The Role of Video Data Sharing in Advancing Education Policy and Practice

David S. Millman

The Databrary (databrary.org) Project

Abstract

Video captures the complexity, richness, and diversity of behavior unlike any other measure. As a result, large numbers of people who study teaching and learning or seek to improve it employ video. Video documents itself to a large degree, and therefore has significant potential for reuse by others. This potential remains largely unrealized because videos are rarely shared. 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 and software tools for coding pose technical challenges. In this thought paper, we will describe how the Databrary data library has overcome the most significant barriers to sharing video within the developmental sciences community, including solutions to maintaining participant privacy, data tagging, and data management. Databrary suggests ways that video and other identifiable data collected in the context of education research might be shared. Widespread open sharing of high value, high impact data will advance education policy and improve practice.

The Role of Video Data Sharing in Advancing Education Policy and Practice

Introduction

The application of insights from research to essential human needs such as improved educational policy and practice depends on the free flow of information. Open data sharing promotes greater transparency and peer oversight, supports inquiry into data collection methods and measurement, stimulates new work, encourages diversity of analysis and opinion, allows for testing of new or alternative hypotheses, facilitates the education of new researchers, permits integrative analyses by creating new datasets from multiple sources, allows future research to build on earlier efforts, and enables exploration of topics not envisioned by the original investigators. Open data sharing increases the impact of public investments in research and leads to better and more effective public policy. Open data sharing is a scientific imperative and common practice in many areas of biomedical, physical, and earth sciences where it has accelerated the pace of discovery. Unfortunately, research on human development and learning remains shrouded in a culture of isolation. Rather than providing direct access to raw data, researchers in the developmental and learning sciences share interpretations of distilled data almost exclusively through publications and presentations. The path from raw data to finding to conclusion cannot be traced or validated by others, nor can new questions be posed by others building on the same, often expensive-to-collect, raw data. To realize significant advances in our understanding of teaching and learning and for those insights to have meaningful impact on educational practice and policy, open data sharing must become an established norm, not the rare exception. Moreover, we must find ways to share the most valuable raw data, the data with the greatest potential for reuse by others, and the data with the greatest potential for yielding meaningful insights into the learning process.

Video is a uniquely rich medium for capturing in real time the interactions that occur in classrooms, muesuems, laboratory, home, and other informal learning settings. Many researchers and practitioners in the developmental, learning, and education sciences collect

video as raw data. Video recordings capture what teachers say and do and what students say and do in ways that offer the potential to discover when, why, and how learning actually works. Still, the open sharing of video data is rare, even within labs or among collaborators. Video data should be conducive to sharing. Video captures behavior that others can readily inspect. It captures multiple dimensions of complex real-world learning settings in ways that others can easily build upon. But, technical, ethical, and cultural barriers have made open sharing of video data rare. Videos come in varied formats and generage large files; these pose conversion and storage challenges. Personally identifying information can be removed from text-based data. Videos may contain faces, voices, names spoken aloud, views or voices of non-participants, views of classroom interiors and sometimes views of the homes of research participants. These elements cannot be removed without reducing the information content and reuse value to others. The collection of video or other identifiable or sensitive information requires approval by a research ethics board. It also requires informed consent or assent from the participants, and possibly parental, teacher, and school district approval. Researchers risk violating participants' privacy if digital images are viewed or released to others without authorization. For these and other reasons, no culture of widespread open sharing of video data has emerged in the education, learning, and developmental sciences.

The Databrary project has built a digital data library specialized for the open sharing of video data from research in these areas of inquiry. In this paper, we will describe how Databrary has overcome the most significant barriers to sharing video, including solutions to maintaining participant privacy, data tagging, and data management. In developing technology and policies that enable the safe and secure sharing of video and other identifiable research data, Databrary suggests ways that video and other identifiable data collected in the context of research on teaching and learning might be shared. Fostering the open sharing and reuse of high value, high impact data like video promises to advance education policy and improve practice.

The Promise of Video Data Sharing

More than 1 billion users worldwide uploade more than 300 hours of video to YouTube every minute (https://www.youtube.com/yt/press/statistics.html). Google, YouTube's owner, has built an immense data infrastructure to upload, store, convert, tag, and stream video. The current scale of video collection in developmental and education contexts is undoubtedly smaller, but it is not small. The Gates Foundation-funded Measures of Effective Teaching (MET) Project, hosted at ICPSR, contains X video segments from X classroom settings, representing tens of terabytes of data. In just over a year of operation, the Databrary digital library has collected more than 7,000 individual videos, representing 2,400 hours of recording, featuring more than 1,800 infant, child, and adult participants. Video data is big, and the interest in recording and sharing video for research, education, and policy purposes has grown.

Video is Uniquely Rich

Video Uses in Education

We record teachers in training and teachers in-service for evaluation. We record classrooms to capture what teachers do and how students respond. We record visitors and docents in museum settings.

The Challenges of Video Data Sharing

Sharing video poses challenges. These include technical infrastructure, research ethics, privacy, and security, and data management. Further, making open data sharing in education research an established norm requires changing understanding about how identifiable data might be shared and building a community of researchers and educators committed to making video data sharing work.

Technical

Video poses two types of technical challenges related to the goal of increasing widespread sharing. One concerns the requirements for storing, sharing, streaming, and preserving large numbers of videos. The other concerns tools for extracting meaningful information from the episodes and behavior videos capture.

Video Data is Big. Video often accompanies other data streams (physiological recordings, brain imaging, EMG, motion tracking, eye tracking, verbal transcripts). Multiple data streams—including two or more camera views—require tools for synchronization and integration and benefit from multivariate time series analyses. Many labs lack the resources to adopt tools for detailed video coding and complex analyses. Furthermore, widespread use of video creates a data explosion: The typical developmental lab collects 8-12 hours of video/week 18 in widely varied formats. Thus, sharing digital video requires substantial storage capacity, powerful search and streaming tools, and significant computational resources for transcoding videos into common, preservable formats.

Tools for Coding Video. After videos are recorded, human coders evaluate videos and apply text-based annotations or numeric codes to videos. Coders use a wide-range of commercial and academic software tools: Transana, StudioCode, V-code, MaxQDA, Noldus Observer, Mangold Interact, and Datavyu, among others. Some use spreadsheets or pencil and paper. The tools have no standard format for data storage or export. Some offer web/cloud based storage (e.g., Transana, V-code), but otherwise coded data may not be easily shared within tools, much less between them. With few exceptions, machine learning has not been applied to video data coding.

Ethics and Privacy

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. A challenge for future education research is to find ways to preserve the value of video for reuse by others while protecting the privacy of research participants. Databrary has developed an access model described below that provides an example of how this might be accomplished.

Data Management

A primary impediment to sharing is researchers' lack of time to find, label, clean, organize, and copy their files into formats that can be used and understood by others15. Even scientists committed to data sharing find it difficult to do. Data management systems that make sharing convenient, reliable—and optimally automatic29—can help.

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. As in other fields22, developmental science is inundated by an explosion of digital data, most of which is inaccessible to other researchers. The data deluge stems from the advent of cheap, high-resolution digital cameras. The average developmental lab collects 12 hours of video/week23. Raw data without information about workflow and data provenance is essentially "orphaned" and cannot be understood by subsequent researchers who weren't part of the original team24. Moreover, video files must be electronically linked with participant permissions and relevant metadata. An effective data management system for video does not currently exist. We will build one to allow researchers within labs to benefit from each other's work, facilitate interdisciplinary collaborations, and provide a solid foundation for open sharing25

Changing Community Practices

Data sharing is an established best practice in some areas of educational research (AERA ethics guide), but it is not necessarily widespread.

Databrary.org

Databrary (databrary.org) is a digital data library specialized for storing and sharing video and research data and metadata. 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. No large-scale repository for sharing video data currently exists. So, we will create a web-based Databrary repository.

System Design

Databary is built on a PostreSQL database using the Scala Play framework and JavaScript. Data are preserved indefinitely in a secure data storage facility at NYU managed by central IT. There is no cost to use the system; an institutional subscription model is under development. All code is hosted on GitHub. Databrary can house video, audio, PDF, spreadsheet, image, and text-based files (along with associated metadata), as well as executable scripts in Ruby, R and Matlab that facilitate data analysis. Video and audio data are transcoded into standard and HTML5-compatible formats, currently H.264+AAC as MP4. Databrary stores other data in native formats (e.g., .doc, .docx, .xls, .xlsx, .txt, .csv, .pdf, .jpg, .png).

Common video and coding file formats are required for users to build on earlier coding efforts, perform integrative analyses across shared files, and search effectively in a shared repository. We will develop tools for converting files from other coding tools into a common format for the Databrary.

Behavior is rich and complex. Research interests and methods are diverse. Naturally, video coding is idiosyncratic. The only universal tags in developmental research are children's age and sex. Thus, OpenSHAPA is powerful and flexible. It is designed to help

researchers understand behavior—or any other time series—using codes of their choosing. As the PI's experience demonstrates, OpenSHAPA will strengthen work within labs. It will also provide the common coding format for files in the Databrary.

Policies for Safe, Secure, & Sharing

Shared videos are maximally informative when participants' faces are visible and their vocalizations are audible. Thus, anonymity cannot be ensured as it is for flat file or imaging data. To address issues of confidentiality, access, and IRB approval for sharing video data, we will develop template data sharing permission forms and contributor and user agreements with input from the community, from human compliance officers, and from leaders of previous data sharing efforts. 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 peopledepicted 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 an Access 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.

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.

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.

Managing Data for Sharing

Raw video is most informative when it is linked with coding spreadsheets, codebooks, protocols, statistical analyses, and manuscripts. A data management system is required to link related files, document workflow and data provenance, and tag files with appropriate permission levels for sharing. Consistent data management practices will strengthen work within labs and among collaborators, enable convenient and reliable file uploading, and enhance the value of shared data in the Databrary.

The 2011 NSF workshop8 and PI Adolph's MacSHAPA/OpenSHAPA workshops over the last decade have revealed widely varying laboratory data management practices. Some minimal standardization across labs is required to make data sharing feasible. In the biomedical sciences, use of laboratory information systems (LIMS) is increasingly common and shows promise for lowering barriers to sharing in other fields15. Consistent with our overall approach, we will research, adapt, and adopt best laboratory management practices and open-source tools from other fields34, and we will prioritize implementation based on the preferences and priorities of the developmental science community.

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.

Based on discussions at the NSF workshop, we expect to see most contributions after

data collection, coding, and analyses are completed. Researchers have widely differing practices for designing codebooks and spreadsheets, managing data collections, coding, analyses, collecting and managing participant data (demographic, anthropometric, etc.), and documenting workflow and data provenance. Our goal for Years 2-3 is to incorporate this more detailed type of information into the data management system. Doing so will strengthen sharing within labs, expand the usefulness of Databrary, and enable more powerful searches. The data management system will be extended to organize and package other files associated with contributors (position, expertise), studies (codebooks, protocols, analysis spreadsheets graphs, manuscripts, presentations, grant proposals, etc.), and participants (coded spreadsheets, test results, demographics, etc.). The same basic ideas will apply— standard file naming conventions and data tagged by contributor, study, and participant level. OpenSHAPA and other applications will communicate with the data management tool. The data management tool will be used to stitch together all the sharable files that comprise a study. From data sharing to best lab practices. The developmental science community is hungry for new practices that will enhance research, as demonstrated by the number of committed contributors. LIMS in other fields embody additional functions. We envision that the data and tool sharing enabled by Databrary will spark interest in other lab management enhancements. For example, the data management system could facilitate participant recruitment and scheduling, task/project management and scheduling, workflow processing, and quality assurance (e.g., reliability testing). We plan to seek community input—via online comments, conference workshops on best lab practices, and other avenues—about which features should be the highest priorities for data management tool developers. Extending the data management tool to incorporate the highest priority features will be a focus of the front-end programmer in Years 4-5.

Building A Community That Embracing Sharing

Data sharing works only when the scientific community embraces it. To make video-based data sharing succeed, we will build a diverse contributor/user community committed to open data sharing. Based on continual community input, we will expand the OpenSHAPA video coding tool, build data management tools, and develop tools for uploading, browsing, and searching data in the Databrary repository. These tools will enable sharing within labs, among collaborators, and among users in the larger community. To avoid pitfalls of previous data sharing efforts, leaders in data sharing, developmental science, and computer science will guide our work through an advisory board. To ensure continued community engagement, Databrary personnel will provide technical support to individual labs and provide information to the community about new data contributions, citations of shared data, and use cases.

Conclusion

As Gesell once noted, cameras can record child behavior in ways that make it "...as tangible as tissue."38 The Databrary project will strengthen traditional ways of conducting video-based research by facilitating better coding, data mining, and data management practices within labs while building the culture and infrastructure to enable a new tradition of open data sharing in developmental science. The Databrary project is potentially transformative. We will increase transparency by sharing the traditionally private activity of data collection, coding, and analysis. We will speed progress by enabling later researchers to benefit from previous work. Ultimately, we will better understand human development and how to transform knowledge into improvements in health.

Acknowledgments

Databrary is based on work supported by the National Science Foundation under Grant No. BCS-1238599, the Eunice Kennedy Shriver National Institute of Child Health

and Human Development under Cooperative Agreement U01-HD-076595, and the Society for Research in Child Development. Any opinions, findings, and conclusions or recommendations expressed in the material contributed here are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, the Eunice Kennedy Shriver National Institute of Child Health and Human Development, or the Society for Research in Child Development.