RESEARCH DATA REPOSITORIES:

The What, When, Why, and How

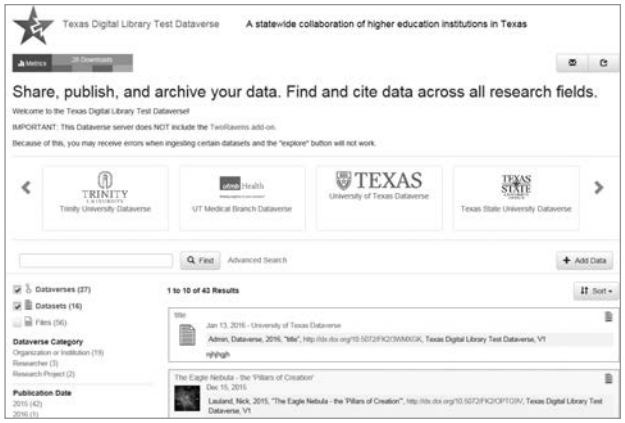
By Ray Uzwyshyn

Like it or not, Big Data, predictive analytics, and data-driven methodologies are the new kids on the block. In our largely technocratic society guided by data-driven decision making, academic libraries are also natural areas for taking up the tasks of research data aggregation, archiving, and synthesis. With these new areas of interest, new library-related positions have become hot commodities. Data scientist, data archivist, data visualization expert, and data curator are all new positions that signal areas of growth. A group of academic, specialized, and research libraries are also becoming serious about the data repositories and infrastructures needed for data curation.

This article provides an overview of the current state of research data repositories, what they are, why your library needs one, and pragmatic steps toward actualization. It surveys the present changing data repository landscape and uses practical examples from a large, current Texas consortial effort to create a research data repository for universities across the Lone Star State. Libraries in general need to think seriously about data repositories to partner with state, national, and global efforts to begin providing the next generation of information services and infrastructures.

Data Research Repositories: Definitions

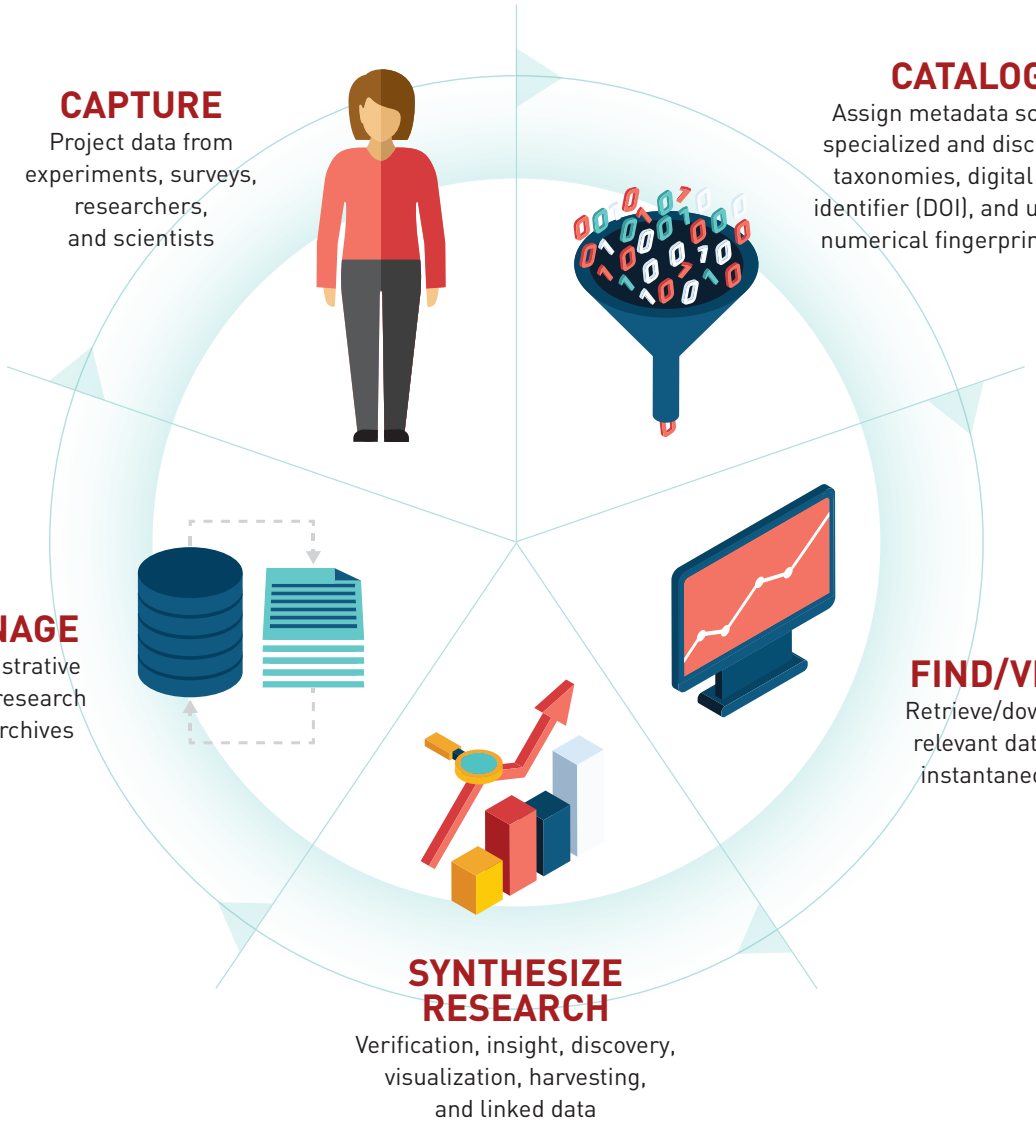
Online research data repositories are large database infrastructures set up to manage, share, access, and archive researchers' datasets. Repositories may be specialized and relegated to aggregating disciplinary data or more general, collecting over larger knowledge areas, such as the sciences or social sciences. Online repositories may also aggregate experts' data globally or locally, collecting a university or consortium of universities researcher's data for mutual benefit. The simple idea is that sharing data improves results and drives research and discovery forward. A repository allows examination, proof, review, transparency, and validation of a researcher's results by other experts beyond the published refereed academic article. Placing research data online allows instantaneous access by a globally dispersed group of researchers to share, understand, and synthesize results. This aggregation and synthesis provide an opportunity for insight, progress, and that uniquely human quest for larger understanding. Data repositories also allow for the publication of previously hidden negative data, essentially experiments that didn't work. This enables other researchers to avoid previous dead ends of those who have tried a path before them to better find their way toward more fertile territory. A global community of experts benefits from online sharing and the aggregation of research data.



Texas Digital Library's Dataverse prototype

The Nuts and Bolts

Data repositories allow long-term archiving and preservation of data by the ingestion/uploading of various data types. This includes simple Excel files, SPSS, and more exotic disciplinary formats (i.e., GIS shapefiles and Genome data-specific formats). Usually, a repository will also provide a permalink strategy for online citation and instant access so that researchers may offer a direct link to their data and ancillary files in the later published article or conference paper. This is usually provided through a digital object identifier (DOI) or universal numerical fingerprint (UNF), which allows later linking of data and possibilities for interoperability and mashing up of data archives. Within data archives, para-textual research material is also stored for later archiving and



The research data repository lifecycle

sharing. Users include social scientists and hard scientists. Data files include spreadsheets, field notes, lab methodology recipes, multimedia, and specific software programs for analyzing and working with the accompanying datasets.

The data repository infrastructure trajectory moves through a lifecycle. It begins with the experiment or research project and initial data capture and progresses to uploading, cataloging, adding disciplinary metadata schema, and assigning DOI and/or UNF (see above). Repositories will typically allow instant searching, retrievability, linking, and downloading of data. As data repositories progress, they will allow synthesis of datasets and data fields to facilitate insight, discovery, and verification. In an online global networked environment, this is accomplished through data harvesting and the possibilities of linked data and data visualization with current applications such as Tableau.

Why a Repository Now?

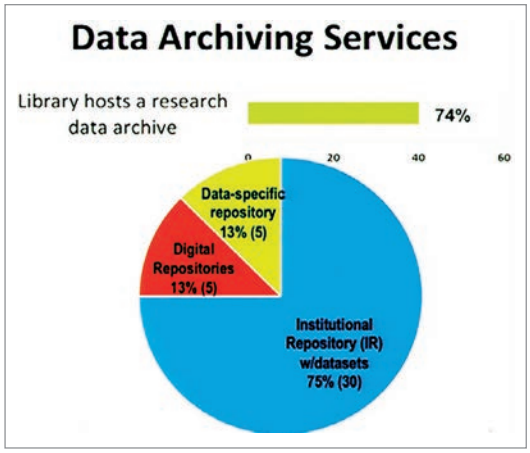
Besides being a good thing for the sharing and verification of data-driven research results, data research repositories are now necessary for university campuses. Placing one's research data online has become mandatory for any researcher wishing to receive grants from any public U.S. agency. This includes the National Institutes of Health (NIH), National Science Foundation (NSF), U.S. Department of Agriculture (USDA), and National Endowment for the Humanities (NEH). The rational is that if a researcher is drawing from the public taxpayers' trough, the research must be publicly accessible through both the article and original data. Sharing this data helps keep the wider economy vital, facilitating healthy competition toward commercialization and dissemination of discovery. If researchers do not have data management plans in place, their chance of obtaining a grant decreases. Currently, a majority of grant-funded researchers do not share data. With recent mandated changes, this situation is

Wordle of the National Science Foundation's grant guide

rapidly changing. Ivy League institutions have already capitalized on it—sharing data leverages and enhances faculty, departmental, and a university's global research standing.

The Current State of Affairs

Among research-intensive institutions and academic research libraries, about 74% provide data archiving services (see figure below). Of this group, only 13% provide data-specific repositories. Another 13% use more general digital repositories, and 74% use temporary stopgaps—text-centric repositories such as DSpace—to accommodate current grant stipulations until new data-centered applications can be put in place. The vast majority of academic libraries lag behind this cohort. In terms of Roger's technological adoption curve, phases of innovators and early adopters of data repositories are complete; we are entering the early majority and primary adoption phases (see the bottom figure). It is a great time to be thinking about a data repository.



The present state of research library data repository services (Fearon, D. and Sallans, A.C.)

Research Data Repository Software

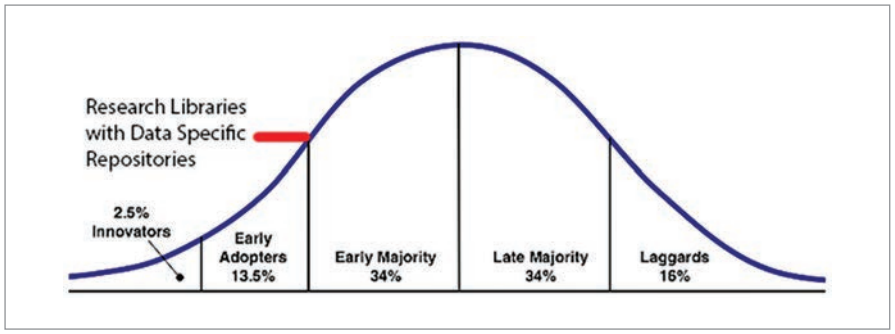
There are currently several possibilities with regard to research data repository software, some specifically created for data (i.e., Dataverse, HUBzero, and Chronopolis), others cobbled together from previous text-based institutional repository/digital library sources (i.e., DSpace, Fedora, and Hydra). The software may be hosted or installed on university servers. Different infrastructures also contain various ranges of data management and data collaboration options. There are both well-established open source software (notably, Dataverse and HUBzero) and proprietary/commercial sources (Interuniversity Consortium for Political and Social Resource [ICSPR], figshare, and Digital Commons).

Repositories, Institutions, and Consortiums

Currently, at Texas State University, we are part of a Texas-wide university effort championed by the Texas Digital Library (TDL) to implement a statewide consortial data repository based on Harvard University's open source solution, Dataverse. Dataverse is a software framework that enables institutions to host research data repositories and has its roots in the social sciences and Harvard's Institute for Quantitative Social Sciences (IQSS).

The Dataverse Network™ Project

Harvard University's Dataverse project: dataverse.org

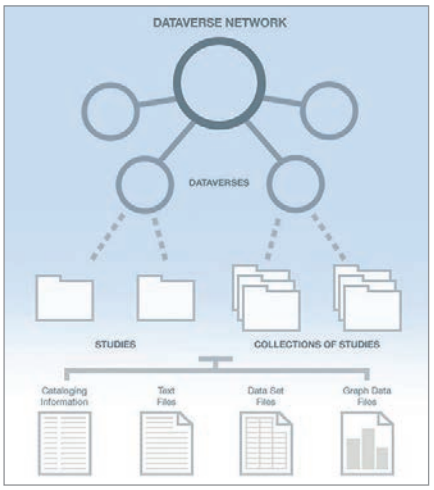


Research data repository library adoption lifecycle 2016

Data Archiving Infrastructure
Primary platform choice

Inst. Repository w/ Data (top 5)	Data-specific Repository
Dspace	Dataverse
Fedora	Chronopolis
BePress Digital Commons	HubZero (customized)
Hydra	DataConservancy
Drupal	Custom repository

Currently used data-specific and institutional repository data archives



Dataverse/Texas Digital Library (TDL) consortial application architecture

Because of Dataverse's largely customizable metadata schema abilities and open source flexibilities, TDL is using it as a data archiving infrastructure for the state (officially scheduled for launch in summer/fall 2016). The software allows data sharing, persistent data citation, data publishing, and various administrative management functions. The architecture also allows customization for a consortial effort for future systemwide sharing and interoperability of datasets for a stronger data research network (see diagram directly above).

If an institution is looking seriously at open source data repositories, other software also worth considering is Purdue's data repository system—HUBzero (hubzero.org)—and a customized instance, Purdue University Research Repository (PURR; purr.purdue.edu). Different from Dataverse's social science antecedents, HUBzero originally began as a platform for hard scientific collaboration (nanoHUB; nanohub.org). In re-creating or customizing one's institution's or consortium's data repository, PURR's

interface is particularly user-friendly. It is worth looking at how different data repositories step their researchers through the data management process via various online examples (see PURR example below).

Other more proprietary data repositories such as figshare, bepress's Digital Commons, or ICPSR are worth looking at, depending on an institution's size, needs, and present infrastructure. As the landscape is changing quickly, an environmental scan is a good idea. A good example scan recently conducted by TDL prior to choosing Harvard's Dataverse is available here: tinyurl.com/h36w93v.

Data Size Matters

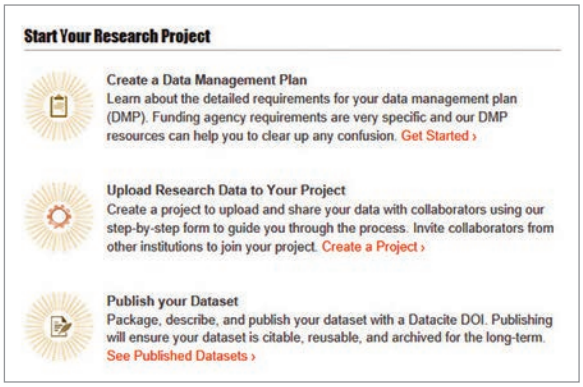
Beyond the specific data repository an institution chooses, another factor that needs to be considered is size of datasets. To generalize, researcher, project, and data storage needs come in all different shapes and sizes. Preliminarily thinking about these factors will be important as an institution moves from implementation and customization to setting policy and data storage requirements.

Research data projects may be divided into size categories of 1) small/medium, 2) large, and 3) very large. For small-to-medium datasets, these are data projects that can be stored on a researcher's current desktop hard drive, typically sets of Excel or other specialized disciplinary data files. These may be uploaded by a researcher, emailed, or transferred through university network drives to a server or the cloud and/or uploaded by a data archivist into a repository. Many of the current data-specific repositories allow researchers' self-uploading processes to begin or facilitate this process.

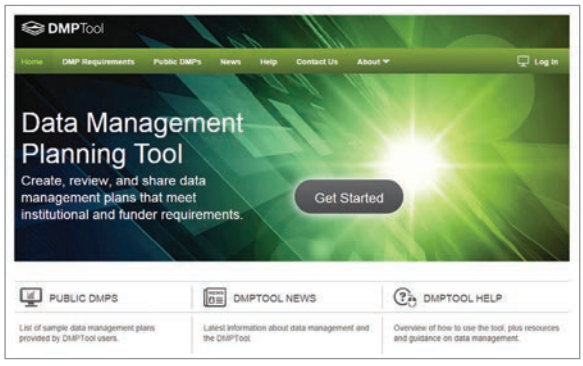
For medium-to-large projects, data may require special back-end storage systems or relationships engendered with core university IT to set up larger storage options (i.e., dedicated network space allocation and RAID). Typically, these types of datasets can still be linked online, but there is a larger weight toward data curation, adding robust metadata for access points and considering logical divisions of datasets/fields in consultation with researchers.

For very large projects, relationships may be required to be engendered with consortial, national, or proprietary data preservation and archiving efforts. For example, TDL partners with both state and national organizations—the Texas Advanced Computing Center (TACC) and the Digital Preservation Network (DPN)—and proprietary solutions, DuraCloud and Amazon Web Services (Amazon Glacier and Amazon S3). Funds become a factor here.

Typically, a university will have a spectrum of researchers with low to very high data storage requirements. Infrastructure bridges should be set up to accommodate the range of possibilities that will arise. Longer-term storage needs that a university or consortial environment anticipates and will require should also be factored in here.



Purdue University Research Repository (PURR), an easy-to-use researcher-centered interface



California Digital Library's DMPTool; dmp.cdlib.org

The many planning aspects involved in the new world of research data repositories

Data Management Planning: The Wide Angle

Data management repositories are an important, but single, piece in any researcher's larger data management plans. Other infrastructure bridges will necessarily involve offices of sponsored research, university core IT, and library personnel working together to build these new paradigms. Fortunately, several good planning tools have been created. A good starting place for planning considerations is the California Digital Library's DMPTool (dmp.cdlib.org). This will help researchers, libraries, and other infrastructure personnel begin thinking and stepping through the multi-tiered process of managing their data.

With an institutional or consortial data repository initially in place and a few key staff members to help researchers navigate this new world, a data repository infrastructure may be enabled. This article has given a whirlwind tour of the fast-changing and now required area of data repositories. A larger presentation with more detailed links and references for further exploration and research is available here: tinyurl.com/jljmmcz.

Ray Uzwyshyn (Ph.D., M.B.A., M.L.I.S.) is the director of collections and digital services for Texas State University Libraries.