Databrary: A digital data library for sharing research video

Rick O. Gilmore, Ph.D.

The Pennsylvania State University  
Department of Psychology  
University Park, PA 16802  
+01-814-865-3664

rogilmore@psu.edu

Lisa M. Steiger

Databrary Project  
196 Mercer Street, Room 807  
New York, NY 10012  
+01-212-998-5800

lisa@databrary.org

Dylan A. Simon, Ph.D.

Databrary Project  
196 Mercer Street, Room 807  
New York, NY 10012  
+01-212-998-5800

dylan@databrary.org

**ABSTRACT**

Video captures the complexity, richness, and diversity of behavior unlike any other tool. As a result, large numbers of researchers who study human or animal behavior employ video. Video documents itself to a large degree, and therefore has significant potential for data re-use. Still, researchers rarely share video. Video often contains information about personal identities, so considerations about research ethics pose challenges to sharing. The relatively large size of video files and diversity of formats pose technical challenges. In this paper, we will describe how the web-based Databrary data library has overcome the most significant barriers to sharing video within the developmental and learning sciences community, including solutions to maintaining participant privacy, data tagging, and data management.

**Categories and Subject Descriptors**

H.2.8 [**Database Management**]: Database Applications – *image databases, scientific databases*

**General Terms**

Management, Documentation, Design, Security, Legal Aspects.

**Keywords**

Data sharing, open science, video, psychology, developmental science.

# INTRODUCTION

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

*JCDL '15*, June 21–25, 2015, Knoxville, TN, USA.

Copyright 2015 ACM 1-58113-000-0/00/0010 …$15.00.

Film and video recording have been a mainstay of research in the developmental sciences for nearly 100 years [CITE]. A typical lab acquires, on average, 12 hours of video per week depicting behavior in natural or laboratory environments [CITE]. These recordings constitute a primary source of raw data. Prior to descriptive or statistical analyses, researchers score or “code” recordings, mostly manually, using commercial software or makeshift tools. However, even the most detailed and laborious video coding does not exhaust the possibilities for analyses by other researchers with different questions. Thus, sharing video promises a wide range use cases, perhaps more than sharing flat-file data or static images. Despite the wealth of untapped information available in research videos and the labor expended in collecting and coding them, developmental researchers rarely share data. In the prevailing culture, investigators conduct research privately within their own laboratories. Data are destroyed or molder away on a shelf after the research is published. The Databrary project aims to change the status quo, making sharing standard practice.

Databrary enables researchers to share video and associated flat-file data for research and educational purposes. For example, a researcher interested in language development might reuse a set of video recordings created originally to study the development of walking. Researchers can conduct integrative analyses by combining data across disparate studies, increase sample sizes, verify the feasibility of proposed projects, compare methods, learn about procedures, and independently verify results. Similarly, investigators can reuse shared videos for educational purposes such as to find illustrative examples for teaching, use exemplars in presentations, show students how to conduct procedures, or provide supplemental materials for reviewers.

Data sharing and open science practices are emerging or accepted norms in the biological, physical and earth sciences with demonstrated benefits for scientific transparency and accelerated discovery [1-4]. Open research practices are gaining traction in the behavioral and social sciences [5-10]. However, open data sharing among developmental researchers is still the exception [11] not common practice. In seeking to expand sharing of video-based data, Databrary faces challenges concerning the interplay among institutional research policies, ethical issues related to obtaining participants’ consent for sharing, and technical authorization systems. This paper describes the current state of the Databrary project including the project organization, how content is acquired, stored, preserved, and accessed, policy considerations, and how the library has affected and hopes to affect the behavioral science research community.

# PROJECT ORGANIZATION

Databrary is a joint project of New York University (NYU) and The Pennsylvania State University (Penn State). The Databrary project began with a workshop on open video data sharing funded by the National Science Foundation (NSF) under Grant No. BCS-1139702 at which developmental scientists, computer scientists, library scientists, and federal agency program officers discussed the promises and challenges of sharing video data. NSF and the National Institute of Child Health and Human Development (NICHD) have provided funding under grant No. BCS-1238599 and Cooperative Agreement U01-HD-076595, respectively. The interdisciplinary team bring expertise in developmental and neural science, information technology, library science, software development, data curation, and community engagement. In addition, a board of expert advisors composed of developmental scientists, library scientists, and leaders of data repositories in the behavioral and social sciences provide guidance.

The project is organized around five core activities.

## Curation

To seed the Databrary, staff identified and selected video datasets for manual curation. The datasets were selected based on availability, presumed interest to developmental science community, suitability for the organization and file structure of the repository, diversity of study type, and on the availability of appropriate sharing permissions (see 2.2 below). Manual curation of archival datasets continues under the direction of a staff member with expertise in library and information sciences.

Nevertheless, we have found manual data curation time consuming and difficult, especially for researchers. The more difficult and time-consuming data sharing becomes, the less likely it will be that researchers participate despite best intentions. So, we have created user-friendly spreadsheet-like interfaces that enable researchers to self-curate datasets as they are collected. At the end of a data collection session, researchers may upload videos and other data files collected along with metadata about participant demographic characteristics or experimental conditions. Databrary transcodes videos to a standard format in the background, and the system alerts the user when the video can be previewed. Researchers may also upload study-level materials -- code books, flat data files used for statistical analyses, images of experimental equipment, links to related external information (e.g., GitHub repositories), etc.

This scheme, which we call "upload-as-you-go", makes the data curation required for sharing part of the normal scientific workflow, not a burdensome after thought. No data are shared until the researcher grants access, but unlike the *ad hoc* organization in a typical investigator's laboratory, on Databrary, the data are organized in a standardized way that could be readily and easily shared with others when the investigator is ready to do so.

## Policies

Sharing video requires special considerations because the data cannot be de-identified without reducing its value. No previously established research data repository has specialized in video data sharing. Therefore, Databrary had to develop a new policy and permission framework for sharing identifiable and sensitive data while simultaneously preserving the rights of research participants.

The framework has two components, both of which build on well-established principles. One extends the idea of informed consent to participate in research by requiring that participants give explicit permission for their identifiable or sensitive data to be shared with other researchers. Databrary makes available on its website template language researchers may use for seeking permission to share, along with video demonstrations and additional guidance. So, researchers who wish to share video must either use the template or demonstrate that participants have granted an equivalent level of permission. The second component involves limiting access to individuals who have prior formal authorization. Researchers must either secure authorization to access Databrary from their institution, typically a university where they conduct independent research, or must have another authorized researcher grant them access. Institution-level authorization requires a formal agreement with Databrary. Fundamentally, the Databrary framework ensures that participants control whether other researchers can view their video (or other identifiable data), and only researchers who pledge to uphold ethical principles, including maintaining participant confidentiality, may access Databrary's sensitive materials.

Databrary developed these policies and agreements in close consultation with Institutional Review Boards (IRBs), sponsored research privacy administrators, and legal staff at NYU and Penn State, with other experts in research data ethics, and with Federal officials. See http://databrary.org/access/guide.html for more details.

## Databrary Software

The centerpiece of the Databrary project is the repository itself. Our team of programmers has developed digital library software, including asset management, workflow, access control, discovery services, and a user interface. This work is closely coordinated with the central NYU IT infrastructure providers for services such as high-capacity storage, high-security storage, and high-performance computing (for video processing). The long-term goal is to build the cyber-infrastructure for seamless search, streaming, uploading, and downloading and to provide an open-source framework for other entities to build on the Databrary software or extend it to other content domains.

## Annotation and Metadata

Video becomes most useful for conducting research when it is transformed into quantitative and qualitative data in a flat-file format. Thus, the Databrary team is enhancing a general-purpose video-coding software application specialized for exploring, annotating, tagging, scoring, and visualizing the video files in preparation for qualitative and quantitative analyses. Toward that end, we have released Datavyu (datayvu.org), a free, open source, multi-platform desktop application. Datavyu is a fork of the OpenSHAPA tool used by some developmental researchers. The long-term goal is to build a web-based version of Datavyu that will provide both data analysis and data management services required for preparing and organizing video material prior to its storage in Databrary.

## Community Practices

The current zeitgeist in the developmental science community is one of privacy and data hoarding. A primary activity of the Databrary project is to introduce the practice of open data sharing to the developmental science community. We are doing so through announcements to developmental research societies, colloquia, and conference presentations. The team participates in workshops to discuss sharing issues and hosts events to teach potential users about sharing and analysis tools. We post links on websites frequented by developmental researchers and participate in user forums. In addition, we have asked leaders in developmental science to set an example for other researchers by publicly committing to open video data sharing. More than a hundred have responded to the call (http://databrary.org/community/contributors.html). Our long-term goal is a self-sustaining community of researchers committed to open science..

# BUILDING THE LIBRARY

Existing data management practices pose challenges for video sharing. Despite similar research methods, study designs vary widely. No two developmental science labs manage data in the same way. Some studies are longitudinal, where researchers observe the same participants at multiple sessions. Some are cross-sectional, where researchers observe each participant at only one session. The timing of observations may be determined by participants’ age (e.g., 4-month-olds, 8-month-olds, 12-month-olds) or abilities (e.g., preverbal, one-word utterances, sentences), experimentally determined variables (e.g., pre- and post-intervention), or other factors. Some labs organize longitudinal data by grouping files first by participant and then by session date. Other labs organize longitudinal data by session and then by participant. Still others organize longitudinal data based on task (book reading, block building, free play). Some researchers institute a central data management system to be followed by all the lab members, providing easier access and greater transparency for the entire lab but not necessarily providing the structure for similar benefits to researchers unfamiliar with the lab’s practices. Other researchers allow their students to keep separate records, making it difficult to share data even within the lab. Some researchers keep videos, metadata, and analyses together, and some do not. Idiosyncratic terms, record-keeping, and data management systems are the norm. Databrary must enable researchers to discover and understand each other’s materials, regardless of the original investigator’s data management system.

Databrary takes several approaches to address this challenge: A flexible data model, informal metadata contributions by users via tagging and commenting, mediated curation, and incentives to deposit materials more consistently. Figure 1 illustrates the principal data model. It accommodates different hierarchies and organizations, such as those described in the examples above.

[ FIGURE 1 ABOUT HERE ]

The full expressive flexibility of the data model is not evident to Databrary users. Instead, users see different aspects of the data model at different times, depending on the context. Researchers organize their materials by acquisition date and time into structures called sessions. A session corresponds to a unique recording episode and contains one or more recordings or related flat files (assets). Researchers can group sessions in different ways, using whatever groupings are appropriate for the particular dataset. These groupings may also identify particular time segments of a recording to distinguish tasks, events, or participants that comprise the session. In the data model, the aggregate of these flexible groupings is called a volume. In practice, researchers combine sessions or segments of sessions within and across datasets to form the raw material that are subsequently described in published articles and presentations. Thus, Databrary contributors can combine sessions or segments within and across datasets with ancillary materials, such as coding manuals, Datavyu spreadsheets, statistical analyses, questionnaires, IRB documents, computer code, sample displays, and links to published journal articles in an aggregate structure investigators call a study. Like datasets, studies are represented as volumes in the data model.

The discovery and browsing interfaces offer content items at the volume (study or dataset) level to users searching the library. The contributor-assigned groupings (records), which may include relevant metadata about participants (measures, such as participant demographic information, domain-specific survey information, location data, condition variables, task descriptions, or other properties), allow a second level of filtering and organization within and between relevant identified studies or datasets. This enables users to quickly identify data that may be relevant to their own interests or research.

Authorized Databrary Investigators can also add keyword tags at the study/dataset, session, or segment level, and can endorse or deprecate keywords that have been suggested by other authorized researchers. These abilities are critical because different researchers may have very different reasons for citing a study or using data. Future implementations may extend the ability of authorized researchers to annotate studies with other kinds of metadata.

The Databrary team recognizes that other data repositories enforce strict metadata ontologies and that doing so may have benefits in sub-domains of research when there is community consensus [12]. We will support standard data coding ontologies among researchers, but only enforce standardization for a small set of standard tags such as study date, participant birth date, and sex. We will also encourage contributors to report race, ethnicity, primary language, language of dataset, and location of session. The Databrary team chose not to require strict metadata ontologies for several reasons. We hope to reduce the pre-deposit curational demands on contributors, encourage the repurposing of shared data, and foster the rapid adoption of data sharing. Developing and achieving agreement on metadata ontologies can take significant amounts of time. Video data are so rich and complex that in many domains, researchers have not settled on standard definitions for particular behaviors and may have little current need for standard tasks, procedures, or terminology. However, we will also encourage users to re-use tags and terminology by suggesting common or similar terms, without confining users to these suggestions. User communities within Databrary may eventually converge on common conceptual and metadata ontologies based on the most common (and commonly endorsed) keyword tags, but standardized ontologies are not necessary for browsing and searching in most of the use cases we envision.

Our experiences curating and ingesting archival datasets have highlighted the considerable value of contributors entering raw video data into Databrary as soon as recordings are acquired. Immediate uploading reduces the workload on investigators, minimizes the risk of data loss and corruption, and accelerates the speed with which materials become openly available. To encourage immediate uploading, Databrary provides a complete set of controls so that researchers can restrict access to their data to only their own labs or to other users of their choosing. Datasets can be shared at a later point when data collection and ancillary materials are complete, whenever the contributor is comfortable sharing, or when journals or funders require it. Databrary has published a Data Sharing Manifesto [13] that explains to researchers the Databrary philosophy. Standards about when data should be shared are evolving. Our philosophy is consistent with concepts and practices in other domains where data sharing is the norm (e.g., www.iedata.org). Planned enhancements to the Datavyu tool will allow contributors to organize video files and metadata as part of the analysis process. This will facilitate the contribution of packaged datasets to Databrary.

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 (HiP) 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 digital video formats in recent years.

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 METS [14] structural metadata 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.

# POLICIES FRAMEWORK

Sharing video recordings poses a unique challenge to existing data sharing policies because videos contain personally identifying information—specifically participants’ faces and voices and often the insides of their homes and classrooms. Sharing personally identifiable information puts research participants and others depicted in recordings at increased risk for loss of privacy. At the same time, blurring or altering original recordings to hide identities undermines or eliminates their value to other researchers. Often, participants’ faces and voices produce the behaviors of interest. So, Databrary has elected to maximize the potential for data re-use by keeping recordings in their original unaltered form. Instead of removing participants’ identities, Databrary restricts access to identifiable or sensitive data to authorized researchers. Further, Databrary provides access only when the people depicted have given permission for their information to be shared with other researchers.

## Restricting Access.

Databrary provides access to shared data only to authorized researchers who have agreed to uphold common practices concerning the responsible and ethical use of identifiable and sensitive data. To become an authorized investigator, applicants must register on the site and electronically sign the Authorized Investigator Agreement, which must also be co-signed by the applicant’s institution. Full privileges will be granted only to researchers with principal investigator (PI) status at their institutions. Other researchers may be granted privileges if they are affiliated with a PI who agrees to sponsor and supervise their application. Initially there will be a manual process to identify the institutional representative—typically the authorizing official of the university—who can co-sign the Investigator Agreement. However, as the user groups at each university expand, Databrary may implement administrative accounts at each institution. This will enable the authorizing official to independently manage the authorizations of individual researchers at her institution.

4.2 Seeking permission to share identifiable data.

Data from a particular session may be stored in Databrary for the contributing researcher’s use whether the records are shared with other scientists or not. When a researcher chooses to share, Databrary makes records openly available to the community of authorized researchers only if the people depicted in the recordings have given permission to release the data for sharing. Thus, Databrary requires that people depicted in recordings grant permission before their information can be shared. Databrary’s policies extend currently accepted principles of informed consent to the situation where participants are granted authority to consent to (or refuse) the release of their identifiable data.

We developed these ideas in close collaboration with the NYU and PSU IRB staff. To formalize the process of acquiring permission, we developed a Participant Release Form Template, based on photo or video release language many researchers use currently. The template release form has standard language that Databrary recommends investigators should use with study participants. This language makes it easy for participants to understand what is involved in sharing their video data, with whom it will be shared, and the potential risks associated with releasing their video and other identifiable data to other researchers. Use of the template also allows for the standardization of language associated with the release of identifiable or sensitive participant data.

## Technical assistance with IRBs.

Some IRBs may deem an investigator’s existing, approved video or photo release form equivalent to the Databrary release. This enables a researcher to share with Databrary recordings they have already collected. However, most researchers will need to modify their research protocols, by adding the Databrary sharing permission procedures, prior to collecting new shareable video data. Databrary staff are available to advise potential data contributors about how to amend existing research protocols so that the information acquired is Databrary-compliant. Protocol amendments involve seeking approval for use of the Databrary template release form and modifying the time period over which collected data will be made available. Specifically, researchers must remove any clauses in research consent documents that require data destruction after some fixed period of time since Databrary intends to store shared data indefinitely.

# IMPACT

The Databrary project has already begun to have an impact. As of mid-January 2015, more than 57 scientists from 35 institutions in North America, the UK, and Europe have received authorization for full access to and data sharing with Databrary. More than 50 additional researchers are in the process of securing institutional approval. In consultation with Databrary staff, many of these researchers have secured or are in the process of securing permission from their research ethics boards to share archival or new data. Thus, the Databrary project has laid the groundwork for change in scientific culture around data sharing within the community of developmental science researchers who are most familiar with it. In a similar vein, Databrary is attempting to lead by example where open science practices are concerned. Databrary is an open source project. The entire code base is available on Github (github.com/databrary) as are all policy documents. The project team consults regularly with other leaders in the open science, data sharing, and data repository communities, and we share best practices among them.

Databrary has also begun to have an impact in the policy arena. The notion that identifiable research data may be shared – under the right circumstances – is not new to Databrary [15,16]. But, Databrary has created a set of policies and template documents that will help IRBs to come to see that seeking permission to share data is merely an extension of the principle of informed consent. Furthermore, Databrary’s Authorized Investigator Agreement combines provisions for data contribution with those guiding data use, part of an overall effort to reduce barriers to sharing and re-use. Our colleagues in the data sharing community tell us that this combination represents an innovation in itself.

The Databrary team has come to understand that laboratory data management practices pose significant challenges to widespread data sharing. Simply put, many researchers deploy workflows that would require significant modification in order for video files and associated metadata to be readily shared. In response, Databrary has expanded its data curation expertise and capacity. However, hand-curating significant volumes of research data in this way is not sustainable in the long term. Building tools that can enable self-curation will be essential to the future success and sustainability of Databrary. As a result, the team has begun to compile best practices for data management and curation that will be folded into future data coding (Datavyu) and data library (Databrary) features. We note that the process of understanding the diversity of data management techniques in developmental science poses unique opportunities for librarians and curators to learn how researchers organize and manage their data for daily use. It also affords researchers opportunities to learn from their library science colleagues about the value of adopting best practices in data organization and management.

# CONCLUSIONS

The Databrary project aims to increase scientific transparency and accelerate discovery in developmental science by creating a user-friendly and powerful infrastructure for researchers to share video and related data. Clearly, sharing video data poses technical and policy challenges, but it presents significant opportunities for accelerating discovery if these challenges can be met successfully. The Databrary project has already made significant strides in identifying and overcoming many of the obstacles, and the tools and infrastructure that we develop promise to enhance data sharing and management practices in the entire behavioral science community.

# ACKNOWLEDGMENTS

This work was supported by the National Science Foundation (BCS-1238599) and the National Institute of Child Health and Human Development (U01-HD-076595-01). The authors gratefully acknowledge the NYU Libraries, Human Connectome Project, the Personal Genome Project, the Inter-university Consortium for Political and Social Research, the Center for Open Science, Dataverse, Data Dryad, and staff from the National Database for Autism Research for their valuable advice and consultation.

# REFERENCES

1. Curry A (2011) Rescue of Old Data Offers Lesson for Particle Physicists. Science 331:694-695. doi:10.1126/science.331.6018.694
2. Integrated Earth Data Applications. (2014). http://www.iedadata.org. Accessed 10 February 2014
3. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P (2009) Data sharing in genomics [mdash] re-shaping scientific practice. Nat Rev Genet 10 (5):331-335. doi:10.1038/nrg2573
4. Overpeck JT, Meehl GA, Bony S, Easterling DR (2011) Climate Data Challenges in the 21st Century. Science 331 (6018):700-702. doi:10.1126/science.1197869
5. Lunshof J, Church G, Prainsack B (2014) Raw Personal Data: Providing Access. Science 343 (6169):373-374. doi:10.1126/science.1249382
6. Adolph K, Gilmore RO, Freeman C, Sanderson P, Millman D (2012) Toward Open Behavioral Science. Psychological Inquiry: An International Journal for the Advancement of Psychological Theory 23 (3):244-247. doi:10.1080/1047840X.2012.705133
7. King G (2011) Ensuring the Data-Rich Future of the Social Sciences. Science 331 (6018):719-721. doi:10.1126/science.1197872.
8. Association for Psychological Science (2012) Special Section on Replicability in Psychological Science: A Crisis of Confidence. Perspectives on Psychological Science 7 (6)
9. Nosek BA, Bar-Anan Y (2012) Scientific Utopia: I. Opening Scientific Communication. Psychological Inquiry 23 (3):217-243. doi:10.1080/1047840X.2012.692215
10. Nosek BA, Spies JR, Motyl M (2012) Scientific Utopia: II. Restructuring Incentives and Practices to Promote Truth Over Publishability. Perspectives on Psychological Science 7 (6):615-631. doi:10.1177/1745691612459058
11. MacWhinney B (2001) From CHILDES to TalkBank. In: M. Almgren AB, M. Ezeizaberrena, I. Idiazabal & B. MacWhinney (ed) Research on Child Language Acquisition. Cascadilla, Massachusetts, pp 17-34
12. Cognitive Atlas. (2014). http://www.cognitiveatlas.org. Accessed 10 February 2014
13. Databrary (2013) Databrary Data Sharing Manifesto. http://databrary.org/user-guide/policies/data-sharing-manifesto.html. Accessed 10 February 2014
14. METS (2014) XSLT-based Metadata Encoding and Transcription Standards Viewers. http://dlib.nyu.edu/metstools/metsviewer/. Accessed 10 February 2014
15. HCP (2014) Human Connectome Project. http://www.humanconnectomeproject.org. Accessed 10 February 2014
16. PGP (2014) Personal Genome Project. http://www.personalgenomes.org. Accessed 10 February 2014