DOLCe Project Report

Connecting the TDR Dataverse to TACC storage

2021-01-21

Context

Project team

Anna Dabrowski	Co-PI, Engineering Scientist Associate	Texas Advanced Computing Center (TACC)	adabrowski@tacc.utexas.edu
Jessica Trelogan	Co-PI, Research Data Services Coordinator	University of Texas Libraries (UT Libraries)	j.trelogan@austin.utexas.edu
Courtney Mumma	Co-PI, Deputy Director	Texas Digital Library (TDL)	c.mumma@austin.utexas.edu
James Myers	Developer	Quantitative Data Repository, Global Dataverse Community Consortium	qqmyers@hotmail.com
Chris Jordan	Manager, Data Management & Collections	Texas Advanced Computing Center (TACC)	ctjordan@tacc.utexas.edu
Nicholas Lauland	Systems Administrator	Texas Digital Library (TDL)	lauland@austin.utexas.edu
Clark Kim	Senior Systems Administrator	Texas Digital Library (TDL)	c.kim@austin.utexas.edu

Stakeholders

University of Texas at Austin, Planet Texas 2050 (PT2050) Grand Challenge
https://bridgingbarriers.utexas.edu/planet-texas-2050/

Planet Texas 2050 supported this project with $15,000 of funding for software development work. Additionally, the project's pilot service was aimed at supporting PT2050 research projects with larger data publication needs. These use cases provided examples for assessing the software implementation and broader service potential.

Texas Advanced Computing Center (TACC)
https://www.tacc.utexas.edu

The Texas Advanced Computing Center designs and operates powerful computing resources, and also provides software, services, and support atop hardware resources. This project used TACC's large-scale data management and storage resources with support from TACC staff.

University of Texas at Austin Libraries (UTL)
https://www.lib.utexas.edu/research-help-support/research-data-services

University of Texas at Austin Libraries provide research data services to the University of Texas at Austin (UT) community, including consultations on data sharing and publication, and managing the local Texas Data Repository (TDR) service for UT researchers.

Texas Digital Library (TDL)
https://www.tdl.org

The Texas Digital Library is a consortium of Texas higher education institutions that builds capacity for preserving, managing, and providing access to unique digital collections of enduring value. The mission of the TDL is to advance and advocate the role of digital libraries and digital scholarly communication technologies that support the research and teaching missions of institutions of higher education in Texas and to promote cooperation, communication, and resource sharing among its members. TDL hosts and maintains software and infrastructure for several services, one of which is the Texas Data Repository (TDR). The TDR includes Dataverse software, which is shared by 9 member institutions including UTL.

Texas Data Repository Steering Committee (TDR SC)
https://www.tdl.org/members/groups/texas-data-repository-steering-committee/

The Texas Data Repository Steering Committee provides strategic planning support and policy oversight for the TDR to ensure focus and direction of programs and services. The TDR SC is composed of Data Repository Liaisons from TDR member libraries who provide the Texas Digital Library with feedback and make decisions about TDR services.

The Dataverse Project and **Global Dataverse Community Consortium (GDCC)**
https://dataverse.org

The Texas Data Repository uses Dataverse, an open-source research data repository software. Software development on the Dataverse Project is led by Harvard's Institute for Quantitative Social Science (IQSS). The Global Dataverse Community Consortium provides international organization to existing community efforts and a collaborative venue for institutions to leverage economies of scale in support of Dataverse repositories around the world.

Summary

The Digital Object LifeCycle—or DOLCe—project connecting the Texas Data Repository (TDR) Dataverse to TACC storage was funded by the University of Texas at Austin's Planet Texas 2050 (PT2050) Grand Challenge program. Three of the stakeholder organizations described above partnered to complete this work: TACC, UTL, and TDL.

Our aim was to facilitate the publication and preservation of larger data in the TDR platform.[1] With PT2050 funding, we focused on offering an option for larger data publication to a limited community of researchers. Our team completed development work on the Dataverse software, which enabled the TDR to manage multiple data stores. We then connected the TDR Dataverse to TACC systems as a location for large data storage, and offered larger data publication services to the PT2050 community of researchers as a pilot.

In this case, "larger data" was defined as digital files with sizes beyond the TDR's standard limit of 4 GB, and datasets larger than 10 GB.[2] The definition was left vague in order to better understand the actual needs of researchers while testing this new functionality.

Overview of activities

1. Connected the TDR Dataverse to TACC storage systems.
    a. Established a new storage allocation for 5 TB on the Corral system at TACC.[3]
    b. Contributed to the Dataverse codebase:
        i. Allowed Dataverse software to manage multiple data stores, and
        ii. Improved the data ingest process with direct file upload and download.
    c. Connected the development TDR Dataverse instance to Corral for testing.
    d. Submitted code to the Dataverse Project for review and approval into Dataverse software version 4.20.
    e. Connected the production TDR Dataverse instance to Corral as part of the upgrade to Dataverse version 4.20 on September 30, 2020, then for an update to version 5.1.1 on November 24, 2020.
    f. Documented technical work on the TDL wiki.[4]
2. Piloted larger data publication with the PT2050 research community.

---

[1] Texas Data Repository: https://dataverse.tdl.org

[2] TDL Texas Data Repository user documentation: https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/289079303/Frequently+Asked+Questions#FrequentlyAskedQuestions-FAQTen

[3] TACC Corral system documentation: https://www.tacc.utexas.edu/systems/corral

[4] TDL wiki project documentation: https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/987365415/Larger+Data+Storage+Project

    a. Solicited users during meetings, in presentations, and through announcements.[5]
    b. Created an information page associated with the PT2050 DataX research portal at TACC.[6]
    c. Published data:
        i. A raster dataset from the PT2050-funded Burkholderia project.
3. Solicited feedback and presented work.
    a. Held conversations with stakeholders:
        i. Meetings with TACC, UTL, and TDL stakeholders.
        ii. Meetings with TDR SC members.
    b. Presented to stakeholders and related communities (see: Slides and recorded presentations).

[5] Including: PT2050 Roundtables, the PT2050 TACC Institute, the PT2050 Research Showcase, and the PT2050 Community Newsletter.
[6] DataX "Publish Data": https://ptdatax.tacc.utexas.edu/publish-data/

Implementation Details

Software development

After a review of Dataverse's existing capabilities and of planned mechanisms for handling larger data, the project undertook development to extend Dataverse's default upload mechanisms to support larger data in a way that would closely mirror how data is currently uploaded. Other mechanisms were either still in early stages of development or required users to follow a more complex upload process.

Specifically, two related efforts were undertaken. First, Dataverse's existing capability was expanded to store files in a remote object store using the industry standard S3 storage protocol. Prior to DOLCe's work, uploads initiated by users (e.g. by dragging files into the upload widget in the Dataverse web interface) would result in files being transferred to the Dataverse server, stored as a temporary file, and retransmitted to the remote S3 store. Limits in Dataverse server disk space, time limits set by various server components, and the processing burden of managing uploads within the Dataverse server made uploads of gigabyte (GB) and larger files problematic. DOLCe's extension transparently altered this process to send files directly from the user's browser to their final storage locations in S3 — without involving the Dataverse server in the transfer or creating intermediate copies. This mechanism works for files uploaded through Dataverse's web interface, as well as with its application programming interface (API). Subsequent development work within the community has further extended this capability to also split files into multiple parts during upload, enabling faster uploads and better error handling. The mechanism implemented can theoretically support upload of individual files as large as 5 TB (testing has been done up to ~100 GB files). Access control in this mechanism is managed through the use of 'pre-signed' URLs. Dataverse provides the user's browser with URLs to upload individual files (or file parts) that each include a cryptographic token authorizing only that specific upload.

Second, Dataverse was extended to allow a single Dataverse instance to manage files in several different storage locations on the backend. With the new capability, Dataverse can be configured, for example, to use one location for most files and one or more additional locations with larger storage capacity for large files. Dataverse administrators can dynamically direct uploads for a given Dataverse/collection to a specific storage location. (Community extensions to this mechanism now support directing uploads for individual Datasets to different locations).

Software deployment

During software development, the TDR's Dev Dataverse instance[7] was configured to add a storage location running MinIO S3[8] on TACC's Corral system. After testing the data upload process and the ability to direct uploads to TACC or the original Amazon Web Service (AWS) S3 store with this instance and documenting the functionality[9], code was contributed back to the larger development community for review and inclusion into the main Dataverse codebase.[10]

The code enabling Dataverse to manage multiple data stores was included in version 4.20 of the software, released April 1, 2020. This release supported "dataverse-level setting of storage location, upload size limits, and supported data transfer methods."[11] Following review by the TDR SC within the TDR Training Dataverse instance, the new version was adopted for the TDR Production Dataverse upgrade on September 30, 2020.

Pilot service

Storage setup

5 TB of storage space was provided by TACC for the PT2050 pilot data publication service. The allocation of storage on the Corral system was handled using TACC's general project and resource allocation process,[12] and was managed by Anna Dabrowski. MinIO S3 was set up by Chris Jordan, with restricted access. Only whitelisted IP addresses could interact with the S3 endpoint, this included addresses of the Dataverse server and project team members, as well as authorized users from within TACC systems.

Data publication

As part of the pilot service, the project team published a 20 GB dataset containing 110 files (https://dataverse.tdl.org/dataset.xhtml?persistentId=doi:10.18738/T8/0NWFY9) within an existing dataverse for the PT2050 Burkholderia research project (https://dataverse.tdl.org/dataverse/burkholderia).

---

[7] Development TDR Dataverse instance: https://dataverse-dev.tdl.org
[8] MinIO documentation: https://min.io
[9] TDL Wiki "Remote Data Storage Design": https://texasdigitallibrary.atlassian.net/wiki/spaces/TDRUD/pages/1001422849/Remote+Data+Storage+Design
[10] Dataverse GitHub pull request #6488: https://github.com/IQSS/dataverse/pull/6488
[11] Dataverse GitHub releases: https://github.com/IQSS/dataverse/releases/tag/v4.20
[12] TACC "Managing and Creating Allocations": https://portal.tacc.utexas.edu/tutorials/managing-allocations

The publication process involved Courtney Mumma acting as a super-admin user to change the data store for the dataverse. Once TACC storage was enabled for this dataverse, Jessica Trelogan created a new dataset record with metadata using the "Add data" functionality in the TDR web interface. The dataset was created without any files.

Since the files for this dataset were located on TACC systems, they were separately uploaded to the dataset by Anna Dabrowski using DVUploader, the command-line bulk file uploader for Dataverse.[13] Then, the dataset with files was published from the Dataverse web interface, using the "publish dataset" button.

Detailed issue resolution timeline

2020-10-09	The first attempt to upload files for this dataset failed because the dataset contained files with the same content (and therefore duplicate MD5 hashes), which Dataverse version 4.20 did not support. Files were correctly sent to the TACC data store using the DVUploader tool, but Dataverse couldn't see or list what it identified as duplicate files based on hashes. Since the structure of this geospatial data required these files—and because the potential for files to be stored and not listed posed a larger concern—the unpublished dataset was removed from the TDR Dataverse, files were deleted from the TACC store, and further testing was paused.
2020-11-24	TDR's Production Dataverse was upgraded to version 5.1.1, which added the ability to manage files with the same content.
2020-11-25	The dataset was recreated and files were successfully uploaded. However, after clicking "publish dataset" within the TDR web interface, the publication process failed, leaving a "*Publish in Progress – The dataset is locked while the persistent identifiers are being registered or updated, and/or the physical files are being validated*" message on the dataset page. The reason for the failure was not fully determined. Dataverse version 5.x added file content validation at the time of publication, and the TDR Dataverse would have worked to retrieve and recalculate the MD5 hash for all files. It's possible that this process was interrupted due to a temporary network problem. It's also possible that the Dataverse server did not have sufficient memory to calculate the MD5 hashes. Analysis of the Dataverse log also shows that it was being tested for security weaknesses by the "UT Dorkbot" at the time, which may have contributed to an unusually high memory load.
2020-12-02	The lock was removed from the dataset by Jim Myers.
2020-12-04	With a second attempt, the dataset was successfully published using the "publish dataset" button within the TDR web interface. With Dataverse retrieving and validating all of the files, the change from "draft" to "published" dataset took approximately 36 minutes.

[13] We used DVUploader version 1.0.9 (and later version 1.1.0). See DVUploader releases: https://github.com/GlobalDataverseCommunityConsortium/dataverse-uploader/releases/

2020-12-08	Jessica Trelogan attempted to download data files using the "Access Dataset" > "Download ZIP" feature in the web interface and encountered issues with downloading files over 2 GB in size, 34 of the 110 files were omitted from the ZIP due to their size. According to Jim Myers, zipping files requires them to be retrieved from the S3 store at TACC to the Dataverse server, written to disk, and then zipped into a file that a user can download. It isn't nearly as efficient as retrieving the large files directly from S3 individually. Dataverse has a configurable size limit for files to be included in zipped downloads that resulted in the largest 34 files being skipped.
2020-12-11	Anna Dabrowski reproduced Jessica's ZIP issue, and also attempted to download files individually. Individual downloads directly from S3 failed. These downloads failed due to firewall restrictions at TACC (see: Storage setup), which had been configured to limit which computers could be used to test large file uploads. A change is being made to avoid this firewall restriction, which will allow public download of large files published in TDR.

Assessment and Recommendations

Setting up new data stores within TDR

It is now technically feasible to connect many backend data stores to the TDR Dataverse. In order to attach a new data store, system administrators need to be involved both at TDL and at the institution controlling the storage site. This technical setup will only be possible with a service framework and documentation that has been approved by the TDL Governing Board.

First, the TDR SC will need to recommend alternative data store support as an extension of the TDR service for TDL member institutions. The TDR SC—or a subcommittee—will summarize the requirements for TDL executives to present to the TDL Governing Board, including a timeline for implementation. Following this, the Board will need to approve the request. After the Board approves the service expansion, the TDR SC will document the service components and work with institutions who wish to implement it locally. TDL will also determine whether there will be a cost for the setup of such a service, since it will require TDL staff labor beyond that which is currently included in the TDR Service.

Our pilot focused on handling large data and used TACC storage systems. However, alternative data stores can be set up with other storage providers and may serve different purposes for TDR member institutions.

Institutions seeking to set up an alternative data store will need to consider:

- A storage provider that supports S3 or Swift (which includes a subset of S3) protocols for Dataverse to manage.[14]
- The cost of storage, including potential fees for file upload and download, and cost for back-up copies.
- Controlling whether and when Dataverse retrieves temporary file copies, to extract metadata and/or creates archival copies in different formats, after files are added to an alternative store.
- The staffing needs and potential cost of managing multiple storage options for different use cases, and tracking storage use.
- Implementing digital preservation workflows for an alternative store.
- Creating data retention policies, as these impact cost and could force deaccessioning requirements.

Additional information from our experience is provided below to support these considerations.

[14] One way to do this is to run MinIO and use existing file storage, as is done at TACC.

TACC as an alternative data store option

TACC may be an option for institutions seeking a large data storage provider. TACC provides 5 TB of free storage on the Corral system to researchers at University of Texas System institutions. This option was used for the pilot service, with an allocation owned by Anna Dabrowski. As the UT Libraries move to provide a larger data publication service to the broader UT community, ownership will be transferred to UT Libraries staff. They will be responsible for the storage allocation, including yearly renewal and requesting additional storage space within the TACC User Portal.[15] The current rate for storage is $95 per TB per year for replicated storage.[16] Additionally, there are no fees for file upload or download.

Institutions seeking to use TACC will need to coordinate with TACC staff to set up the appropriate type of store, and manage system administrator communication between TDL and TACC staff. Please also note, TACC storage will lack the backend digital preservation workflows otherwise implemented by TDL within the TDR AWS S3 store. While TDR supports a preservation workflow to create archival dataset copies in a backend digital preservation system (Chronopolis[17]), current plans exclude large datasets stored at TACC from this workflow. This decision reflects both storage cost concerns and a recognition that the workflows have not been tested with very large datasets.

Managing multiple data stores

After a data store is connected, a TDR super-admin user will be able to adjust the settings for dataverses (introduced in version 4.20 for the Dataverse web interface) or datasets (introduced in version 5.1 as an API call) to send content to the alternative storage. For TDR, super-admin capabilities are restricted to TDL staff. Institutions with multiple data stores will therefore need to contact TDL using a Helpdesk ticket (support@tdl.org) to request changes to the store at either level.

By default, alternative data stores will be configured to use direct uploads. Once the store is chosen for a dataverse or dataset, files can be directly uploaded to the S3 endpoint (MinIO or AWS) using the DVUploader Command Line utility, or the Dataverse web interface. Currently, the pilot TACC storage has a maximum file size of 100 GB (as compared to 4 GB for TDR overall), and no ingest limit.

---

[15] TACC User Portal: https://portal.tacc.utexas.edu

[16] Replicated storage maintains a second copy of data in a geographically distinct storage system in case the first copy is lost through a system failure, or a physical disaster. It is not intended to protect users from mistakes or allow point-in-time recovery of old "versions" of data, but is typically used as part of a disaster recovery plan.

[17] Chronopolis: http://libraries.ucsd.edu/chronopolis/

Depending on preferences and use cases, institutions may want to consider configuring, or supporting further development work on, processes completed after upload to particular stores. This includes:

Thumbnail creation: Thumbnails are created on demand when datasets are displayed and cached (not during upload). This may be something that should be size limited in a future release.

Metadata extraction: Variable-level metadata are extracted up to the ingest size limit on the data store, which is currently not set. There is also a separate, system-wide setting to limit the size of particular tabular formats to be ingested (currently set a 0 which disables tabular ingest).

Mimetype analysis: Analysis only looks at the file extension for direct upload files, restoring detection based on file contents is expected in a future Dataverse release. Most such detections will only need to see a few to 1000 bytes from the beginning of the file, and can be efficient with large files.

Derivative file creation: Same as metadata extraction.

Unzipping: Unzipping is omitted for direct upload, and there are no plans to change this. The ZIP file would need to be pulled into the Dataverse server for unzipping, and then sent back to the data store. It would probably be more efficient to use a normal upload if this is desired.

Full-text indexing: Not currently enabled, this does have a system-wide size limit that can be set.

Previews: Previewers are separate programs and do their own downloads as they require. Neither the Dataverse mechanism to launch previewers nor the individual previewers that exist have any size limits in place. There has been some community interest in having such limits due to issues that arise with large files.

It may also be worth considering what would be necessary, both technically and from a policy perspective, to enable moving files for published datasets and dataverses from one data store to another. Although this is currently possible, it involves a system administrator transferring the files and then updating the database. This situation could arise if a TDR member institution would like to move existing files to an alternative data store that they've set up, or in the case that an alternative store is used and needs to be removed in the future.

Providing sustainable services locally

This project was the result of several repeated requests from researchers for larger data publication support. However, uptake of the pilot service for PT2050 researchers has been

slow. Even though the team discussed this work and requested data in multiple venues associated with the PT2050 initiative, many research teams were not ready to publish data or identified other venues for data publication. In the near-term, UTL will continue offering the pilot service to PT2050 researchers under the initial 5 TB allocation for the project.

In order to fully understand the implications of rolling out and sustaining this service for a wider research community, we are also expanding the offer to select research groups that have requested larger data support. As the 5 TB TACC allocation begins to fill up, we plan to offer the service more widely starting with targeted outreach to an expanded community of researchers at UT known to have larger data. Meanwhile, we are considering a fee-for-service model to cover additional direct storage costs, and some changes to our deposit model and policies in order to roll out the service beyond the pilot. No plans to advertise the service more widely will be made until mechanisms for cost recovery, tracking, and sustaining the service have been developed and tested.

Based on this limited pilot, we identified the following direct and indirect costs in addition to the technical and policy implications that may arise from offering this service. All of these should be considered carefully in light of each TDR member's capacity, choice of storage provider, and perceived need.

Direct costs

Unlike the storage provided by TDL as part of the basic membership agreement, each TDR member institution will need to determine and pay the cost of alternative data stores independently. As mentioned above, TACC provides storage at a direct cost to UT system members of $95 per TB per year. Because we cannot predict if this fee will fluctuate in future, accurate cost estimates for the long-term are difficult. Currently UTL plans to assume that this direct cost will continue to decrease, but review mechanisms are recommended to ensure that costs are covered on the off-chance that storage costs increase or that service providers change.

In addition, TDL's digital preservation service for TDR excludes data in external stores. If this is deemed important for the TDR member institution or for individual research data collections, implementing a similar TDL digital preservation option for remote storage would entail additional direct costs including: $2,500 per year for the Digital Preservation Service for TDL members[18], storage, ingress and egress fees as described on the TDL website[19], and an initial investment in onboarding and developing workflows (which have not been tested for very large datasets), documentation, and a retention policy.

[18] Texas Digital Library's base membership and service costs, https://www.tdl.org/members/membership/. Accessed January 15, 2021
[19] Texas Digital Library's Digital Preservation Service storage allocation and costs, https://www.tdl.org/duracloud/cost/. Accessed January 15, 2021.

Other direct costs that may need to be addressed include fees for upload/download, migration if required by service providers, and extra security or infrastructure that may be required to maintain access to the external store.

UT Libraries are currently considering a fee-for-service option for researchers on a case-by-case basis (each case with its own service-level-agreement). We are not likely to charge additional service fees, but to charge only for the direct cost of storage to the best of our ability to predict it. We are considering creating Service Level Agreements (SLAs) for each depositor who pays for this service, with a 10-year minimum guarantee of access. This would allow researchers to budget for larger data publication at the project planning stage and pay for the long-term storage costs up front, but hand over the responsibility of stewardship (including ongoing payments and allocation maintenance) to UTL.

Indirect costs

Beyond the cost of storage, there are a number of staffing considerations and policy decisions that require further thought and possible cost recovery.

Deposits
Clear delineation of which data go where will be needed in order to initiate a deposit, as well as to track and monitor the use of storage resources. This may have larger implications for deposit models, especially for institutions that are offering a completely unmediated service, and may require additional workflows and staffing considerations.

UTL is considering implementing a review step prior to publication as part of a larger exploration of adding curation as a service to improve the value of deposits. This step would facilitate conversations about data size and choices related to appraisal, selection, and retention.

Agreements with depositors
TDR institutional members will also need to assume responsibility for tracking storage usage, backups, maintenance (including cost, but also service agreements, etc). Individual members will have to develop and document policy around guarantees for access to data in those stores and for which data should go there.

When talking about the service, important points will include:

- Reference to a comprehensive policy and schedule for data retention;
- Emphasizing careful attention to the appraisal and selection of deposits prior to using the service;
- Highlighting that the TDR Dataverse is a publication platform, rather than a data storage solution, and that the TDL will manage its infrastructure and support separately from the

storage management.

Agreements with storage providers

Bi-monthly TDR reports can facilitate tracking the use of multiple storage options. These reports could include information about which dataverses are using which storage option, how much space their files use, and for how long data have been stored. This would help TDR liaisons fulfill agreements with researchers (especially those charged for the service up front) and would enable communication with storage providers. These workflows will need to be developed and costs for implementation considered.

Beyond the DOLCe Project

Through this work, we aim to enable:

- The University of Texas at Austin Libraries to offer a sustainable larger data publication option as part of their services for local researchers.
- TDR member institutions to implement their own alternative data storage options.
- Global Dataverse Community Consortium members to manage multiple data stores with their Dataverse instances.
- Future work across our organizations improving research data management, publication, and preservation, such as:
    - Pushing and pulling data across TACC systems and software platforms;
    - Chronopolis for distributed long-term digital preservation accommodating large datasets.

Recommendations for further improvements

At this point, the developments initiated by DOLCe are fully embraced by the Dataverse community. Multiple stores and direct uploads to S3 are in use, and have been incorporated into plans, by many community members including Harvard University (Harvard maintains a Dataverse open to users around the world and uses the new capabilities to provide larger data support for Harvard-affiliated projects). As noted previously, the community has supported further expansion of these capabilities and they are expected to be maintained as part of the open-source Dataverse project.

Additional use cases could possibly be handled with more development work. Some that have been discussed across the Dataverse community include:

- Additional management options, such as automatically directing large files to a specific store (versus, or in addition to assigning storage locations by dataverse or dataset).
- Adding reporting and maintenance tools that could help in identifying abandoned uploads (users neither “save” nor “cancel” an upload in progress), moving files between stores, etc.
- Adding direct upload support for ‘auxiliary files’ — a recently added capability in Dataverse that allows uploads of files containing metadata about or subsets and/or alternate formats of a data file that are linked to the data file itself.
- Supporting ingest and other file processing (such as future capabilities to perform spam and/or virus checking) at/near a remote store rather than in the Dataverse server.
- Support for creating archival Bags for larger datasets, potentially via the use of ‘holey’ Bags where large files are referenced via URL rather than being included in the Bag directly.

- Improvements in the scalability of Dataverse to support larger numbers of files in a dataset, helping to support larger datasets composed of many files.

Related work

There are several other efforts that include support for larger data, which started prior to DOLCe-supported work. Some of these may be useful as additional or alternatives for TDR large data support in the future. They include:

Data Capture Module (DCM)
DCM provided the earliest support for larger data, it is still listed as 'experimental' in the Dataverse guides. This mechanism involves users making uploads to intermediate storage using the 'rsync' command and the implementation of a DCM server that periodically checks for new files and transfers them to an S3 store configured in Dataverse. This is the only other mechanism included in the current Dataverse release[20].

Trusted Remote Storage Application (TRSA)
TRSAs are an anticipated class of applications that connect with Dataverse and manage large and/or sensitive data files at remote locations. TRSAs may offer an alternative for institutions wishing to manage data files locally. They would involve user interaction with the TRSA as well as Dataverse, which could allow functionality such as supporting computation colocated with the data (or, for sensitive data, providing users access to files within a non-networked secure 'enclave'). A proof-of-concept TRSA implementation was created by the Odum Institute as part of the Impact Project. That prototype is now moving towards production with a redesign incorporating the multistore capability developed through DOLCe and adapting the mechanisms created for direct S3 uploads to allow remote files managed by the TRSA to be registered with Dataverse.

Globus integration
Globus' FTP-based file transfer services are widely used in communities managing very large data files and large numbers of data files due to their support for efficiency and error tolerance. Two independent efforts have created prototype integrations that allow use of Globus for uploading and/or downloading files for Dataverse datasets. Due to differences in the access-control models between Dataverse and Globus, these projects have not yet reached production status. A combined effort to create a production Globus integration has been initiated and, as with TRSA, the multistore and direct upload mechanisms initiated by DOLCe are enabling a redesign that will simplify implementation and help manage the access-control differences between Dataverse and Globus.

Data download

[20] Dataverse "Big Data Support": https://guides.dataverse.org/en/5.1/developers/big-data-support.html

While the DOLCe work has focused primarily on improving support for uploading large files, downloading large files can also be challenging. Dataverse already supported direct download of files from S3 prior to DOLCe and this mechanism allows download of individual files up to (and beyond) the 100 GB limit currently set for the TACC S3 store. However, Dataverse also supports downloading a ZIP file containing all files within a dataset. The mechanism for this has been to dynamically create the ZIP file in the Dataverse server from temporary files (these must be retrieved from a remote S3 store). This mechanism is compute and resource intensive and Dataverse has implemented a size limit to exclude larger files from the ZIP, which was encountered in testing.

With the Dataverse version 5.2 release, Dataverse can now optionally use a separate application to create ZIP files. Use of this application, in its current state, could allow TDR to support inclusion of larger files in the ZIP dataset download. Running the application at TACC could also reduce network charges to/from the TDR instance on AWS.

There is also work being done to create a new API that will simplify the download of all files within a dataset to local storage using existing web download applications. These applications are capable of retrieving a list of files from a URL and then downloading all files in the list. The work in Dataverse is to provide URLs where such lists can be retrieved. Such a capability would provide a practical alternative to Dataverse's zipped dataset download that could be recommended by TDR.

Reference Documents

Slides and recorded presentations

- Research Data Access and Preservation (RDAP 2020) Summit slides: https://osf.io/a836r/
- Dataverse Community Meeting 2020, Remote Storage & Large Datasets session recording: https://www.youtube.com/watch?v=LHyiA3JeiwE&feature=youtu.be
- PT2050 Research Showcase 2020 presentation: https://vimeo.com/473106210