Open Video Data Sharing Can Transform Education Research

Rick O. Gilmore

The Pennsylvania State University, The Databrary Project

Karen E. Adolph, David S. Millman

New York University, The Databrary Project

Author Note

Rick O. Gilmore is in the Department of Psychology, The Pennsylvania State University, University Park, PA 16802, rogilmore@psu.edu. Karen E. Adolph is in the Department of Psychology at New York University. David S. Millman is with the NYU Library. 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.

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 employ video. Video documents itself to a large degree. This presents significant potential for reuse by others. The potential remains largely unrealized because videos are rarely shared. Video contains information about personal identities. This poses challenges to sharing. The large size of video files, diversity of formats, and incompatible software tools pose technical challenges. We describe how the Databrary data library has overcome the most significant barriers to sharing video within the developmental sciences community. Databrary has developed solutions to maintaining participant privacy, storing, streaming, and sharing video, and for managing video datasets and associated metadata. The Databrary experience suggests ways that video and other identifiable data collected in the context of education research might be shared. We envision a data intensive science of teaching and learning, with video as its core, that allows educational experiences to be tailored to students in ways that big data promises to personalize medicine. The creation and support of repositories that enable the open sharing of dense, richly informative, high value, and high impact data about teaching and learning will help realize this ambitious vision.

Open Video Data Sharing Can Transform Education Research

Introduction

Open data sharing can help to translate insights from scientific research into applications serving essential human needs. Open data sharing bolsters transparency and peer oversight, encourages diversity of analysis and opinion, accelerates the education of new researchers, and stimulates the exploration of new topics not envisioned by the original investigators. Data sharing and reuse increases the impact of public investments in research and leads to more effective public policy. Although many researchers in the developmental, learning, and education sciences collect video as raw research data, most research on human learning and development remains shrouded in a culture of isolation (?, ?). Researchers share interpretations of distilled, not raw data, almost exclusively through publications and presentations. The path from raw video to research findings to conclusions cannot be traced or validated by others. Other researchers cannot pose new questions that build on the same raw materials. This paper describes how the Databrary data library has overcome the most significant barriers to sharing video within the developmental sciences community. It highlights how open video data sharing might improve scientific practice and advance research on learning and development.

The Promise and Challenge of Video

Video is a uniquely rich, inexpensive, and adaptable medium for capturing the complex dynamics of behavior. Researchers use video in home and laboratory contexts to study how infants, children, and adults behave in natural or experimenter-imposed tasks (?, ?). Researchers record videos of students in classrooms (?, ?) to understand what teachers do and how students respond. Because video closely mimics the multisensory experiences of live human observers, recordings collected by one person for a particular purpose may be readily understood by another person and reused for a different purpose. Moreover, the success of YouTube and other video-based social media demonstrates that web-based video storage and streaming systems are now sufficiently well developed to satisfy large-scale demand. The question for researchers and policymakers is how to capitalize on video's potential to improve teaching and learning.

The answer requires overcoming significant technical, ethical, practical, and cultural challenges to sharing research video. *File sizes and diverse formats present special challenges* for sharing. Video files are large (one hour of HD video can consume 10+ GB of storage) and come in varied formats (from cell phones to high-speed video). Many studies require multiple camera views to capture desired behaviors. Research video creates a data explosion: A typical lab studying infant or child development collects 8-12 hours of video/week (?, ?). Thus, sharing videos requires substantial storage capacity and significant computational resources for transcoding videos into common, preservable formats.

Technical challenges involved in searching the contents of videos present barriers to sharing. Videos contain rich and diverse information that requires significant effort by human observers to extract. Researchers make use of videos by watching them and, using paper and pencil or more automated computerized coding software, translating observations into ideas and numbers. In many cases, researchers assign codes to particular portions of videos. These codes make the contents of videos searchable by others, in principle. However, researchers focus on different questions from varied theoretical perspectives and lack consensus on conceptual ontologies. So, in practice, most coded data are not easily shared. Although human-centered video coding capitalizes on the unique abilities of trained observers to capture important dimensions of behavior, machine learning and computer vision tools may provide new avenues for tagging the contents of videos for educational and developmental research (?, ?, ?, ?, ?, ?).

Open video sharing must overcome *ethical challenges* linked to sharing personally identifiable data. Although policies exist for sharing de-identified data, video contains easily identifiable data: faces, voices, names, interiors of homes and classrooms, and so on. Removing identifiable information from video severely diminishes its reuse value and poses additional burdens on researchers. So, open video sharing requires new policies that protect the privacy of research participants while preserving the integrity of raw video for reuse by others.

Open video sharing faces practical *challenges of data management*. Developmental and education research is inundated by an explosion of data, most of which is inaccessible to other researchers. Researchers lack time to find, label, clean, organize, and copy their files into formats that can be used and understood by others (?, ?). Study designs vary widely, and no two labs manage data in the same way. Idiosyncratic terms, record-keeping, and data management practices are the norm. Few researchers document workflows or data provenance. Although video requires minimal metadata to be useful, video files must be electronically linked to what relevant metadata exist including information whether participants have given permission to share.

Perhaps the most important *challenge is cultural*–community practices must change. Most researchers in the education, learning, and developmental sciences do not reuse their own videos or videos collected by other researchers; they neither recognize nor endorse the value of open sharing. Contributing data is anathema and justifications against sharing are many. Researchers cite intellectual property and privacy issues, the lack of data sharing requirements from funding agencies, and fears about the misuse, misinterpretation, or professional harm that might come from sharing (?, ?, ?). Data sharing diverts energy and resources from scholarly activities that are more heavily and frequently rewarded. These barriers must be overcome to make data sharing a scientific norm.

Databrary.org

The Databrary project has built a digital data library (http://databrary.org) specialized for open sharing of research videos. Databrary has overcome the most significant barriers to sharing video, including solutions to maintaining participant privacy, storing, streaming, and sharing video, and for managing video datasets and associated metadata. Databrary's technology and policies lay the groundwork for securely sharing research videos on teaching and learning. In only a year of operation, Databrary has collected more than 7,000 individual videos, representing 2,400 hours of recording, featuring more than 1,800 infant, child, and adult participants. Databrary has more than 100 authorized researchers representing more than 60 institutions across the globe. Video data is big data, and the interest in recording and sharing video for research, education, and policy purposes continues to grow.

The Databrary project (databrary.org) arose to meet the challenges of sharing research video and to deliver on the promise of open data sharing in educational and developmental science. With funding from NSF (BCS-1238599) and NIH (NICHD U01-HD-076595), Databrary has focused on building a data library specialized for video, creating data management tools, crafting new policies that enable video sharing, and fostering a community of researchers who embrace video sharing. Databrary also developed a free, open-source video annotation tool, Datavyu (http://datavyu.org). The project received funding in 2012-2013, began a private beta testing phase in the spring of 2014 and opened for public use in October 2014.

System Design

The Databrary system enables large numbers of video and related files to be uploaded, converted, organized, stored, streamed, and tagged. Databrary is a free, open-source (http://github.com/databrary) web application whose data are preserved indefinitely in a secure storage facility at NYU. Databrary can house video and audio files, along with associated materials, coding spreadsheets, and metadata. Video and audio data are transcoded into standard and HTML5-compatible formats. This ensures that video data can be streamed and downloaded by any operating system that supports a modern browser. Copies of original video files are also stored. Databrary stores other data in their original formats (e.g., .doc, .docx, .xls, .xlsx, .txt, .csv, .pdf, .jpg, .png).

The system's data model embodies flexibility. Researchers organize their materials by acquisition date and time into structures called *sessions*. A session corresponds to a unique recording episode featuring specific participants. It contains one or more videos and other file types and may be linked to user-defined metadata about the participants, tasks or measures, and locations. A group of sessions is called a *volume*. Databrary contributors may combine sessions or segments with coding manuals, coding spreadsheets, statistical analyses, questionnaires, IRB documents, computer code, sample displays, and links to published journal articles.

Databrary does not enforce strict ontologies for tagging volumes, sessions, or the contents of videos. 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 standardized tasks, procedures, or terminology. Indeed, standardized ontologies are not necessary for many use cases. Databrary empowers users to add keyword tags and to select terms that have been suggested by others without being confined to the suggestions. Moreover, Databrary encourages user communities within Databrary to converge on common conceptual and metadata ontologies based on the most common keyword tags, and to construct and enforce common procedures and tasks wherever this makes sense.

Future challenges include enhancing the capacity to search for tagged segments inside of videos. Some search functionality exists in the current software, with more extensive capabilities on the near horizon. A related challenge involves importing files from desktop video coding tools. This will allow for the visualization of user-supplied codes independent of the desktop software deployed in a particular project. We envision a parallel set of export functions that permit full interoperability among coding tools. The priority will be to create interoperability with tools using open, not proprietary file formats. Databrary also recognizes the need to develop open standards and interfaces that enable Databrary to link to and synchronize with outside sources that specialize in other data types.

Policies for Safe and Secure Video Sharing

Policies for openly sharing identifiable data in ways that securely preserve participant privacy are essential for sharing research video. Databrary does not attempt to de-identify videos. Instead, we maximize the potential for video reuse by keeping recordings in their original unaltered form. To make unaltered raw videos available to others for reuse, Databrary has developed a two-pronged access model that (a) restricts access to authorized researchers, and (b) enables access to identifiable data only with the explicit permission of participants.

To gain access to Databrary a person must register on the site. Applicants agree to uphold Databrary's ethical principles and to follow accepted practices concerning the responsible use of sensitive data. Each applicant's institution must co-sign an access agreement. Full privileges are granted only to those applicants with independent researcher status at their institutions. Others may be granted privileges if they are affiliated with a researcher who agrees to sponsor their application and supervise their use. Ethics board or IRB approval is not required to gain access to Databrary because many use cases do not involve research, but IRB approval is required for research uses. Once authorized, a user has full access to the site's shared data, and may browse, tag, download for later viewing, and conduct non- or pre-research activities.

Unique among data repositories, the Databrary access agreement authorizes both data use and contribution. However, users agree to store on Databrary only materials for which they have ethics board or IRB approval. Data may be stored on Databrary for the contributing researcher's use regardless of whether the records are shared with others or not. When a researcher chooses to share, Databrary makes the data openly available to the community of authorized researchers.

In addition to restricting access to authorized researchers, Databrary has extended the principle of informed consent to participate in research to encompass permission to share data with other researchers. To formalize the process of acquiring permission, Databrary has developed a Participant Release Template (?, ?) with standard language we recommended for use with study participants. This language helps participants to understand what is involved in sharing video data, with whom the data will be shared, and the potential risks of releasing video and other identifiable data to other researchers.

Managing Data for Sharing

When researchers *do* share, standard practice involves organizing data after a project has finished, perhaps when a paper goes to press. This “preparing for sharing” after the fact presents a difficult and unrewarding chore for investigators. It makes curating and ingesting datasets challenging for repositories, as well. Databrary has chosen a different route to curation.

We have developed a data management system that empowers researchers to upload and organize data as it is collected. 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. The system employs familiar, easy-to-use spreadsheet and timeline-based interfaces that allow users to upload videos, add metadata about tasks, settings, and participants, link related files, and assign appropriate permission levels for sharing. To encourage immediate uploading, Databrary provides a complete set of controls so that researchers can restrict access to their own labs or to other users of their choosing. Datasets can be openly shared with the broader research community 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.

Building a Community

Data sharing works only when the scientific community embraces it. From the beginning, Databrary has sought to cultivate a community of researchers who support data sharing and commit to enacting that support in their own work flows. Our community building efforts involve many interacting components. They include active engagement with professional associations, conference-based exhibits and training workshops, communications with research ethics and administration staff, talks and presentations to diverse audiences, and one-on-one consultations with individual researchers and research teams. These activities are time and labor-intensive, but we believe that they are critical to changing community attitudes toward data sharing in the educational and learning sciences. Looking ahead, it will be critical to engage funders, journals, and professional organizations in the effort to forge community consensus about the importance, feasibility, and potential of open video data sharing.

Conclusion

Imagine a time in the near future when researchers interested in studying classroom teaching and learning can mine an integrated, synchronized, interoperable, open and widely shared dataset. The components include video from multiple cameras, eye tracking, motion, and physiological measurements, and information from both historical and real-time student performance measures. Imagine that this classroom-level data can be linked with grade, school, neighborhood, community, region, and state-level data about education practice, curriculum, and policy. Then, imagine training a cadre of experts with skills in the data science of learning and education who are sensitive to privacy, confidentiality and ethical issues involved in research involving identifiable information. We empower these learning scientists to extract from the data meaningful insights about how educational practice and policy might be improved. In short, imagine a science of teaching and learning that can be personally tailored to individuals in ways analogous to the impact of big data on medicine. The barriers to realizing this vision are similar to those that confront the vision of personalized medicine – the development of technologies that enable data to be collected, synchronized, tagged, curated, stored, shared, linked, and aggregated; policies and practices that ensure security and individual privacy; and the cultivation of professional expertise needed to turn raw data into actionable insights.

As Gesell once noted, cameras can record behavior in ways that make it "...as tangible as tissue" (?, ?). The Databrary team contends that video has a central role to play in efforts to make tangible the anatomy of successful teaching and learning. In fact, we argue that video can be the core around which other measures of teaching and learning cluster. This requires reducing barriers to sharing video and fostering new community values around data sharing that make it indispensible. The Databrary project has built technology and policies that overcome many of the most significant barriers to widespread sharing within the developmental sciences community. Databrary suggests ways that video and other identifiable data collected in the context of education research might also be shared. Technologies and policies for providing secure access to videos for broader use cases will have to be developed, tools that allow desktop coding software files to be seamlessly converted to and from one another will have to be perfected, and ways of synchronizing and linking disparate data streams will have to be created. Equally important, communities of scholars dedicated to collecting, sharing, and mining education-related video data will have to be cultivated. But, we believe that the widespread sharing of high value, high impact data of the sort that video can provide promises to achieve this ambitious vision to advance education policy and improve practice. Databrary is working toward a future where open video data sharing is the norm, a personalized science of teaching and learning is the goal, and what optimizes student learning is as tangible as tissue.