JLSC

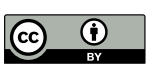
ISSN 2162-3309 | JLSC is published by the Pacific University Libraries | http://jlsc-pub.org

Volume 3, Issue 2 (2015)

Researcher-Library Collaborations: Data Repositories as a Service for Researchers

Andrew S. Gordon, David S. Millman, Lisa Steiger, Karen E. Adolph, Rick O. Gilmore

Gordon, A. S., Millman, D. S., Steiger, L., Adolph, K. E., & Gilmore, R. O. (2015). Researcher-Library Collaborations: Data Repositories as a Service for Researchers. *Journal of Librarianship and Scholarly Communication,* 3(2), eP1238. http://dx.doi.org/10.7710/2162-3309.1238



ISSN 2162-3309 10.7710/2162-3309.1238

PRACTICE

Researcher-Library Collaborations: Data Repositories as a Service for Researchers

Andrew S. Gordon
Databrary Information Engineer, New York University
David S. Millman
Databrary Co-Investigator / Director of Digital Library Technology Services, New York University
Lisa Steiger
Databrary Community Liaison and Project Manager, New York University
Karen E. Adolph
Director of Databrary / Professor of Psychology and Neural Science, New York University
Rick O. Gilmore
Associate Director of Databrary / Associate Professor of Psychology, Pennsylvania State University

INTRODUCTION New interest has arisen in organizing, preserving, and sharing the raw materials—the data and metadata—that undergird the published products of research. Library and information scientists have valuable expertise to bring to bear in the effort to create larger, more diverse, and more widely used data repositories. However, for libraries to be maximally successful in providing the research data management and preservation services required of a successful data repository, librarians must work closely with researchers and learn about their data management workflows. **DESCRIPTION OF SERVICES** Databrary is a data repository that is closely linked to the needs of a specific scholarly community—researchers who use video as a main source of data to study child development and learning. The project's success to date is a result of its focus on community outreach and providing services for scholarly communication, engaging institutional partners, offering services for data curation with the guidance of closely involved information professionals, and the creation of a strong technical infrastructure. **NEXT STEPS** Databrary plans to improve its curation tools that allow researchers to deposit their own data, enhance the user-facing feature set, increase integration with library systems, and implement strategies for long-term sustainability.

Received: 03/01/2015 Accepted: 05/13/2015

Correspondence: Karen E. Adolph, Databrary Project, New York University, 196 Mercer Street, Room 807, New York, NY 10012, karen.adolph@nyu.edu

INTRODUCTION

Databrary (2015a) is a web-based repository for storing and sharing videos collected as raw data for research on child development and learning. Databrary is designed to meet researchers' needs. It is not an isolated project initiated by the university library or central IT department. The project focuses on a specific scholarly domain, the developmental and learning sciences, and on a particular data type—video. The National Science Foundation (BCS-1238599) and the National Institute of Child Health and Human Development (U01-HD-076595) provided initial funding. Databrary began accepting contributions in early 2014 and opened for general use in October 2014. In less than a year of operation, the number of institutions with authorized users grew to 69, with 121 individual, authorized investigators from North America, Europe, South America, and Australia. As of May 2015, 35 individual contributors representing 25 different universities contributed video data or excerpts, totaling approximately 2,700 hours of video.

Most researchers who study child learning and development or classroom teaching collect video as raw data, but various barriers have prevented them from openly sharing these materials. One barrier is ethical. Although personally identifying information can be removed from text-based data while preserving its integrity and reuse potential, the same is not true of video. Raw research videos typically contain faces, voices, names spoken aloud, and the interiors of participant's homes and classrooms. Personally identifying information cannot be removed from video without violating the integrity of the data and reducing its potential for reuse. Furthermore, the collection of video and other identifiable or sensitive information requires approval by a research ethics board and informed consent from participants. The consent process formalizes a promise by the researchers to protect participants' identities from disclosure. Researchers risk breaking this promise if digital images are viewed or released to others without authorization. Sharing video poses technical, procedural, and cultural challenges as well.

Despite the challenges associated with sharing, video has significant potential for reuse. Video uniquely captures the complexity and richness of behavior, and it is largely self-documenting. Thus, videos recorded in one research setting for one purpose can be used by others for different purposes. As such, sharing video has significant potential for reuse if participants grant permission to share, contributors trust that their data will be treated with proper care, and video data find a suitable home.

To realize the promise of video data sharing, Databrary has focused on reducing barriers and on forging community consensus. Project innovations include the development of policies to enable sharing of identifiable data, the creation of technical infrastructure that

implements secure sharing, active curation and data management tools, easy citation of video data and related materials, and the promotion of practices that encourage researchers to share videos when they are most comfortable.

Experts in the NYU Library and project staff with training in library and information science have played critical roles in creating the infrastructure for sharing video. Through intense interactions with researchers in the developmental and learning science communities, Databrary's library and information science experts have learned about the diverse curation requirements of datasets collected through very different lab practices, and have determined how to represent those datasets in a standard fashion for future access and reuse. Accordingly, the Databrary project offers insights about ways that libraries and librarians may engage with scholars in other topic domains to serve emerging demands for sharing research data. In this paper, we discuss how Databrary established itself as a data repository that works closely and frequently with researchers. We describe how Databrary found solutions to many of the barriers that limit video sharing through close interaction with its target community. Finally, we outline future enhancements to Databrary that will further this work.

LITERATURE REVIEW

Creating and operating research data repositories pose many challenges. Structural features—how the repository is integrated within the host institution and how it interfaces with content producers—are critical. Databrary has strong ties to the NYU library but operates autonomously as a domain repository.

Collaborations between libraries and research entities or content producers are common. For example, Purdue Libraries and Information Technology at Purdue collaborate on the Purdue University Research Repository (Witt, 2012). The Inter-university Consortium for Political and Social Research (ICPSR) emerged from the Institute for Social Research at the University of Michigan to encompass a consortium of institutional partners (Lyle, 2014). Yale's Office of Digital Assets and Infrastructure collaborates with the Institution for Social and Policy Studies on an open data repository in the social sciences (Peer & Green, 2012).

Curation of research data (Ogburn, 2010) requires new policies and practices for communication and engagement with the scientific community. Purdue Libraries, for example, developed a process for interviewing researchers about their data curation needs when contributing to data repositories because, in Carlson's (2012) words, "services that do not align with real-world needs of researchers will not be used" (p. 12). Librarians have strengths in metadata creation and standardization across domains. Their involvement can help to reduce barriers to sharing that often plague data dissemination, particularly where

research cultures differ widely (MacMillan, 2014). A librarian who works with researchers and content producers can provide material description and documentation services that are informed by domain expertise. This results in higher quality research outcomes (Federer, 2013). Librarians can also assist researchers to deploy data management tools and implement best practices that make it easier for data owners to describe and prepare their data for preservation (McLure, Level, Cranston, Oehlerts, & Culbertson, 2014).

Developing successful data repositories also requires new practices to manage workflows involving technology and metadata creation. Research data include diverse materials related to scholarly process as well as to products. Researchers are increasingly held responsible for documenting and sharing the raw data from their research along with its products and derivatives (Heidorn, 2011; Greenberg, White, Carrier, & Scherle, 2009). The inclusion of process-oriented materials means that research data differ from other products of research such as journal articles and books, and it raises questions about how libraries should respond (Wickett, Sacchi, Dubin, & Renear, 2012; Hourclé, 2008). Representing research data outside of its original context risks making the data more difficult to interpret (Borgman, 2012). How should libraries represent datasets in a manner that allows them to be cataloged, preserved, cited, and understood by others?

Propagating research data is fast becoming a core component of scientific communication, but bridges between researchers and repository staff must be built to facilitate that propagation (Castelli, Manghi, & Thanos, 2013). A survey of repository staff workers in Australia and New Zealand illustrates some of the challenges ahead. For example, building digital data repositories requires library workers to develop new skills related to repository software and skills to communicate with IT departments (Simons & Richardson, 2012). Library staff also need to develop working knowledge about semantic web-based metadata schemas such as the Resource Description Framework (RDF), multimedia file formats, and access concerns such as copyright legislation and open access standards (Simons & Richardson, 2012). Libraries are a natural place for these new practices to emerge, but implementation is not trivial. Beyond new roles and practices, institutional changes may be required such as the reorganization of roles, positions, and the development of library infrastructure to support research data curation (Giarlo, 2013).

Databrary's experience echoes many of these points. Staff with expertise in library and information science have found that by interacting closely with researchers, they are able to provide curation services more closely tailored to individual needs. Providing those services, however, has required a range of new technical skills pertaining to software development. In turn, staff who lack library or information science backgrounds have learned about curation, preservation, and metadata. The result is a team with multidisciplinary expertise that is better suited to provide high quality data curation services.

DESCRIPTION OF SERVICES

Databrary's services cluster around several themes: a focus on community outreach and the provision of services for scholarly communication, engaging institutional partners, offering services for data curation with the guidance of information professionals, and developing and maintaining a strong technical infrastructure. We discuss each of these in turn.

Community Outreach and Scholarly Communication

From the outset, Databrary sought to connect with the research community in a meaningful way, in part by hiring staff who understand both the professional needs of researchers and the requirements for preserving information and facilitating access to it. Early on, the project team came to understand that researchers were more likely to share video data with colleagues who were part of the same scholarly community—people who held the same understanding about the sensitivities involved in sharing identifiable data related to children and families. At the same time, the team understood that it would have to change prevailing sentiments about the feasibility of sharing videos openly and bring knowledge about data curation and preservation practices to researchers unfamiliar with these topics. This led to a decision to hire two staff, one with specific responsibilities for community outreach and a second with experience in library and information science. These team members and Databrary's technical staff interact directly with researchers, providing hands-on support at every stage of the researcher's interactions with Databrary. Staff assist with initial user registration, consult with research ethics boards, and manage data curation. The Databrary team actively seeks out new potential contributors and datasets, and the team has established partnerships with some of the primary scholarly associations in the developmental and learning sciences: the Society for Research in Child Development, the International Congress on Infant Studies, the Cognitive Development Society, and the American Educational Research Association.

Databrary has also attempted to forge a consensus on professional values concerning the questions of what materials to share, when in the research life cycle materials should be shared, who should share, and how one should acknowledge use of shared videos and other materials. Databrary's Data Sharing Manifesto (Databrary, 2015b) articulates that all researchers should share as much material as they are comfortable sharing and have permission to share, researchers should share as early in the research life cycle as possible, and researchers should properly acknowledge all materials contributed by others that inform their research products. To support proper citation behavior, Databrary provides valid uniform resource identifiers (URIs) in a standard format for datasets as a whole and for subcomponents within them. The system also connects to publications associated

with a dataset via persistent identifiers such as Digital Object Identifiers (DOIs). Library and information science experts have been instrumental in shaping the design and implementation of these features.

Finally, through communication with researchers, Databrary staff learned about the important connections between data privacy requirements, trust in the security of the repository, and a potential contributor's support for open data sharing. Databrary has had to create policies and technical systems to protect data privacy and establish trust. Sharing identifiable research data requires that the Databrary system restricts access to materials on the basis of the permissions granted by individual participants and on the level of sharing granted by the researcher. Databrary offers several permission levels, allowing researchers to share data only with their own labs, in bilateral relationships with specific individual Databrary researchers, with the entire community of authorized Databrary researchers, or with the general public. Because of this, Databrary staff work closely with data contributors to determine how a dataset's original distribution restrictions, usually governed by a research ethics board, correspond to Databrary's access levels. As such, privacy considerations constitute a significant component of the curation process.

Institutional Positioning

Databrary has established relations with a diverse "internal" community as well. The project relies on several collaborations and partnerships within NYU, the host institution. These enable Databrary to navigate swiftly between the University Library and the community of researchers on campus and elsewhere. The system employs a hybrid technical architecture, developed initially by the NYU Libraries and the central IT organization. The Databrary web application uses central IT servers and storage. The Libraries and central IT, in turn, guarantee the preservation of Databrary collections indefinitely, even if project funding is interrupted. This partnership requires that Databrary follow digital preservation best practices. This model is new for the Libraries and central IT, but it represents a desired direction for enhanced central support for research data repositories across the University.

Another significant partnership is with the Office of Sponsored Programs. Normally, this office does not work closely with projects after funding has been received. However, the Office of Sponsored Programs has been an active partner in developing new policies for granting access. The office acts as a model university Authorized Organizational Representative, a role that is critical in the legal and policy framework Databrary developed for sharing between institutions. Similarly, the General Counsel's office, ordinarily a strictly administrative office that challenges or defends legal issues, is an active partner in developing the legal and policy framework for inter-institutional sharing.

A document called the Databrary Access Agreement enables inter-institutional sharing (Databrary, 2015c). This agreement must be signed by an authorizing official, commonly referred to as an Authorized Organizational Representative in the U.S. context, or someone that has the authority to affirm the enforcement of research practices on behalf of institutions elsewhere. This is typically the director of the institution's Office of Sponsored Programs. Then, an officer at the institution can authorize individual researchers within the institution to access and share data in Databrary. Researchers agree to treat data from Databrary with the same standards of care and ethical concern that would apply to data they collect themselves, to respect the desired release preferences of people depicted on videos, and to supervise the use of Databrary videos and other materials by students and staff under their guidance. The agreement permits both access to the data and, with ethics board approval, contributions. To our knowledge, this combination of user-contributor privileges makes the agreement novel, and, like other aspects of the Databrary project, it emerged as a way to reduce barriers to sharing that the team discovered in engaging with the target scholarly community. By bringing together and creating collaborations among various administrative entities in new ways, Databrary has also influenced university administrative processes.

Curation

A primary purpose of Databrary's community outreach efforts, internal partnerships, and policy framework is to secure contributions of raw research videos. Databrary supports *after-the-fact* and *active curation*, or what Giarlo (2013) refers to as *post hoc* and *sheer curation*, respectively. After-the-fact curation consists of ingesting datasets after all data collection is complete, typically after all study products (research papers, analyses, etc.) are created. After-the-fact curation usually involves considerable time and energy on the part of the data owner to convey the essential aspects of their dataset for ingestion and significant assistance from a library and information science professional. Active curation involves tools built into Databrary that enable researchers to organize and manage their raw data and metadata online in the midst of its collection. Databrary built a web-accessible user interface that allows researchers to enter study metadata and upload their videos after each data collection. Making active curation a regular part of a researcher's workflow then makes sharing a quick (one button press) and final step.

A central challenge in developing a data repository is defining a metadata schema that will accept a wide variety of datasets while adding a level of standardization to allow deposits to be easily searched (Hourclé, 2008; Orchard, 2014). Through working with researchers, Databrary learned that requiring only a minimal amount of metadata to make a dataset understandable by the designated community was preferable to making exhaustive data descriptions mandatory. The latter burdens researchers, reducing their incentive to

participate. Moreover, developmental and learning sciences support a diverse range of research topics, and, with few exceptions, no common metadata ontologies have emerged. As a result, Databrary chose to create a system that defines minimum requirements for metadata, but supports the addition of information beyond that minimum even after a dataset has been deposited. This approach standardizes the internal representation of datasets while facilitating discovery and sharing from the outset. It also lays a foundation for the emergence of stricter metadata standards as researchers achieve consensus within the user community.

After-the-fact curation. Communication with researchers is a key component in the curation of data already collected. These types of datasets include recordings collected relatively recently and those collected many years or decades ago (often requiring digitization from tape or film). Early in the curation process, Databrary staff and researchers discuss the datasets and how the target contributors envision the representation of their data inside the repository. These discussions also inform the ongoing development of the metadata schema, ensuring that it continues to meet the diverse needs of a wide range of individual labs.

Databrary's model for seeking permission to share data is new, so most data eligible for after-the-fact curation were gathered under a different set of provisions. Communication between Databrary staff and the researcher helps to forge a mutual understanding about how to interpret pre-existing restrictions in a way compatible with Databrary's policies and ethical principles. Gathering the permissions takes considerable curation effort because access restrictions are essential metadata and apply to the study volume, sessions (i.e., analytic units within studies), and individual video files. Problems encountered and solved in the process of curating new datasets inform the process of bringing in new contributions.

After a dataset has been approved for ingesting and the contributor has been authorized for access, staff gather, organize, and prepare the data. Staff review the data for personal information Databrary does not wish to upload, such as recordings of residential addresses or Social Security numbers. In the case of older video collections, where relevant metadata may have been lost or never documented, staff also review videos for any relevant metadata related to participant tasks or conditions. In these circumstances, staff with expertise in library science and behavioral science work side by side.

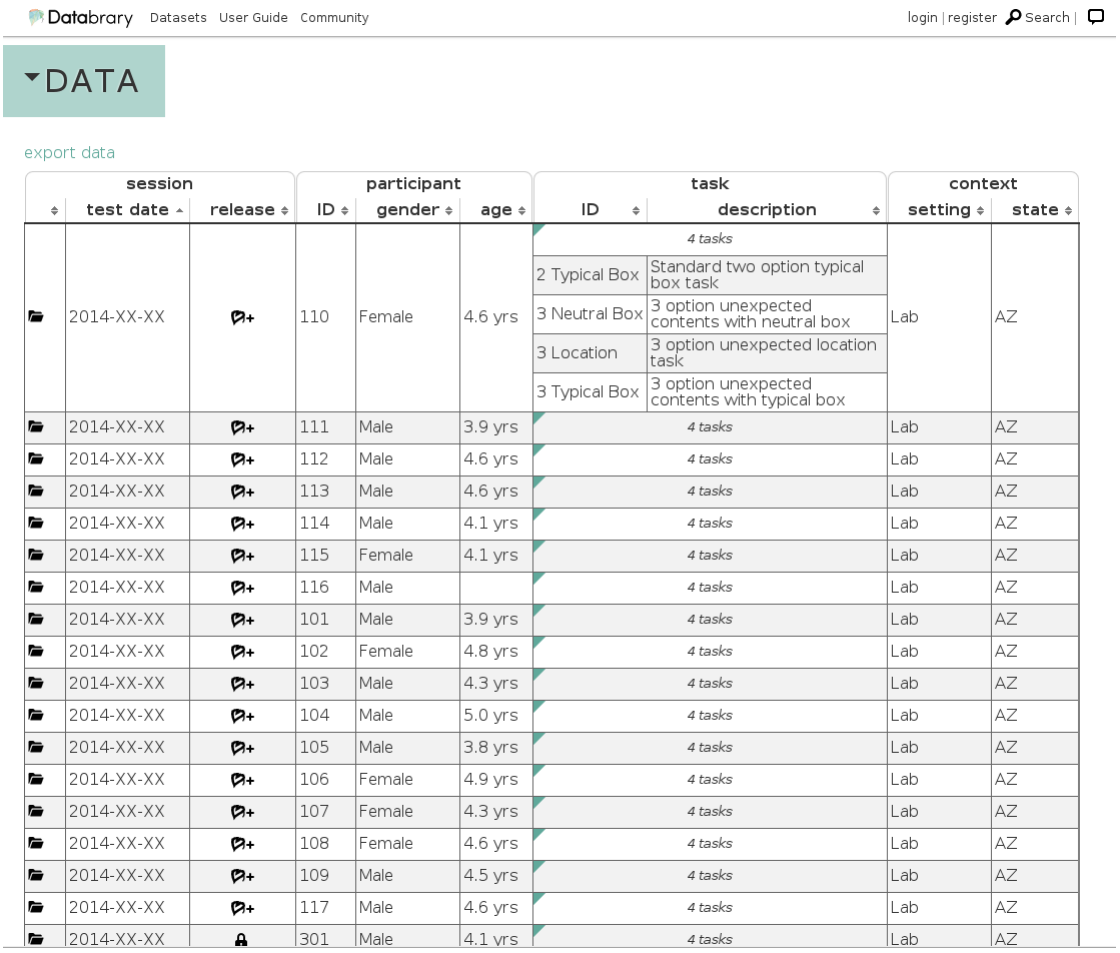
Finally, after all the metadata have been organized into a set of comma-separated value (CSV) files, and video files have been uploaded to a staging server for ingest, a set of scripts merge the metadata into a JavaScript Object Notation (JSON) file which is then submitted via the web application. This initiates the uploading of the video assets, the creation of

research sessions and records, transcoding of video files to a standard format, and clipping of video assets to remove identifying information where specified in the JSON file. After upload, the original and resulting video assets are stored on the long-term preservation location within NYU's ITS data centers.

Active curation. The curation of data after its collection requires significant resources. Thus, researchers naturally balk at the prospect of preparing data for sharing after a study has ended. Researchers who invest a lot of time in interpreting and processing their data are less likely to share (Borgman, 2012), and after-the-fact data curation does not scale well (Giarlo, 2013). From the beginning, Databrary was intended to be an active repository in which users browse, comment on, excerpt, cite, modify, deposit, and reuse data. To realize this vision, Databrary needed to provide tools to assist researchers with managing and preserving research data from early in the research life cycle. To be useful, the tools would have to reflect and augment common practices for data collection and management in this field of research.

The decision to make active curation a priority emerged from Databrary's focus on reducing the barriers to data sharing faced by its target research field. To make active curation compelling for researchers to use, Databrary needed to craft interfaces that were familiar to them. The team incorporated insights drawn from observations of data management practices in a sample of laboratories. From these, the team created a set of data management features that empower researchers to upload data with accompanying metadata as each study unfolds—essentially, "upload-as-you-go."

The insight that the observational session is a basic analytic unit of behavioral science (Bakeman & Quera, 2012) inspired the decision to create a spreadsheet interface that focuses on sessions. This spreadsheet interface (see Figure 1, following page) allows for entering, editing, and viewing session-level metadata (e.g., participants, experimental conditions or treatments, grouping variables, tasks, session access levels, etc.). Most researchers use desktop spreadsheets for precisely this purpose in their own labs, making the interface and functionality transparent to users. Databrary has also implemented a timeline for uploading, viewing, and tagging video assets related to sessions. The timeline view is designed to look and operate like video-coding software such as Datavyu, Mangold Interact, and Noldus Observer, which many researchers in developmental science use to code videos for behaviors of interest (see Figure 2, page 11). The timeline allows users to upload video files, position them to reflect the temporal order of each component of a study session (i.e., metadata records and files), and annotate video sections with user-generated tags. These tags become additional metadata indices for search and discovery.



Databrary Datasets User Guide Community login | register Search |

DATA

export data

session			participant			task		context	
	test date	release	ID	gender	age	ID	description	setting	state
	2014-XX-XX		110	Female	4.6 yrs	*4 tasks* 2 Typical Box 3 Neutral Box 3 Location 3 Typical Box	 Standard two option typical box task 3 option unexpected contents with neutral box 3 option unexpected location task 3 option unexpected contents with typical box	Lab	AZ
	2014-XX-XX		111	Male	3.9 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		112	Male	4.6 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		113	Male	4.6 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		114	Male	4.1 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		115	Female	4.1 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		116	Male		*4 tasks*		Lab	AZ
	2014-XX-XX		101	Male	3.9 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		102	Female	4.8 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		103	Male	4.3 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		104	Male	5.0 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		105	Male	3.8 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		106	Female	4.9 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		107	Female	4.3 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		108	Female	4.6 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		109	Male	4.5 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		117	Male	4.6 yrs	*4 tasks*		Lab	AZ
	2014-XX-XX		301	Male	4.1 yrs	*4 tasks*		Lab	AZ

Figure 1. Spreadsheet metadata interface for a dataset hosted by Databrary (Fabricius, 2014). For transparency purposes, Databrary exposes as much metadata about a study as possible without sharing sensitive or identifiable information.

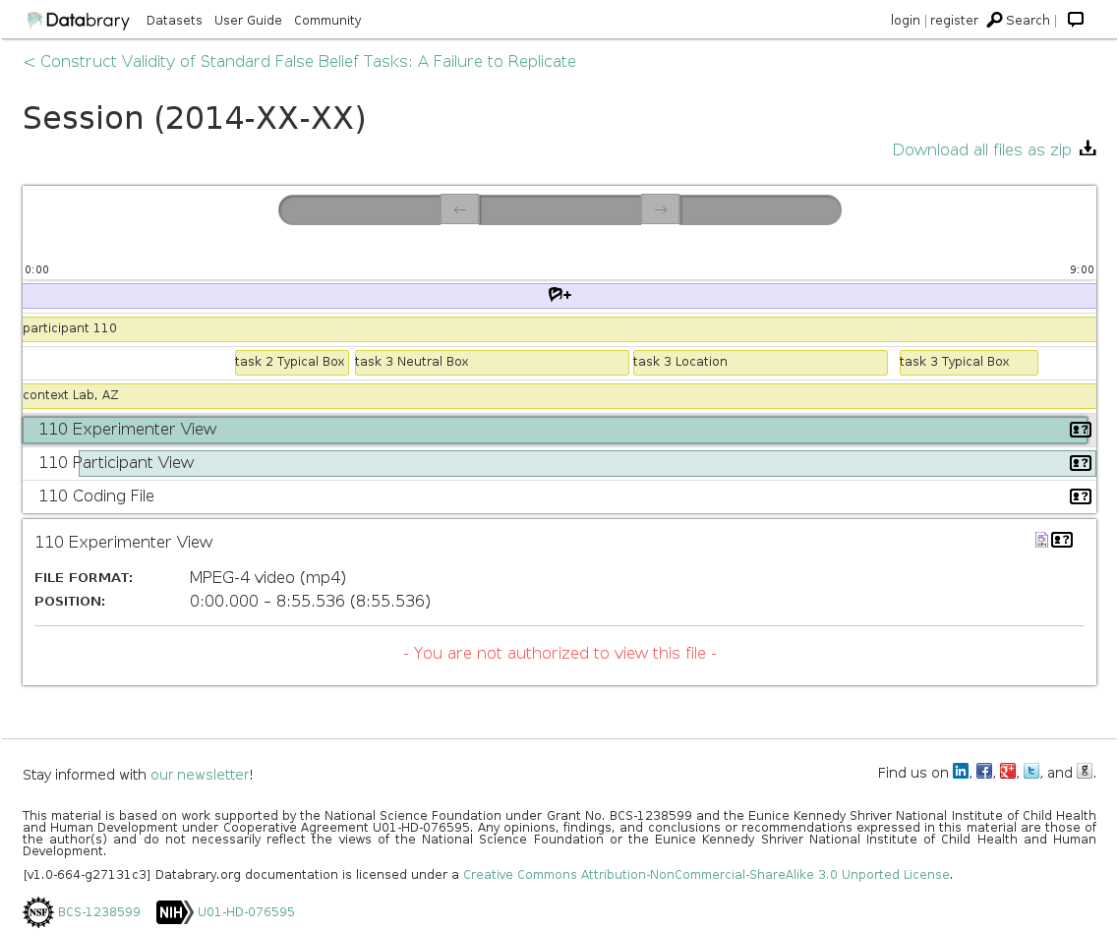


Figure 2. Timeline for one of the sessions in a dataset hosted by Databrary (Fabricius, 2014). This session has been shared with the Databrary community, but not with the public. Still images from the video are hidden and a warning message is shown. Authorized users can view, download, and tag the video.

Achieving a deep familiarity with the practices of researchers in the target domain also enabled Databrary to create a representational model for data that most researchers understand and a data management workflow similar to existing practices, but strengthened by the web-based interface. Moreover, the team acknowledges that Databrary's target research domain is what Borgman (2015) termed "small" or "little science." That is, the target audience reflects a domain characterized by localized and heterogeneous data management practices instead of a community-wide set of standard practices. As such, Databrary anticipates that the use of a standard metadata tool will contribute to the harmonization of metadata tags over time and greater standardization of data management practices, including the possibility of standardized ontologies. If it works as intended, active curation will reduce significant barriers to sharing. As a result, active curation will accelerate the pace of contributions and inclusion of new data contributors.

To encourage new research data management practices, it is not enough to build a piece of technology and hope that researchers will use it. Helping contributors to navigate the site, upload their data, and reuse other researchers' data is also a core function of Databrary's ongoing community outreach. Furthermore, staff use these outreach opportunities to gather feedback and gain a better understanding of what improvements will make Databrary a useful part of the target communities' data management workflows. Evaluation of whether active curation works should include a comparison of how many datasets are added by researchers and how many are ingested after the fact by Databrary staff and how much time active curation saves both researchers and liaison librarians compared with after-the-fact curation. Gathering feedback from users through more targeted usability testing would also promote understanding about what works and what needs improvement.

Technical Infrastructure

Because many of the system requirements for Databrary were novel and specific to the particular target domain, the team opted to build a new application rather than adapt an existing tool. The result is an open-source (Github, 2015) web application built in Haskell on the Play Framework to support a responsive user interface, a complete application program interface (API), and high-performance streaming. The backend is a PostgreSQL relational database. The user interface is built primarily on the AngularJS JavaScript framework, and all data access is performed through an open JSON API.

NYU Libraries have played a critical role in advising the development team about storage and computing technologies available within the NYU IT system and in helping to negotiate access to and cost-models for IT services. As part of the curation process, Databrary stores at least two versions of each item of Databrary video content: a copy for access and the

received original file if it was digital, or a 10-bit YUV digital preservation copy if the original version was not digital. Currently, the access version format is H.264 with AAC audio in an MPEG-4 container, although we expect the appropriate video formats to change over time, as has been the case with many recent digital video formats. The system uses NYU's High Performance Computing (HPC) cluster to transcode videos upon ingest using FFmpeg.

For preservation, the original file (if digital) or the preservation copy will be stored in a long-term preservation repository managed jointly by the NYU Libraries and the central Information Technology Services unit. This repository ensures that each content item has a Metadata Encoding and Transmission Standard (METS) file that associates the digital asset with its metadata. It stores files in two mirrored and geographically distributed locations and a third copy on offsite tape; it performs regular fixity checks; and it provides a format migration capacity, in the event that a stored format becomes at risk of obsolescence.

NEXT STEPS

The Databrary team continues to build upon the lessons learned during the project's design and initial rollout. Priorities for the next several years include improving active curation capabilities, developing feature enhancements, building more extensive integration with other services, and planning for long-term sustainability.

Improving Active Curation

As the user community grows, Databrary will continue to update and codify its curation and collection development processes. The more datasets Databrary ingests, the more staff can refine the metadata schema to represent that diversity. The user interface for active curation is still new, so staff plan to continue gathering feedback from users to improve these tools. Databrary aims to strike a balance between representing datasets as researchers want to represent them and maintaining a structure that makes information useful to and discoverable by others.

Enhancing Databrary's Feature Set

With Databrary established as a working service, staff will add enhancements to help researchers to search and access materials. Full-text search is becoming relatively trivial with off-the-shelf search engines like Apache Solr or Elasticsearch, but search is not trivial for video data. Higher level descriptions of video data can assist viewers in finding relevant content, but creating metadata that describes video file content, especially on a frame-by-frame basis, poses real challenges. By extending the video tagging and annotation tools on

the session timeline, Databrary will allow researchers to add metadata that will be useful for others to identify interesting segments of video. Similarly, the team will enhance tools for researchers to create their own excerpts—small, illustrative clips gleaned from larger video files—that contain a salient event or example of a phenomenon. With the permission of the participant, investigators can share excerpts with other scholars or use them in the classroom and at conferences. Excerpts also become a means within the repository for finding and selecting datasets that have a conceptual relationship. Because many investigators who collect video do so in conjunction with other temporally dense data streams—physiological measures (e.g., heart rate, brain activity), body motion, or eye gaze position—we will explore ways to link Databrary's video assets to external repositories storing these measures, or where feasible, provide internal support for them. Finally, Databrary has developed its own desktop coding tool, Datavyu (2015), and plans to incorporate ways to read and write Datavyu files in the web interface. In addition, Databrary plans to read and write files compatible with other prevalent video coding/annotation tools used in the developmental and learning science communities. This will allow researchers to more easily share video coding/annotation data with their video data regardless of the coding software they use in their own lab.

Integration with Other Services

Databrary plans to strengthen its connection with existing library services (i.e., the library catalog and other aggregate searches over existing data repositories). Databrary is well positioned to provide interoperability with library-based metadata schemas (such as export of data packages cross-walked to Dublin Core) and to implement standards such as the Open Archives Initiative–Protocol for Metadata Harvesting (OAI-PMH). This will allow for the automated incorporation of data that researchers add to Databrary into federated library searches with other domain-specific data repositories.

Additionally, by providing a refined API and assigning DOIs to volumes, Databrary will provide libraries and other information systems the opportunity to tap into stored datasets in a more customized fashion. Minting DOIs for datasets in Databrary will also allow data to be cited in future journal articles. This helps contributors by making measurable the scholarly impact of deposited data.

Planning for Long-term Sustainability

Currently, Databrary does not charge users for storage, curation, or reuse services. The NSF and NICHD grants bear the cost. Sustaining domain specific research data repositories on project-specific grants is common, but the model has flaws. Databrary is part of a consortium

of domain specific repositories led by the ICPSR that has called for new, more sustainable funding models (Ember et al., 2013). In the meantime, the project team continues to develop plans for long-term sustainability of Databrary, with focus on the ArXiv (ArXiv, 2015) and ICPSR institutional subscription models, storage volume/curation load based fees-for-service, and professional society partnerships.

CONCLUSION

Library practitioners are engaged in active discussions about the appropriate role of libraries in the collection and management of research data. Databrary offers a working model that demonstrates how a research data repository can benefit from interacting closely with the research community. Databrary shows that a diverse team of experts can devise novel policy, technical, and curatorial solutions to problems encountered in fostering wider data sharing. The project also demonstrates that being strategically and structurally attached to library systems through management, staff, and technology is an important ingredient in building a successful repository.

We do not assume that all data repositories will be able to replicate the exact process undertaken by Databrary. Larger scale data repositories that serve multiple fields of research may lack the available staff to shadow researchers in every represented domain. Nevertheless, the development of data repositories will require new practices (Heidorn, 2011; McLure et al., 2014; Simons & Richardson, 2012; MacMillan, 2014). It will require the work of information professionals equipped with new skill sets that allow them to translate the needs of the library to research teams. Finally, it will require leaders who are capable of navigating between repository, policy, and library workflows and are committed to understanding the work of researchers who may not have the time, motivation, or capability to properly preserve their data for the long term.

ACKNOWLEDGMENTS

This work was supported by NSF (BCS-1238599) and NICHD (U01-HD-076595) to Karen Adolph and Rick Gilmore. The authors gratefully acknowledge the NYU Libraries for their valuable advice and consultation.

REFERENCES

ArXiv (2015). *arXiv.org e-Print archive*. Retrieved June 1, 2015 from http://arxiv.org

Bakeman, R., & Quera, V. (2012). Behavioral observation. In H. Cooper, P. M. Camic, D. L. Long, A. T. Panter, D. Rindskopf, & K. J. Sher (Eds.), *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics* (pp. 207-225). Washington, DC, US: American Psychological Association. http://dx.doi.org/10.1037/13619-013

Borgman, C. L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science & Technology*, *63*(6), 1059-1078. http://dx.doi.org/10.1002/asi.22634

Borgman, C. L. (2015). *Big data, little data, no data: Scholarship in the networked world*. MIT Press.

Carlson, J. R. (2012). Demystifying the data interview: Developing a foundation for reference librarians to talk with researchers about their data (English). *Reference services Review*, *40*(1), 7-23. http://dx.doi.org/10.1108/00907321211203603

Castelli, D., Manghi, P. & Thanos, C. (2013). A vision towards scientific communication infrastructures: On bridging the realms of research digital libraries and scientific data centers. I*nternational Journal on Digital Libraries*, *13*(3/4), 155-169. http://dx.doi.org/10.1007/s00799-013-0106-7

Databrary (2015a). *Databrary: An open data library for developmental science*. Retrieved June 1, 2015 from http://databrary.org/

Databrary (2015b). *Data sharing manifesto*. Retrieved June 1, 2015 from http://databrary.org/access/policies/data-sharing-manifesto.html

Databrary (2015c). *Databrary access agreement*. Retrieved June 1, 2015 from http://databrary.org/access/policies/agreement.html

Datavyu (2015). *Datavyu: Video coding and data visualization tool*. Retrieved June 1, 2015 from http://datavyu.org

Ember, C., Hanisch, R., Alter, G., Berman, H., Hedstrom, M., & Vardigan, M. (2013). Sustaining domain repositories for digital data: A white paper. *Workshop on Sustained Domain Repositories for Digital Data*. Retrieved June 1, 2015 from http://datacommunity.icpsr.umich.edu/sites/default/files/WhitePaper_ICPSR_SDRDD_121113.pdf

Fabricius, W. (2014). Construct validity of standard false belief tasks: A failure to replicate. *Databrary*. Retrieved June 1, 2015 from http://databrary.org/volume/98

Federer, L. (2013). The librarian as research informationist: a case study (English). *Journal of the Medical Library Association*, *101*(4), 298-302. http://dx.doi.org/10.3163/1536-5050.101.4.011

Giarlo, M. J. (2013). Academic libraries as data quality hubs. *Journal of Librarianship and Scholarly Communication*, *1*(3), eP1059. http://dx.doi.org/10.7710/2162-3309.1059

Github (2015). *Databrary on Github*. Retrieved June 1, 2015 from https://github.com/databrary/

Greenberg, J., White, H. C., Carrier, S., & Scherle, R. (2009). A metadata best practice for a scientific data repository. *Journal of Library Metadata*, *9*(3-4), 194-212. http://dx.doi.org/10.1080/19386380903405090

Heidorn, P. B. (2011). The emerging role of libraries in data curation and e-science. J*ournal of Library Administration*, *51*(7/8), 662-672. http://dx.doi.org/10.1080/01930826.2011.601269

Hourclé, J. A. (2008). FRBR applied to scientific data. *Proceedings of the ASIST Annual Meeting*, *45*(1). http://dx.doi.org/10.1002/meet.2008.14504503102

Lyle, J. (2014). ICPSR: A consortial model to advance and expand social and behavioral research. *027.7*, *2*(1), 19-29.

MacMillan, D. (2014). Data sharing and discovery: What librarians need to know. *Journal of Academic Librarianship*, *40*(5). 541-549. http://dx.doi.org/10.1016/j.acalib.2014.06.011

McLure, M., Level, A. V., Cranston, C. L., Oehlerts, B., & Culbertson, M. (2014). Data curation: A study of researcher practices and needs. *portal: Libraries & the Academy*, *14*(2), 139-164. http://dx.doi.org/10.1353/pla.2014.0009

Ogburn, J. L. (2010). The imperative for data curation. *portal: Libraries and the Academy*, *10*(2), 241–246. http://dx.doi.org/10.1353/pla.0.0100

Orchard, S. (2014). Review: Data standardization and sharing—The work of the HUPO-PSI. *BBA - Proteins and Proteomics*, *1844*(1, Part A), 82-87. http://dx.doi.org/10.1016/j.bbapap.2013.03.011

Peer, L., & Green, A. (2012). Building an open data repository for a specialized research community: Process, challenges and lessons. *International Journal of Digital Curation*, *7*(1), 151-162. http://dx.doi.org/10.2218/ijdc.v7i1.222

Simons, N., & Richardson, J. (2012). New roles, new responsibilities: Examining training needs of repository staff. *Journal of Librarianship & Scholarly Communication*, *1*(2), eP1051. http://dx.doi.org/10.7710/2162-3309.1051

Wickett, K. M., Sacchi, S., Dubin, D., & Renear, A. H. (2012). Identifying content and levels of representation in scientific data. *Proceedings of the ASIST Annual Meeting*, *49*(1). http://dx.doi.org/10.1002/meet.14504901199

Witt, M. (2012). Co-designing, co-developing, and co-implementing an institutional data repository service. *Journal of Library Administration*, *52*(2), 172-188. http://dx.doi.org/10.1080/01930826.2012.655607