**Active Curation for Video Research Data**

# LOSING RESEARCH DATA DUE TO LACK OF CURATION AND PRESERVATION

Child development researchers have a problem: Their primary source of data is video, but they lack standard tools and practices across different labs for organizing, storing, and ensuring long term preservation, which can result in lost or corrupted data. Databrary has been funded, designed, and deployed to solve this problem. In this case study, we tell you how we did that. More generally, the processes and strategies employed here can be used to inform self-curation in other academic disciplines.

## Video Data in Child Development Research

For most child development researchers, video recordings of children’s behavior provide the backbone of their research programs. To collect these data, researchers record children in laboratory or natural settings and then analyze the videos for behaviors, themes, and events of interest. Video recordings represent a variety of research designs: the observations can focus on experimental manipulations of behavior or naturalistic, spontaneous expressions of behavior; recordings can span a few minutes or several hours; participants can be observed once or multiple times; observations can include a single child, multiple family members, peer groups, classrooms, or larger public groups; and the scientific questions are limited only by the researchers’ imaginations. By applying user-defined tags to portions of video, the recordings are transformed into quantitative data (e.g., frequency counts, rates, and durations) or qualitative data (e.g., ethnographic descriptions, conversation analyses, and narratives). Finally, the processed data are subsequently used for drawing inferences about child development.

Compared to other forms of research data (flat-file tabular data, imaging data, textual data), video has unique potential for reuse. One unique aspect of video is that it is largely self-documenting. Viewers can see what is happening and how the events were recorded. The raw data are more readily interpretable and carry more visible information about the original context of their original collection. A second unique aspect of video is that the recordings contain a wealth of information that is potentially usable by a broad audience. Multiple researchers can use the same raw research videos to explore entirely different phenomena. Therefore video data reuse may involve questions outside the scope of the original study.

Despite these advantages, video data sharing and reuse is not the norm in developmental science. Instead, most researchers collect videos for a single study and, after the study is complete, allow the data to molder away on a hard drive or set of tapes in a cabinet. Given the potential of video for reuse in research, the Databrary video research library (databrary.org) opened for general use in October 2014 to enable sharing and reuse of raw research video data among developmental scientists. Databrary is a web-accessible repository, with access permissions set differentially to authorized researchers and the public, depending on the dataset. The project is housed at New York University and is closely aligned with both the university’s Libraries and Information Technology Services. Databrary provides user-centered software for researchers to curate their video, metadate, and related documents in order to manage, store, and share these assets with their community.

## Post Hoc Curation and Active Curation

Post-hoc curation (i.e., after all of the data have been collected) is the most common way that researchers contribute data to domain repositories, and 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 (Giarlo, 2013). To prepare the data for deposit, researchers must revisit data that they have already collected, analyzed, and stored away and now annotate and describe them for the philanthropic 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, so information professionals were needed to process the collection and prepare it for sharing (Heidorn, 2011; Federer, 2013). To avoid these barriers to sharing, Databrary aimed to design tools that enable researchers to actively curate their own data during each phase of the research lifecycle, while continuing to support post-hoc curation when appropriate. Active or self curation in Databrary allows researchers to organize and manage their own data as they collect it, incentivizing researchers to store and preserve their data for sharing and reuse and relieving them of the burden of having to organize and describe their data once the study is alread complete.

# 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 good understanding of researchers’ workflows—in particular, the path from video data collection to storage. We started by interviewing of a handful of representative researchers and their staff at NYU and other institutions who regularly collect video data . Participants in these interviews were enlisted [from an existing network of researchers...?]. We hoped that the interviews would inform us about what researchers might want from a service that would help them to organize, manage, and eventually share their videos. However, the interview results were only minimally informative. Most researchers had not previously considered how to prepare their data for sharing and reuse—even for reuse within their own laboratories.

We realized that we needed to observe researchers’ current practices and the tools they used (or lacked) in their labs so that we could support active video curation for developmental science. In order to obtain an understanding of how researchers collect, organize, and analyze their videos and metadata we collected a sample of data from each researcher we had originally reached out to in order to determine similarities and differences among datasets. We learned that child development research is characterized by a wide diversity of practices both within and across labs. As a result, datasets are heterogeneously described and organized, which increases the time required to prepare these data for post-hoc ingesting into a repository.

# ACTIVE CURATION AND PRESERVATION IMPLEMENTED IN DATABRARY

## Databrary

Active curation techniques would address this problem, but first the researchers needed a standardized method for organizing and describing different types of datasets. To accomplish this goal, we designed Databrary from the beginnning to be a user-facing data repository that accommodates a diversity of existing data management practices. Into the system we built the basis for a standardized metadata schema that is comprised of categories that we found many researchers in the field to agree on as important to describing videos for finding and understanding (e.g., participant details, study conditions and tasks, geographic location, language of the participant, etc.). Our observations resulted in a system, described in more detail below, that uses a data model that reflects how researchers already view the different components of their video datasets and employs interfaces with which researchers are already familiar.

## Data Model

For active curation tools to make sense to our intended user base, we chose to adopt researchers’ language and organizing principles (Martinez-Uribe & Macdonald, 2009). Developmental researchers call the analytic units of their studies “sessions” (Bakeman & Quera, 2012). A session is essentially a recording period. Within each session, we assign the general term “record” to the metadata that describes a session. Records include information about participants, activities, and researcher-defined conditions and groups. The predefined records available in the upload interface were drawn from what we observed to be the most common pieces of metadata used across multiple labs.

## Interfaces

More important, we needed to craft interfaces that were easy to use and familiar to the community. Spreadsheets are a common tool employed across labs to record session metadata. As such, 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 1). Rows of the spreadsheet correspond to individual sessions. The columns correspond to basic categories of metadata describing that session, which helps researchers manage their own data and assists other researchers to search and find videos of interest to them. Column categories in the spreadsheet are customizable and can be applied as needed to the researcher's study.

<**FIGURE 1 HERE** – Gordon\_CaseStudy\_Fig1.png. Title: Databrary Metadata Interface Example

CAPTION: Spreadsheet metadata interface for a dataset hosted on Databrary (Fabricius, 2014). Databrary exposes as much metadata about a study as possible without sharing sensitive or identifiable information.>

In addition to allowing researchers to add and modify record metadata through this interface, we provided tools to enhance a researchers’ ability to generally analyze their data. Allowing users the ability to switch the view of the spreadsheet based on a category of interest gives them the power to easily gain new insights into their data at the click of a button (Figure 2). As such, active curation allows researchers to add their data to a repository as they work with it and additionally provides them with ways to understand their data better as they collect it.

**<FIGURE 2 HERE** – Gordon\_CaseStudy\_Fig2.png. Title: Databrary Task View Interface Example CAPTION: Spreadsheet metadata interface displaying the task view (Fabricius, 2014). Users can quickly switch between views to see a summary of how many sessions and which sessions contain a certain record.>

Finally, most developmental researchers use desktop video coding software [HOW TO CITE?] such as [Datavyu, Mangold, Noldus] to annotate videos for analysis. Thus, we implemented a timeline view for managing the videos and metadata within sessions that is similar to this commonly used desktop software (Figure 3). On the timeline, 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, with keywords and tags. The timeline and tagging functionality further enrich the metadata to help other researchers find and make sense of the video data contained in Databrary on a granular level.

**<FIGURE 3 HERE** – Gordon\_CaseStudy\_Fig3.png. Title: Databrary Timeline Interface Example CAPTION: Timeline for one of the sessions in a dataset hosted by Databrary (Adolph, 2014). Users can access video assets in the browser and data owners can manage their data using the timeline interface.>

# POSSIBILITIES FOR ACTIVE CURATION AND PRESERVATION IN OTHER AREAS OF RESEARCH

Databrary is still being actively developed. Now that it has been in use by its intended community for almost a year, we will conduct targeted usability testing in order to get a better understanding of what is working, what is missing, and what needs improvement. Though child development is not the only only discipline where video or audio is at the center of its data collection and analysis workflows. Researchers in Anthrolopology, Ethology, Sociology, and Linguistics collect large amounts such time series data, as well. [...so what...]

**Bibliography**

Adolph, K. (2014). Social and motor play on a playground. Databrary. Retrieved August 10, 2015 from http://dx.doi.org/10.17910/B77P4V.

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

Fabricius, W. (2014). Absence of Construct Validity in Standard False Belief Tasks. Databrary. Retrieved August 10, 2015 from http://dx.doi.org/10.17910/B7Z300.

Federer, L. (2013). The librarian as research informationist: a case study (English). Journal of the Medical Library Association, 101(4), 298-302.

Giarlo, M. J. (2013). Academic Libraries as Data Quality Hubs. Journal of Librarianship and Scholarly Communication 1(3):eP1059. doi: 10.7710/2162-3309.1059

Heidorn, P. B. (2011). The Emerging Role of Libraries in Data Curation and E-science. Journal of Library Administration, 51(7/8), 662-672. doi: 10.1080/01930826.2011.601269

Martinez-Uribe, L., & Macdonald, S. (2009) User Engagement in Research Data Curation. In Borbinha, J.L., Kapidakis, S., Papatheodorou, C., Tsakonas, G. (Eds.), Research and advanced technology for digital libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27 - October 2, 2009. Proceedings. (pp. 309-314) Berlin: Springer-Verlag. doi: 10.1007/978-3-642-04346-8