Working with Authors to Curate Their Data

Andrew S. Gordon, Lisa Steiger, and Karen E. Adolph from the Databrary Project based at New York University provide an example of working with the author at this stage in the next case study.

Losing Research Data Due to Lack of Curation and Preservation

Andrew S. Gordon, Lisa Steiger, and Karen E. Adolph

Child development researchers have a problem: Their primary source of data is video, a medium with boundless potential for reuse; but researchers lack the community practices and infrastructure to make reuse possible. Researchers lack tools and standard practices across labs and institutions for organizing, storing, and ensuring long-term preservation of their video data, thus precluding reuse. This case study describes how Databrary developed a solution to the problem of data loss by enabling child development researchers to curate and describe their own research videos in an online repository that allows for sharing and reuse and provides long-term preservation. More generally, the processes and strategies described here can inform active, researcher-driven curation in other academic disciplines.

VIDEO DATA IN CHILD DEVELOPMENT RESEARCH

For most child development researchers, video recordings of children's behavior serve as the raw data for their research programs. Video is the medium of choice because it captures the richness and complexity of children's behavior easily and with high fidelity. Moreover, as a medium, video is especially well suited for analyzing behavior because it allows researchers to manipulate time. They can watch recorded behaviors unfold at various playback speeds, stop the video to focus on a particular event, and loop portions of the recording to better understand an event.

The general process is to record children's behavior and then score and analyze the recordings. The methods for collecting, annotating, and analyzing video footage are boundless. Videos can focus on experimental manipulations or natural, spontaneous expressions of behavior. Recordings can span a few seconds or several hours. Children can be observed once or across multiple time points. Observations can include a single child, multiple family members, peer groups, or entire school classrooms. With the recordings in hand, researchers then apply standard or user-defined tags to segments of the videos to generate quantitative data (e.g., frequency counts, rates, and durations) and qualitative data (e.g., ratings, ethnographic descriptions, conversation analyses, and narratives). Finally, researchers analyze the processed data to make inferences about child development.

Compared with other forms of research data (e.g., flat-file tabular data, imaging data, textual data), video offers unique potential for reuse. One aspect of research video that makes it eminently reusable is that important information about the raw data is directly accessible. The recordings convey visible and audible information about the original context in which they were collected. Viewers can see what is happening, what the researchers did, and what the children did. Thus, new investigators can reuse research videos with little additional information or documentation to perform integrative analyses or increase the diversity and size of their sample. A second related aspect of research video is the richness of the data that it contains. A close-up view of a child's face, for example, contains information about the child's facial expressions, vocalizations, and patterns of looking behavior, and how these various behaviors unfold over time. A wide shot view of a child at play could be used to study the development of walking or the development of social interaction. Thus, investigators can reuse existing raw research videos to explore entirely different phenomena and to ask new questions outside the scope of the original study.

Despite the unique advantages of video, video data sharing and reuse is not the norm in developmental science, and few platforms exist to motivate developmental researchers to publish their video data. Instead, most researchers collect videos for a single study and upon completion of the study, they allow the data to molder away on a hard drive or in a stack of DVDs in a cabinet. The process of collecting video data from children is expensive and time-consuming—data collection requires appropriate recording equipment, and lab staff must identify potential participants, schedule the appointments, run the data collection, and process the data. Thus, the field incurs a great loss when the use of a data set is limited to the scope of a single study in a single lab, and the life of a data set is only as long as the life of the media it is stored on.

Given the rich potential of video for research reuse and the importance of providing long-term preservation of these assets, we launched the Databrary repository in October 2014 to enable sharing, reuse, and indefinite preserva-

tion of raw research videos among child development researchers. Databrary is a Web-accessible repository, with access permissions set differently for authorized researchers and the public, depending on the data set and individual participant consent. The project—funded by the National Science Foundation (BCS-1238599) and the National Institute of Child Health and Human Development (U01-HD-076595)—is housed at New York University's Institute of Human Development and Social Change and collaborates closely with the university's libraries and the Information Technology Services.

POST HOC CURATION VERSUS ACTIVE CURATION

Post hoc curation (i.e., curation after all of the data have been collected and analyzed) is the most common way that researchers contribute data to domain repositories. So we initially assumed that this would be the primary means of acquiring data in Databrary. However, post hoc curation is hugely time-consuming and cumbersome to the data contributor.[12] To prepare the data for deposit, researchers must revisit data that they have already collected, analyzed, and stored away and now organize and describe it for the purpose of sharing. We quickly learned that the required commitment of time and personnel exceeded what most researchers were willing to do. Moreover, researchers lacked the expertise to prepare the data for ingest, so information professionals were needed to process the collection for sharing.[13]

To preempt these barriers to sharing, Databrary implemented tools to enable researchers to actively curate their own data immediately following each data collection—to organize, describe, and store the data at an early phase of the research life cycle, rather than as a burdensome final step at the end of the project. To encourage researchers to use these tools to organize, describe, and share their data, we built them with a strong emphasis on integrating the language and interfaces that our intended community was already using.

BUILDING A SYSTEM OF ACTIVE CURATION TO SUIT RESEARCHERS' NEEDS

Determining the best way to build active curation tools for the developmental science community required a clear understanding of researchers' workflows—in particular, the path from video data collection to storage of the video files and metadata. We started by interviewing a handful of representative researchers and their staff at NYU and other institutions who regularly collect video data and who represent the diversity of research in the developmental science community. The

interviews were unstructured and were intended to elicit details about researchers' current data management workflows and practices. We hoped that the interviews would inform us about what researchers might want from a service that would help them to organize, manage, store, and eventually share their videos. However, the interview results were only minimally informative about what researchers would want. Most researchers were not able to tell us how the various tasks of their day-to-day data management amounted to an explicit workflow, and it was evident that many of them had not previously considered how to prepare their data for sharing and reuse—in most cases, even for reuse within their own laboratories.

We realized that we needed information science professionals and domain experts working together to observe researchers' current practices and the tools they used (or lacked) in their labs. These observations allowed us to make inferences about the best ways to support active video curation for child development researchers. To obtain an understanding of how researchers collect, organize, and analyze their videos and metadata, we gathered a sample of data from each researcher we had originally interviewed to determine similarities and differences among data sets and lab practices. We learned that child development research is characterized by a wide diversity of data management practices, both within and across labs. As a result, data sets are heterogeneously structured and organized, which significantly increases the time required to prepare this data for post hoc ingesting into a repository.

Despite this diversity, we observed that researchers across labs care about emphasizing certain aspects of their research, such as what tasks were involved in a data collection (e.g., toy play, book reading, answering a set of question, watching displays on a computer monitor), whether the data collection was preliminary (i.e., intended to work the kinks out of the method) or part of the target dataset, and whether a participant had to be excluded from analyses for particular reasons (e.g., fussy or sleeping baby, equipment failure, experimenter error). Developing and implementing an approach to active curation required a focus on the evident similarities across different data sets and lab practices to gain traction with the intended research community.

Finally, we understood that our intended user group would require tangible incentives for adopting new practices for managing and storing their research data in a central repository. Notably, major disincentives include the cost to laboratories in terms of extra time and effort for post hoc curation and the financial cost of storing and serving video files. Thus, Databrary eliminated the cost of post hoc curation by replacing it with active curation tools that make trivial the cost of uploading and managing video recordings. Databrary creates incentives for active curation by providing free, unlimited storage; the repository acts as a convenient lab server to facilitate communication among lab members and with collaborators off site. After a study is complete and the contributors are ready to do so, sharing the data with the entire Databrary community is as simple as the push of a button.

ACTIVE CURATION IMPLEMENTED IN DATABRARY

HOW RESEARCHERS VIEW THEIR DATA

Active curation reduces friction for researchers contributing to a repository by merging the effort of describing and storing videos for sharing and reuse with data collection and organization. But first the intended community of researchers needed a more standardized process for organizing and describing different types of data sets. To accomplish this goal, we designed Databrary from the beginning to be a user-facing data repository that accommodates the diversity of existing data management practices we observed. The system relies on a metadata schema that reflects how researchers already view the different components of their video data sets (participant demographics, study conditions and tasks, geographic location, language of the participant, and so on).

For active curation tools to make sense to our intended user base, we chose to adopt researchers' language and organizing principles.[14] Developmental researchers call the analytic units of their studies "sessions."[15] A session is essentially a recording period. Within each session, we assign the general term *record* to the metadata that describes a session. Records comprise the important information about participants, activities, and researcher-defined conditions and groups as outlined in the metadata schema. The predefined records available in the upload interface were drawn from what we observed to be the most common types of metadata used across multiple labs, the metadata required for reporting by journals, and the metadata required by the major granting agencies.

INTERFACES THAT THE COMMUNITY ALREADY USES

Critically, to enable researchers to use our active curation service, we needed to craft interfaces that were easy to use and already familiar to the community. Spreadsheets are a common tool employed across labs to record session metadata. As a result, we designed a web application view that allows users to upload, modify, and manage session metadata into a spreadsheet, with features such as auto-completion, field pre-population, bulk editing, and suggested entries for convenience (figure 3.7). Rows of the spreadsheet correspond to individual sessions. The columns correspond to basic records describing the sessions, which helps researchers to manage their own data and assists other researchers in searching and finding videos of interest to them. Column categories in the spreadsheet are customizable and can be applied as needed to the researcher's study.

session			participant	task	context	file
name	test date	release	age	description	setting	name
	2014-XX-XX		4.6 yrs	4 tasks 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	3 files
	2014-XX-XX		3.9 yrs	4 tasks	Lab	3 files
	2014-XX-XX		4.6 yrs	4 tasks	Lab	3 files
	2014-XX-XX		4.6 yrs	4 tasks	Lab	3 files

FIGURE 3.7
Databrary metadata interface example. Spreadsheet metadata interface for a data set hosted on Databrary. Databrary exposes as much metadata about a study as possible without revealing sensitive or identifiable information (as determined by the participant and the data contributor). The availability of the metadata differs depending on the permission level of the user attempting access. Source: William Fabricius, "Absence of Construct Validity in Standard False Belief Tasks," Databrary, 2014, accessed November 10, 2015, doi:10.17910/B7Z300.

In addition to allowing researchers to add and modify record metadata through this interface, we provided tools to enhance researchers' ability to analyze their data. Allowing users to create, save, and share summaries of their metadata—such as number of participants by group, condition, or gender—gives them the power to quickly and easily gain insights into their data as they collect it (figure 3.8).

task	session			
description	name	test date	release	summary
3 option unexpected contents with neutral box	32 sessions			
3 option unexpected location task	32 sessions			
Standard two option typical box task	32 sessions			
3 option unexpected contents with typical box	32 sessions			
No task		2014-XX-XX		participant, context Lab

FIGURE 3.8
Metadata summary view. Display of summary information for the metadata from a data set. (Source: Catherine Tamis-LeMonda, "Language, Cognitive, and Socio-emotional Skills from 9 Months until Their Transition to First Grade in U.S. Children from African-American, Dominican, Mexican, and Chinese backgrounds," Databrary, 2013, accessed November 16, 2015, doi:10.17910/B7CC74.) This example shows the number of sessions for several tasks. Users can drag and drop metadata records to explore their data sets at a higher level of analysis.

Finally, most researchers annotate their research videos with a set of user-defined codes using desktop coding software such as Databrary's Datavyu, The Language Archive's ELAN, Mangold's Interact, Noldus's Observer, or the University of Wisconsin's Transana.[16] Thus, we implemented a time line view for managing the videos and metadata within sessions that is similar to these commonly used desktop coding tools (figure 3.9). On the time line, researchers can stream video files and visualize how video data, session metadata, and other files relate to each other temporally and thematically. Researchers can also use this interface to annotate an entire video file, or specific segments of video, with keywords and tags. The time line and tagging functionality further enrich the metadata to help other researchers to find and make sense of the video data contained in Databrary on a granular level.

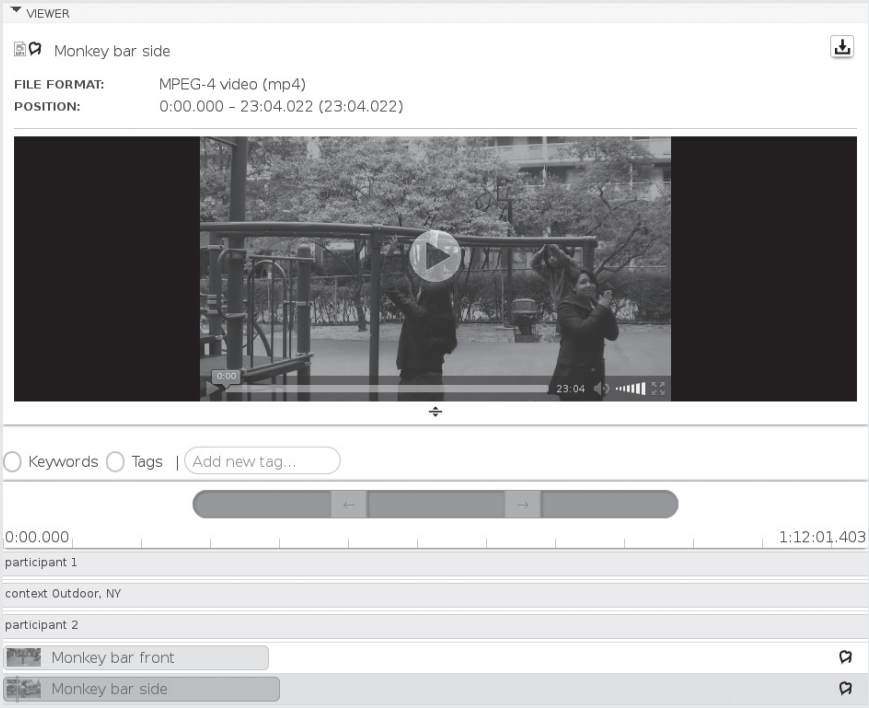

FIGURE 3.9
Databrary time line interface example. Time line for one of the sessions in a data set hosted on Databrary. (Source: Karen E. Adolph, "Social and Motor Play on a Playground," Databrary, 2014, accessed November 10, 2015, https://nyu.databrary.org/volume/9/slot/6113/-?asset=9607, doi:10.17910/B77P4V.) Users can access video assets in the browser, and data owners can manage their data using the time line interface. Video Image License: http://creativecommons.org/licenses/by-nc-sa/4.0.

POSSIBILITIES FOR ACTIVE CURATION AND PRESERVATION IN OTHER AREAS OF RESEARCH

In conceiving Databrary, we came to understand that valuable research data is susceptible to loss because it is not traditionally collected for the purposes of being preserved and shared with other researchers. We also knew that the task of preparing data toward these ends is cumbersome. Researchers need services, tools, and a centralized infrastructure to make preserving and sharing their research data a viable process. Our starting assumption that researchers are more motivated to use a service that is similar to their existing workflow allowed us to develop tools to facilitate and incentivize active curation so that researchers might prepare their own data for contribution to a repository that is a shared, community resource. Of course, child development is not the only discipline where video or audio data is central to the research workflow. Researchers in education, social anthropology, ethology, sociology, and linguistics also collect significant numbers of research video and audio files. Similar projects serving other fields may need to determine the metadata schema and interfaces that work best for their community. Other fields and institutions will also want to decide the scale at which they expect their repository solutions to function. Finally, they will also have to determine the ethical policies and protections (e.g., what to collect, what to share, with whom, and how) that best suit the nature of their research and the type of data they collect.

Different data types have different repository needs, but research video has specific needs and requirements for its curation and management. As a result, active curation may be possible in other fields of research with video or audio files at their center. The Databrary example demonstrates what is critical to this endeavor: Make curation similar to and seamless with researchers' current workflows; use language and interfaces that are based on researchers' current practices; and provide researchers with the infrastructure that allows them to organize, describe, manage, and store their data products so that they may be shared with their colleagues now and in the future. Most important, eliminate disincentives and build in incentives (e.g., centralized storage, collaborative tools, and curatorial assistance) for data sharing and reuse.

3.6 Consider the File Formats

In this step, identify all file formats in the data submission and note their restrictions. Curators may also want to verify the technical metadata of the files (e.g., resolution, audio/video codec) that may limit the files for various reuse purposes. When appropriate, curators may choose to transform the data files into open,