**Self Curation for Video Research Data**

Andrew S. Gordon, Lisa Steiger, and Karen E. Adolph – Databrary Project, New York University

**INTRODUCTION**

Video recordings of children’s behavior are an essential part of research in developmental science. 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; children 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). 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 interpretable and issues of data provenance can be mitigated. A second unique aspect of video is that the recordings contain a wealth of potentially useable information. Multiple researchers can use the same raw research videos to explore entirely different phenomena about children’s behavior. Video reuse can involve questions outside the scope of the original study.

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 completed, allow the data to molder away on a shelf or hard drive. Given the potential of video for research reuse, the Databrary (databrary.org) video research library was created to enable sharing and reuse of raw research video data among developmental scientists. Databrary is a web accessible library, with access provided differentially to authorized researchers and the public, depending on the dataset and the participants’ permissions. The library is housed at New York University and is closely aligned with the university’s Libraries and Information Technology Services. Databrary provides user-centered software and curation services for researchers to manage, store, and share their video data, metadata, and related research documents.

Because developmental science lacks a tradition of open data sharing, historically, most video data do not contain appropriate permissions from participants to be shared. Thus, a primary mission of Databrary is to expedite *future* contributions of research data, rather than to focus on previously collected data. Moreover, most researchers do not want to share their research videos until their study is completed and the paper is in press. Our aim, therefore, is to facilitate a rapid transition from private use to open sharing.

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 lack the expertise to prepare the data, so information professionals are 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, while continuing to support post-hoc curation when appropriate. *Active*, or self, curation allows researchers to organize and manage their own data as they collect it, ensuring that the dataset is ready for sharing at the end of their study with no added burden or cost.

**TOWARD ACTIVE CURATION**

Determining the best way to build active curation tools for the developmental science community required an understanding of researchers’ workflows—in particular, the path from video data collection to storage. We started by enlisting the help of a handful of representative researchers and their staff who regularly collect video data, and we asked them a prepared set of interview questions. We hoped that the interviews would inform us about what researchers would want from a service that helped them to organize, manage, and eventually share their videos. The interviews were only minimally informative. Most researchers had not 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 design the services and tools that would best suit active video curation for developmental science. We needed to obtain an understanding of how researchers collect, organize, and analyze their videos and metadata. We also collected a sample of each researcher’s data 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 addresses this problem by creating a standardized method for bringing in different types of datasets. We were able to fashion a user-facing repository that accommodates a diversity of existing data management practices. Our observations resulted in a system with 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.

**IMPLEMENTATION**

**Data Model**

For active curation tools to make sense to our intended user base, we had 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 any piece of 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 across multiple labs.

**Interfaces**

More important, we needed to craft interfaces that were transparent 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 to upload, modify, and manage session metadata that looks like 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 (e.g., participant details, study conditions and tasks, geographic location, language of the participant) describing that session, which helps researchers to manage their own data and assists other researchers to find videos of interest to them. Categories in the spreadsheet are customizable and can be applied as needed to the researcher's study.

[**FIGURE 1 HERE** - spreadsheet\_vol\_98.png. CAPTION: Spreadsheet metadata interface for a dataset hosted on Databrary (Fabricius, 2014). For transparency purposes, 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 metadata through this interface, we provided tools to enhance researchers’ ability to work with 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 in addition, provides them with ways to understand their data better as they collect it.

[**FIGURE 2 HERE** - task\_view\_vol\_98.png. 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.]

Most developmental researchers use desktop video coding software to annotate videos for analysis. Thus, we implemented a timeline view for managing the videos and metadata within sessions that is similar to the desktop software (Figure 3). On the timeline, researchers can stream video files and visualize how videos, session metadata, and other files relate to each other temporally and thematically. Researchers can also use this interface to annotate entire video files, or to annotate specific segments of video files with keywords and tags. The timeline and tagging functionality further enrich the metadata and helps other researchers to find and make sense of the video data contained in Databrary on a granular level.

[**FIGURE 3 HERE** - timeline\_vol\_9.png. 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.]

**NEXT STEPS**

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 user experience testing to get a better understanding of what is working, what is missing, and what needs improvement. At the same time, continuing to perform staff-guided curation on datasets when applicable gives us a better idea of the different needs of our user base and what features to add to active curation.

**REFERENCES**

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