTIC)	12 months	No specific solution[2]
DOE	Public Access Gateway for Energy and Science (PAGES)	12 months	Varies by office[2]
DOT	N/A	N/A	To be released
FDA[3]	PMC	12 months	Disciplinary data repositories, where available[2]
NASA	NASA branded PMC portal	12 months	NASA archives, or other repository[2]
NIST	PMC interface	12 months[5]	EDI registry of datasets, Developing a Common Access infrastructure[2]
NIH[3]	PMC	12 months	Multiple solutions + Data Discovery Index
NOAA	NOAA Institutional Repository, using CDC Stacks	12 months	Multiple solutions short term + NOAA Data Centers for data "worthy" of long term preservation
NSF	PAGES	12 months	An "appropriate repository" [2]
USDA	USDA public access archive system (PubAg)	12 months	USDA registry of datasets, other repository options[2]
USAID	N/A	N/A	USAID repository: Development Data Library, or other
VA	PMC	12 months	Partner with HHS, NIH, FDA, and DoD on "effective mechanisms" [2]

[1] https://www2.icsu-wds.org/files/ostp-public-access-memo-2013.pdf

[2] Adapted from Columbia University Libraries, http://scholcomm.columbia.edu/open-access/public-access-mandates-for-federally-funded-research/

[3] Will require data management plans (DMPs)

[4] Exploring a data commons solution through HHS auspices. Additionally, data management costs may be included in the budget.

[5] NIST reserves right to shorten or extend the embargo period

Specific Policies of Public Access Mandates from Federal Agencies

Agency for Healthcare Research & Quality (AHRQ)

Implementation plan: http://www.ahrq.gov/funding/policies/publicaccess/index.html

(A) Authors will be required to deposit publications in the PubMed Central database.
(D) DMP required. Data will be submitted to a commercial repository.

Assistant Secretary for Preparedness and Response (ASPR)

Implementation Plan: http://www.phe.gov/Preparedness/planning/science/Pages/AccessPlan.aspx

(A) Authors will be required to deposit publications in the PubMed Central database.
(P) DMP required. In-scope digital scientific data sets resulting from research projects must be deposited in a recognized scientific data repository capable of long-term preservation of the data and open access to the public within 30 months from the creation of the data set or upon publication of a peer reviewed publication based on the data set, whichever is sooner. Additionally, a metadata document for the data, using common core metadata, must be submitted to ASPR, and will be made publicly available on data.gov, and other appropriate sharing locations such as phe.gov. Also, new awards will not be given unless terms of previous awards are met, including the conditions detailed in data sharing and management plans.

Center for Disease Control

Implementation Plan: http://www.cdc.gov/od/science/docs/Final-CDC-Public-Access-Plan-Jan-2015_508-Compliant.pdf

(A) Authors must submit final, peer-reviewed journal manuscripts to the CDC Stacks repository using the National Institute of Health Manuscript Submission (NIHMS) system, upon acceptance of the manuscript.
(D) DMP required (Appendix B of the above linked document, will eventually be electronically fillable). Minimal data must be released at the time of article publication, with more detailed data released according to CDC standard research data release timeline; all data intended for release, regardless of publication, should be made accessible within 30 months of the end of data collection. Researchers should use the repositories available to them, including the National Center for Health Statistics (NCHS) or CDC WONDER; other options are under development.

Department of Defense (DOD)

Implementation plan: http://dtic.mil/dtic/pdf/PublicAccessMemo2014.pdf

(A) Authors must submit final, peer-reviewed journal manuscripts to the Defense Technical Information Center (DTIC) system upon acceptance for publication.
(D) DMP required. Digitally formatted scientific data sets should be stored and publicly accessible to search, retrieve, and analyze; publicly releasable primary data, samples, and other supporting materials created or gathered in the course of work should be publicly accessible at no more than incremental cost and within a reasonable time.

Department of Energy (DOE)

Implementation plan: http://energy.gov/sites/prod/files/2014/08/f18/DOE_Public_Access Plan_FINAL.pdf

(A) Discoverability and access to version of record publications will be made possible through the portal and search interface tool, the Public Access Gateway for Energy and Science (PAGES), and in cases

where the publisher-hosted version of record is not publicly accessible, the DOE will provide access to accepted manuscripts in publicly accessible repositories, of which the DOE's OSTI repository may be one.
(D) DMP required (templates). Different offices will have different requirements for storage and public access link, notably the EERE which has not yet released its requirements.

Department of Transportation (DOT)

Planning phase: http://www.transportation.gov/open/plan-chapter3#sec3-2-1

(A) This document does not address publication of major research findings.
(D) A plan for data is to be released in early 2015, with a target implementation of October 2015.

Food and Drug Administration (FDA)

Implementation plan: http://www.dot.gov/open/plan-chapter3#sec3-2-1

(A) Authors will be required to deposit final peer-reviewed versions of articles in the PubMed Central database.
(D) DMP required. Researchers are expected to make data accessible in discipline specific repositories, where available, at the time of article publication.

National Aeronautics and Space Administration (NASA)

Implementation plan:
http://science.nasa.gov/media/medialibrary/2014/12/05/NASA_Plan_for_increasing_access_to_results_of_federally_funded_research.pdf

(A) Publications will be made available through a NASA-branded portal to the National Institute of Health's PubMed Central ® (PMC) platform, following the NASA-sponsored author's submission of an exact copy of the as-accepted manuscript or the publisher-transmitted copy of the Version of Record.
(D) DMP required. The requirement for public access to sharable data may be met by including data with the publication as supplementary material, through NASA archives, or through other means, and means of access should be indicated in the published article.

National Institutes of Health

Implementation Plan: http://grants.nih.gov/grants/NIH-Public-Access-Plan.pdf

(A) Authors must submit final, peer-reviewed journal manuscripts to PMC.
(D) DMP required. While data may be deposited in any of the many already existing public repositories, using community standards of data collection and description, NIH is also funding the development of a data discovery index, and will continue to explore the development of a data commons.

National Institute of Standards and Technology

Implementation Plan: http://www.nist.gov/data/ – as applied to unclassified research projects

(A) Authors must submit either the version of record or the final accepted peer-reviewed manuscript upon acceptance for publication, plus the associated public access archive system metadata through NIST's PMC interface, all of which should be publicly available within 12 months of publication, although NIST reserves the right to shorten or lengthen this embargo period.
(D) AN 'effective' DMP is required, which should address all digital data as defined by OMB Circular A-110, and explicitly address data that will support publications. Under the guidance provided in theProject Open Data component of OMB memorandum M-13-13, metadata for existing data should conform to the schema posted at https://project-open-data.cio.gov/ and be submitted to the NIST Enterprise Data Inventory (EDI), which is visible at http://www.data.gov. NIST will continue to "track and

respond to changes in digital technologies" as it develops the Common Access Platform (CAP) for data distribution. Data should be made available 12 months following publication of the associated article.

National Oceanic and Atmospheric Administration (NOAA)

Implementation Plan: http://docs.lib.noaa.gov/noaa_documents/NOAA_Research_Council/NOAA_PARR_Plan_v5.04.pdf

(A) Authors must submit the final accepted peer-reviewed manuscript, in an accessible format, upon acceptance for publication to the NOAA Institutional Repository. These materials must be made publicly and freely available within 12 months.
(D) Data Sharing Plan will be required. Data refers to "digitally formatted scientific data resulting from unclassified research supported wholly or in part by Federal funding;" numerical model outputs and software or tools required to ingest or read data in the formats offered are included in this definition. Data must be made available with article publication for supporting data, or within one year of collection for other data. NOAA will employ short term access solutions such as SHARE and participation in developing an interagency Research Data Commons based on FAIR principles in addition to long term preservation at NOAA data centers.

National Science Foundation (NSF)

Implementation Plan: http://www.nsf.gov/pubs/2015/nsf15052/nsf15052.pdf Summary: http://www.nsf.gov/pubs/2015/nsf15051/nsf15051.pdf

(A) Authors must submit either the version of record or the final accepted peer-reviewed manuscript to the DOE's Public Access Gateway for Energy and Science (PAGES) repository in PDF/A format that should be available for download, reading, and analysis free of charge no later than 12 months after initial publication, with machine-readable metadata available at initial publication.
(D) A 2-page DMP is required. "All data resulting from the research funded by the award, whether or not the data support a publication, should be deposited at the appropriate repository as explained in the DMP."

US Agency for International Development (USAID)

Implementation Plan: N/A

(A) N/A
(D) http://www.usaid.gov/sites/default/files/documents/1868/579.pdf Section 579.3.3.: The Development Data Library (DDL) is one part of the strategy to increase public access to data. Researchers may submit data to DDL, or if it is submitted to another repository, they "must submit a notice to the DDL, providing details on where and how to access the data, in accordance with the instructions found at www.usaid.gov/data." (http://blog.usaid.gov/2014/10/announcing-usaids-open-data-policy/)

US Department of Agriculture (USDA)

Implementation plan: http://www.usda.gov/documents/USDA-Public-Access-Implementation-Plan.pdf

(A) Effective January 2016, authors of publications accepted for publication on or after this date will submit to the USDA public access archive system (PubAg) all final peer-reviewed journal manuscripts once the manuscript is accepted for publication, or the final published article, provided the author has the right to submit the published version.
(D) Phased approach, with mainstream implementation targeted for 2016-2017 and DMPs to be

required, likely starting January 2016. USDA will support a registry of datasets, and are continuing to explore other repository options.

US Department of Veterans Affairs (VA)

Implementation plans: http://www.va.gov/ORO/Docs/Guidance/Plan_for_Access_to_Results_of_VA_Funded_Rsch_02_14_2014.pdf and PMC Deposit (http://www.research.va.gov/resources/policies/public_access.cfm)

(A): Authors will be required to deposit publications in the PubMed Central database upon acceptance of publication, and make available within 12 months of publication.
(D): Clinical Trial information will continue to be submitted to and be available fromhttps://clinicaltrials.gov/. For other types of digital research data, a DMP is required and, "VA will seek partnerships with HHS, NIH, FDA, and DoD to identify and share effective mechanisms" for public accessibility under both open and controlled access conditions.

Additional Summary Information on the Federal Mandates

A crowdsourced table created by a community of academic data librarians provides additional information on the emerging federal mandates[6]. The living spreadsheet is available at http://bit.ly/FedOASummary.

[6] Whitmire, Amanda; Briney, Kristin; Nurnberger, Amy; Henderson, Margaret; Atwood, Thea; Janz, Margaret; Kozlowski, Wendy; Lake, Sherry; Vandegrift, Micah; Zilinski, Lisa (2015): A table summarizing the Federal public access policies resulting from the US Office of Science and Technology Policy memorandum of February 2013. figshare. http://dx.doi.org/10.6084/m9.figshare.1372041. Retrieved 22:10, Jul 24, 2015 (GMT).