ISSN 2161-3974

JeSLIB 2021; 10(3): e1208
https://doi.org/10.7191/jeslib.2021.1208

Journal of eScience Librarianship
putting the pieces together: theory and practice

Full-Length Paper

"(Hyper)active Data Curation: A Video Case Study from Behavioral Science

Kasey C. Soska[1], Melody Xu[1], Sandy L. Gonzalez[1], Orit Herzberg[1], Catherine S. Tamis-LeMonda[1], Rick O. Gilmore[2], and Karen E. Adolph[1]

[1] New York University, New York, NY, USA
[2] The Pennsylvania State University, University Park, PA, USA

Abstract

Video data are uniquely suited for research reuse and for documenting research methods and findings. However, curation of video data is a serious hurdle for researchers in the social and behavioral sciences, where behavioral video data are obtained session by session and data sharing is not the norm. To eliminate the onerous burden of post hoc curation at the time of publication (or later), we describe best practices in active data curation—where data are curated and uploaded immediately after each data collection to allow instantaneous sharing with one button press at any time. Indeed, we recommend that researchers adopt "hyperactive" data curation where they openly share every step of their research process. The necessary infrastructure and tools are provided by Databrary—a secure, web-based data library designed for active curation and sharing of personally identifiable video data and associated metadata. We provide a

Correspondence: Kasey C. Soska: kasey.soska@nyu.edu

Received: April 8, 2021 **Accepted**: June 4, 2021 **Published**: August 11, 2021

Data Availability: All of the methods are openly shared. Additional information at end of article.

Disclosures: The authors report no conflict of interest.

Abstract Continued

case study of hyperactive curation of video data from the Play and Learning Across a Year (PLAY) project, where dozens of researchers developed a common protocol to collect, annotate, and actively curate video data of infants and mothers during natural activity in their homes at research sites across North America. PLAY relies on scalable, standardized workflows to facilitate collaborative research, assure data quality, and prepare the corpus for sharing and reuse throughout the entire research process.

Curation and Data Sharing

Many social and behavioral scientists report that data sharing is required by their funders or journals (Gewin 2016; McKiernan et al. 2016). Yet, they also report that data sharing is onerous, time-consuming, difficult, costly, and unrewarded, and most researchers do not share (Alter & Vardigan 2015). The sticking point is data curation. Effective curation makes shared data findable, accessible, interoperable, and reusable—the so-called FAIR guidelines (Wilkinson et al. 2016). Thus, social and behavioral researchers must ensure that their data are uploaded in a software-agnostic format in an easily accessible repository, accompanied by the relevant metadata and clear guidelines for access, credit, and citation (Gordon, Steiger, & Adolph 2016; Vines et al. 2014). However, FAIR guidelines exceed the curation capacities of most researchers, because typical data management practices and sharing incentives make curation onerous (Krzton 2018).

Challenges of "Post-hoc" Curation

Data curation in most social and behavioral research—when it occurs at all—typically occurs when the manuscript is published, or even later when an outside researcher requests the data. Such "post-hoc" curation—conducted after the study ends—is fraught with problems. At the end of a study, cleaning, labeling, and collating data for sharing can seem a burdensome chore, rather than an integral part of the research process. Data sharing as an onerous last step in dissemination often leads to an "upload and dump" mentality. File labelling is inconsistent, and the data provenance is impoverished (for example, see meager labeling of "video" data at https://osf.io/search/?q=video&filter=file&page=1). Indeed, the lag (often years long) between the start of data collection and the publication of the manuscript makes provenance uncertain: Researchers' recall for naming conventions, inclusion/exclusion criteria, links among data elements, and details on administration of the protocol degrades over time. Moreover, participant permission to share data is difficult to obtain after data collection ends, because contact information becomes obsolete or participants are reticent to share data from a session they no longer remember (Gilmore, Adolph, & Millman 2016).

In addition, most behavioral scientists consider shared data as supplemental materials to accompany a published manuscript rather than treating the shared dataset as the principal publication that fostered the manuscript as an offshoot. The former view—data as supplemental—may limit which data researchers share. Manuscripts typically include only a subset of participants; others were pilots, did not meet inclusion criteria, did not complete all tasks or sessions, and so on. Some tasks or experimental conditions may have yielded null results or been replications and therefore excluded from the manuscript. However, these "extra" data could be valuable for other researchers to replicate a method (e.g., by seeing pilots and data collection sessions that did not work) or in secondary data reuse. Indeed, behaviors that might exclude participants from one task (e.g., infants fussing) may be exactly what another researcher wants to study (Gilmore & Adolph 2017).

Active Curation is the Solution to the Problems of Post-hoc Curation

Here we suggest new methods and technologies to relieve some of the burdens of data curation, improve data management, and thus make data sharing a welcome and integral part of behavioral research. Instead of post hoc data curation, we advocate for "active" (i.e., upload as you go) data curation, an approach endorsed by library scientists (Akmon, Hedstrom, Myers, Ovchinnikova, & Kouper 2018; Myers & Hedstrom 2014). Indeed, we take active curation a step farther. We suggest that best practices require "hyperactive" data curation, where each step of the research life cycle involves considerations for data sharing.

We offer a case study from behavioral science involving video data collection across sessions and associated annotation done at scale. Our story involves three key players: (1) video collected as sharable research data and video created as documentation for training purposes; (2) Databrary (www.databrary.org), a restricted-access video data library designed for active data curation; and (3) the Play & Learning Across a Year (PLAY) Project (www.play-project.org), a large-scale effort involving collaborative data collection and annotation among 70 research teams across North America. We describe how we planned from the outset to openly share personally identifiable and potentially sensitive video data with the research community by making hyperactive data curation the backbone of the research project.

Making Data Curation "Hyperactive"

Hyperactive curation expands on active curation in two ways. First, researchers should consider sharing everything. They should plan to share every protocol decision, all training materials, every data collection session (whether included in final analyses or not), all raw and processed data, all analytic scripts, and so on (Macleod, Collins, Graf, Kiermer, & Mellor 2021). Video documentation can supplement protocols to increase transparency and reproducibility (Adolph 2020). Second, researchers should consider sharing at every step. Figure 1 outlines the process of hyperactive curation—with steps applicable to any behavioral research project (left column) and the curation acts and technologies specific to the PLAY Project (right column). As each instance of study-wide materials is finalized (the protocol, annotation manual, questionnaires, etc.), it enters a curation pipeline (blue rows in Figure 1) whose endpoint is a recognized data repository. Likewise, as soon as each piece of data is collected, it is curated (green rows in Figure 1). Any data collected by human hands (or instruments involving calibration) require assurance of data quality and adherence to the protocol before processing and analyses; as quality assurance decisions are noted and data are processed, they are curated (yellow rows in Figure 1). Each time researchers annotate or augment the raw data—with qualitative descriptions or formal annotation to generate quantitative outcome measures—annotations linked to raw data are curated (orange rows in Figure 1). At the time of publication (red rows in Figure 1), sharing of excerpts, exemplars, and links to the curated dataset involves a mere button press.

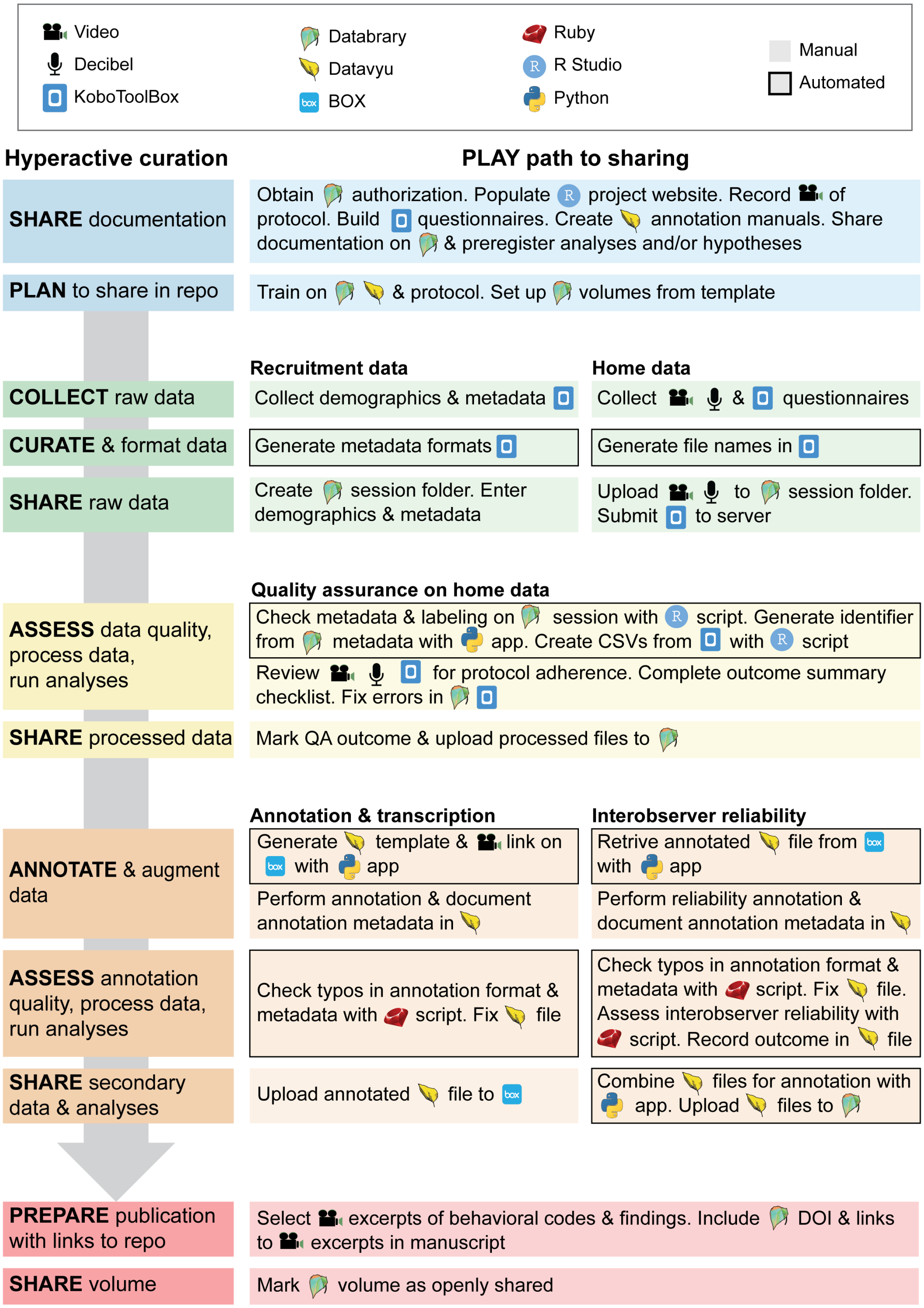

Video
Decibel
KoboToolBox
Databrary
Datavyu
BOX
Ruby
R Studio
Python
Manual
Automated
Hyperactive curation
PLAY path to sharing
SHARE documentation
Obtain authorization. Populate project website. Record of protocol. Build questionnaires. Create annotation manuals. Share documentation on & preregister analyses and/or hypotheses
PLAN to share in repo
Train on & protocol. Set up volumes from template
Recruitment data
Home data
COLLECT raw data
Collect demographics & metadata
Collect & questionnaires
CURATE & format data
Generate metadata formats
Generate file names in
SHARE raw data
Create session folder. Enter demographics & metadata
Upload to session folder. Submit to server
Quality assurance on home data
ASSESS data quality, process data, run analyses
Check metadata & labeling on session with script. Generate identifier from metadata with app. Create CSVs from with script
Review for protocol adherence. Complete outcome summary checklist. Fix errors in
SHARE processed data
Mark QA outcome & upload processed files to
Annotation & transcription
Interobserver reliability
ANNOTATE & augment data
Generate template & link on with app
Retrive annotated file from with app
Perform annotation & document annotation metadata in
Perform reliability annotation & document annotation metadata in
ASSESS annotation quality, process data, run analyses
Check typos in annotation format & metadata with script. Fix file
Check typos in annotation format & metadata with script. Fix file. Assess interobserver reliability with script. Record outcome in file
SHARE secondary data & analyses
Upload annotated file to
Combine files for annotation with app. Upload files to
PREPARE publication with links to repo
Select excerpts of behavioral codes & findings. Include DOI & links to excerpts in manuscript
SHARE volume
Mark volume as openly shared

Figure 1: Steps in hyperactive curation (left column) and specific curation actions in the PLAY project (right column). Icons denote technologies. Outlined boxes denote automated processes.

Deciding to Curate Hyperactively

Researchers in the social and behavioral sciences (and in related fields, such as biomedicine and education) typically curate data only for internal use or limited-scope sharing. They build training manuals, label and store raw and processed data, augment data through annotations, and conduct analyses prior to publication. The key difference with our hyperactive approach is considering a broader audience at every step to shrink the time gap to open sharing. That is, before any data collection commences, researchers should begin curating their materials—perhaps when piloting, applying for funding, or pre-registering hypotheses.

Our curation approach emerged as part of a large, multi-site research initiative, so the workflows are particularly amenable to other collaborative projects but can scale down to a single laboratory. The workflows can apply to any research field where a reproducible protocol is used to collect data across sessions, participants, or sites. And compared to existing curation practices, we have found these methods to accelerate the pace of data sharing, improve collected data quality, and reduce researcher burden.

Planning for Hyperactive Curation: The Five Ws

To plan for hyperactive curation, we offer as a heuristic five Ws (left column of Table 1). We illustrate the value of the heuristic using the PLAY Project as a case in point (right column of Table 1).

Why curate and share materials, methods, and primary research data for a particular project? Any sharing requires well-curated data, associated metadata, and study-wide materials—whether sharing only among the members of the research team, with known colleagues, or open sharing with the larger research community. However, knowing that you want to share is insufficient. An adequate curation plan must meet the goals of sharing. In the case of PLAY, the goals were to facilitate reuse for teaching and research, promote transparency and reproducibility, and ensure the quality of the primary data and the secondary data annotations.

What data should be curated is intimately tied to the goals for sharing. If the goal is to provide illustrative exemplars for teaching or training, then short video excerpts could be prepared with associated notes or voice-over. If the goal is to enable replication and increase transparency, then sharing should focus on methods—full session-long videos; study-wide materials; deidentified processed data; annotated, reproducible analysis scripts; and the provenance of each data file. If even a single participant agrees to share, their data can serve as a representative exemplar of the entire protocol. And if the intent is to enable widespread data reuse, then sharing should focus on the participant data—entailing participant permission from as many sessions as possible to share the data in the rawest form possible (original videos from data collection sessions,

Table 1: Answering the five Ws of active curation for the PLAY Project.

Why are data and associated materials being curated and shared?	• Sharing with collaborators across 50 institutions • Reuse of all videos, questionnaires, annotation spreadsheets, and other data by authorized Databrary researchers • Total transparency of methods, materials, and questionnaires for replication • Reproducibility of curation and workflow tools • Quality assurance on annotation schemes by linking annotations directly to videos
What data and associated materials to curate and share?	• All videos from every session • All questionnaire results and metadata • All transcription and behavioral annotation spreadsheets • All collection materials, surveys, and annotation manuals • Full exemplar session showing researcher performing protocol with participant family
Where to share data and associated materials?	• Databrary volume for study-wide collection materials • Databrary volumes for data from each session (allows sharing of potentially identifiable sessions because participant and researcher faces, voices, and homes will not be obscured) • Project website (standardized training materials given to researchers with details of all methods)
When to curate and share data and associated materials?	• Data collection and annotation protocols shared during collection process (publicly shared on Databrary and project website) • Video and associated data curated during collection process (to "private" Databrary volume) • Video and associated data shared after embargo for PLAY group members to prepare publications
Who will curate and share data and associated materials?	• PLAY staff curate protocol materials • Site researchers curate video data after collection • PLAY staff perform quality assurance on video data • PLAY staff curate associated annotation files • Project PIs share all data and oversee project

raw physiology data, questionnaire data for every item, etc.). Sessions that fail to meet quality assurance for the primary project can be marked as such and shared. If securing participant permission to share is impossible, when using video, the researcher can blur participant faces, remove audio, or censor instances of identifiable information (see https://nyu.databrary.org/volume/1116 and Ossmy & Adolph 2020). Motivations for sharing dictate the depth and breadth of materials to be curated. However, there is no harm in sharing more data than needed to meet the goals for sharing. Well-curated, complete datasets are valuable; fragmented pieces in a "data dump" are not. All of the PLAY data and methods are shared because the goals for sharing include teaching/training, research reuse, transparency, and reproducibility of high-quality data.

Where should data be shared? The answer depends on the data types. Databrary enables open sharing of personally identifiable video data for reuse, teaching/ training, and transparency. Repositories such as the Open Science Framework, Mendeley Data, and the Qualitative Data Repository are file-type agnostic but not designed for identifiable video. Repositories for specific research fields (e.g., TalkBank, National Database for Autism Research, OpenNeuro) may require adherence to data dictionaries with specific file types, organization structures, metadata, and labels. Knowing the types, permissions, and accepted structures for the intended repository means data can be actively curated in the appropriate format at the outset. The PLAY data are shared in Databrary because it is designed for sharing potentially identifiable research video.

When should data be curated and shared? With hyperactive data curation, the job starts before data collection begins, and sharing can occur at any time and need not be all or none. For example, full research protocols and sharing of questionnaires, annotation manuals, and other research materials can be shared at any time—even during piloting with protocol changes tracked via version control on a website or wiki. To prevent others from "scooping" findings, sharing timelines can vary—when filing a registered report, during review, when the article goes to press, after an embargo period, or later. Many repositories, such as Databrary, allow researchers to keep their datasets "private" or unlisted, or to share only an overview description or portions of data until the contributor opens up the entire dataset or particular files for wider sharing. PLAY shares study-wide materials at the outset; primary data and secondary annotations are shared after an embargo period.

Who will curate and share the data? The dataset owner (typically the principal investigator) is responsible for authorizing sharing in a repository. However, students and staff can serve as data stewards by curating primary research data throughout collection and annotation. Library partners can guide workflow and curation practices to offload planning by unseasoned data curators. Researchers can work with IT departments, computer scientists, and institutional support staff to develop custom apps, use repository APIs, or fork existing workflows for automated data transfer and quality assurance. Active curation can scale from one research team to multi-site projects where facets of curation are delegated to numerous researchers. PLAY data are curated by site researchers and central PLAY staff, and sharing decisions are handled by the PLAY project directors.

To build a reproducible workflow, the process should be automated whenever possible to avoid idiosyncrasy and reduce human error. Moreover, every research endeavor will face curation challenges that depend on the primary data types (behavioral, physiological, survey), whether data are collected en masse or across sessions, the data format required by the host repository, and protection of participant privacy and confidentiality.

Using and Curating Video in Behavioral Research

To illustrate the process of hyperactive curation, we use video data as a model system. Video recording is common among researchers in the social and behavioral sciences. Researchers use video as a primary source of data (to record participant behavior for later annotation or transcription) or as backup for live data collection. Researchers also use video to document research procedures (including computer-based tasks), train research staff, or illustrate research findings for teaching or presentations. Video—as both data and documentation—facilitates scientific transparency and speeds progress when openly shared and reused.

Video as Research Data

Since the advent of film, behavioral scientists used cinematic recordings (now digital video) to capture participant, group, or animal activity for later annotation, transcription, and analyses (Adolph 2020; Gesell 1935). Video annotation can entail summary scores or ratings of an entire session or experimental condition, notations for specified time epochs (e.g., every 10 seconds), or frame-by-frame micro-codes that identify the onset/offset of target behaviors and categorize each event. Transcripts of participant speech include verbatim serial records of who said what, speech utterances (bounded by pauses or conceptual clauses), or utterances time-locked to their location in the video. Typically, a subset of the video data is independently annotated by another researcher as quality assurance of the annotation scheme and implementation (termed "inter-observer reliability"). And finally, the annotations and transcripts are processed into flat-file spreadsheets for statistical and graphical analyses.

Compared to other data formats (questionnaires, physiological data, etc.), video is so rich in detail about behavior and the surrounding context that it is uniquely suited for research reuse (Adolph 2016; Gilmore & Adolph 2017; Gilmore, Kennedy, & Adolph 2018; Suls 2013). The same research team can reuse their own video data to ask new questions (e.g., Adolph et al. 2012; Karasik, Adolph, Tamis-LeMonda, & Zuckerman 2012; Karasik, Tamis-LeMonda, & Adolph 2011; Karasik, Tamis-LeMonda, & Adolph 2014) and new research teams can conduct secondary research—often asking questions in domains never considered by the original data contributors (e.g., Friedman et al. 2014; Gilmore, Raudies, & Jayaraman 2015; Messinger et al. 2017). Researchers can reuse the entire dataset, portions of the parent dataset, or combine multiple datasets to address their questions (e.g., Han & Adolph 2021; Soska, Robinson, & Adolph 2015). Thus, video data curation (including the video files, annotation spreadsheets, participant information, etc.) lays the foundation for sharing and reuse among a single research team or with the wider research community.

Video as Documentation for Transparency, Training, and Teaching

Written descriptions of methods and results are the bread and butter of behavioral science. But compared to text and still images, video captures more nuance and details of who did what and how and where they did it (Adolph 2016, 2020; Suls 2013). Thus, video documentation of procedures, testing displays, and findings offers greater transparency and thereby greater support for reproducibility than do text and still images (Gilmore & Adolph 2017).

Video is a tremendously useful tool for training and instruction. Indeed, textbooks offer video collections of staged “mock-ups” of classic studies; standardized assessments provide video examples to train new administrators; and researchers retain a bank of lab videos for internal training. Although staged mock-ups can be informative, real research data are the most powerful teaching tools; commercial video demos are typically proprietary with restricted access; and lab training videos are often unsuitable or unavailable for broader use.

Well-curated, findable video documentation in an accessible repository will increase the reach of the research (Gilmore, Adolph, Millman, & Gordon 2016). Video excerpts are a fast, efficient way to demonstrate aspects of the data collection protocol (e.g., instructions given to participants), illustrate operational definitions of behavioral annotations (e.g., what behaviors count as a “object interaction”) or transcriptions (e.g., which infant vocalizations are “babbles” and which are “words”), and to highlight research findings and exceptions. Full, unedited videos of actual participants reveal the entire data collection process, including parts the original researcher may take for granted, but that are critical for reproducibility (e.g., how to position infants for remote eye-tracking).

Special Challenges in Curating and Sharing Video Data: Databrary

Video data pose unique challenges for curation and sharing. File sizes are large and file formats become quickly outdated. Ensuring participant privacy is challenging because video typically contains personally identifiable information—faces are visible, voices are audible, and the interiors of people’s homes may be revealed. Often vulnerable populations are involved (e.g., children, people with disabilities, people with concerns about immigration status). The Databrary library addresses these concerns. Databrary provides unlimited storage, automatically transcodes videos into standard, preservable formats (while retaining the original files), and limits access to authorized investigators (Gilmore et al. 2018).

Databrary's ethical policy framework relies on an access agreement signed by researchers and their home institutions. As authorized investigators, researchers can function as data contributors, data reusers, or both. The access agreement protects the privacy of participants' data and protects the rights of the authorized investigators (Gilmore et al. 2018). Data contributors can upload files with different permission levels (https://databrary.org/support/irb/release-levels.html)—shared only with the original research team, shared openly with authorized investigators, usable in teaching and with non-authorized learning audiences, or publicly accessible. For sessions permissioned only for sharing with the original researchers, videos and identifiable data (birthdate, etc.) can be marked "private," while anonymized data such as demographics and questionnaire responses can be shared with authorized researchers.

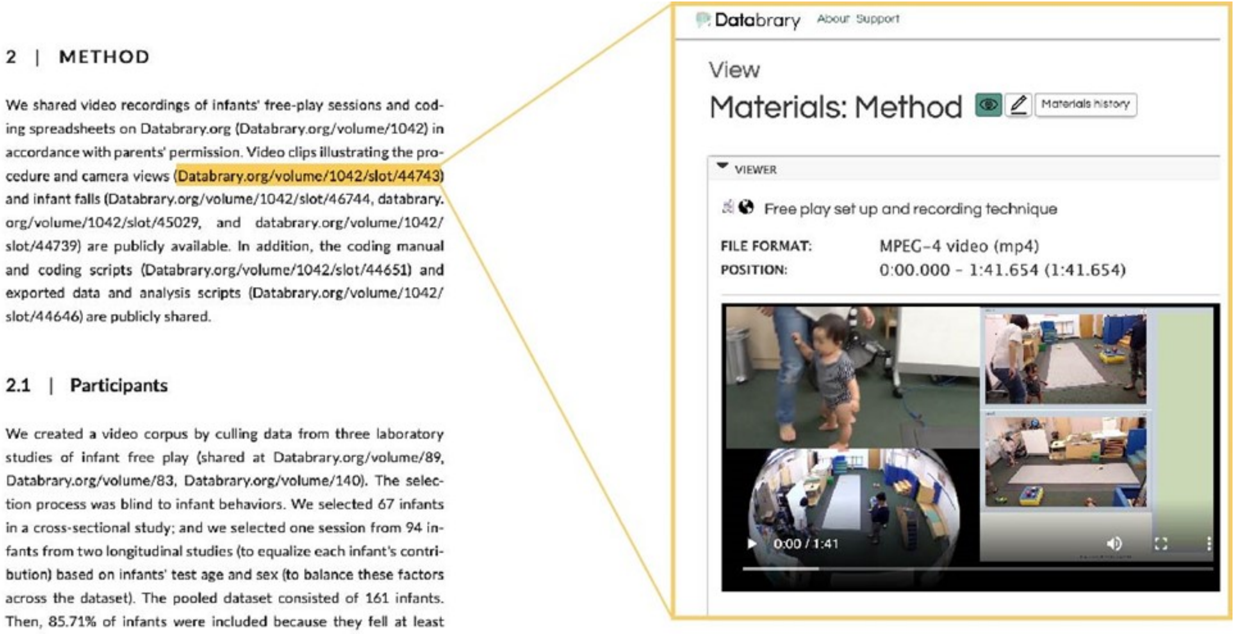
Metadata are curated in spreadsheet form through structured fields (e.g., birthdate, sex, spoken language, disability status) and freeform groups (e.g., tasks, conditions, exclusion/inclusion criteria). Primary data from each data collection (videos and other associated data from annotation, survey responses, physiological recordings, etc.) are stored in "session" folders. Documentation files that apply to an entire dataset (protocols, videos of methods, blank surveys, annotation manuals, etc.) are shared in a "materials" folder.

Sharing Curated Video Data in Publications

With hyperactive curation, preparing the video dataset to be linked with a publication or set of publications is straightforward. Databrary automatically generates a DOI for the dataset at its creation to make the dataset findable on its own. The DOI should be included in publications and credited whenever the dataset is referenced or reused. To make a dataset findable from the publication, we suggest streamlining the typical pathway of electronic links—from manuscript to supplemental materials on a publisher's website behind a paywall, to the dataset repository, to the actual videos or other data files. Instead, we encourage researchers to use live links in the publication that take readers directly to the dataset, materials, or individual videos (Figure 2). Another option is to replace traditional image figures in the article with video "figures" (see Adolph 2020) or link out to a project website (as we did in this article). In addition to sharing video data and excerpts, researchers should embed live links to analysis files, annotation manuals, automation scripts, and curation tools for others to use. Moreover, the hard work of curation and sharing can be recognized by listing shared datasets on the researcher's curriculum vitae and grant applications.

Case Study: "Hyperactive" Curation in the PLAY Project

PLAY leverages the power of video for documentation and reuse in a hyperactive curation workflow. Each step in the curation and sharing process from PLAY project planning through final publications and sharing are illustrated in Figure 1 (right column). PLAY will produce a first-of-its-kind openly shared corpus of hour-long videos of 1000+ infants and mothers during natural activity in the home with full



2 | METHOD

We shared video recordings of infants' free-play sessions and coding spreadsheets on Databrary.org (Databrary.org/volume/1042) in accordance with parents' permission. Video clips illustrating the procedure and camera views (Databrary.org/volume/1042/slot/44743) and infant falls (Databrary.org/volume/1042/slot/46744, databrary.org/volume/1042/slot/45029, and databrary.org/volume/1042/slot/44739) are publicly available. In addition, the coding manual and coding scripts (Databrary.org/volume/1042/slot/44651) and exported data and analysis scripts (Databrary.org/volume/1042/slot/44646) are publicly shared.

2.1 | Participants

We created a video corpus by culling data from three laboratory studies of infant free play (shared at Databrary.org/volume/89, Databrary.org/volume/83, Databrary.org/volume/140). The selection process was blind to infant behaviors. We selected 67 infants in a cross-sectional study; and we selected one session from 94 infants from two longitudinal studies (to equalize each infant's contribution) based on infants' test age and sex (to balance these factors across the dataset). The pooled dataset consisted of 161 infants. Then, 85.71% of infants were included because they fell at least

View

Materials: Method

Free play set up and recording technique

FILE FORMAT: MPEG-4 video (mp4)
POSITION: 0:00.000 - 1:41.654 (1:41.654)

Figure 2: Example of a live link in a publication (left side) that points to the location on Databrary of a component of the shared video dataset (right side). Adapted from Han, D. & Adolph, K.E. 2021. "The impact of errors in infant development: Falling like a baby." *Developmental Science* January. https://doi.org/10.1111/desc.13069

speech transcriptions and behavioral annotation. Additional data include parent-report questionnaires across a range of domains in infant development, detailed video tours of each home, decibel recordings of ambient noise, and detailed demographics. Given the size and scope of the project, curation was planned and integrated into the workflow from the outset to ensure quality assurance, distribute videos to annotation sites, make the curated data available to the 70 PLAY labs, and later openly share the corpus on Databrary. PLAY also aims to demonstrate the feasibility of total transparency across the entire data collection and annotation process.

Curation Planning for Piloting and Grant Application

In webinars and conferences, PLAY researchers jointly determined the data collection protocol (including technical and procedural specifications for the videos), data types, criteria for inclusion based on adherence to the protocol (percent allowable off-camera segments, missing questionnaire data, etc.), detailed video behavioral annotation and transcription schemes, and inter-observer reliability guidelines for annotation. The process is documented in shared videos (https://nyu.databrary.org/volume/254). Before piloting began, each PLAY investigator obtained a Databrary institutional agreement to become an authorized investigator.

Curating Methods for Training, Transparency, and Reproducibility

All aspects of the PLAY protocol are openly shared (blue rows in Figure 1). We built a public website in R Markdown (www.play-project.org) to document recruitment, collection, annotation, and participant demographics by site. Training of each site is conducted virtually, with recordings of the training shared back to the researchers. The data collection protocol (https://www.play-project.org/collection.html) contains text descriptions with accompanying videos for each aspect of recruitment and data collection in both English and Spanish. We publicly shared full videos of a typical data collection showing what the researcher records (https://nyu.databrary.org/volume/876/slot/55651/-?asset=337405). We also publicly shared full videos showing what the researcher does during recruitment and data collection to illustrate how to administer the protocol from start to finish (https://nyu.databrary.org/volume/876/slot/55651/-?asset=337402). We produced a video of the steps to curate videos on Databrary (https://nyu.databrary.org/volume/876/slot/35422/-?asset=312164). Annotation definitions, exceptions, and exemplar behaviors are documented in digital annotation manuals (https://www.play-project.org/coding.html).

Curation Workflow and Tools

PLAY distributes the responsibility of collection, annotation, and curation across the participating labs to minimize the burden on individual labs. Figure 3 illustrates project oversight and pathways from training, through data collection, annotation, quality assurance, and sharing. Data collection is spread over 30 research sites across the United States (blue boxes in Figure 3). Video annotation of infant and mother behaviors is dispersed across 48 labs (orange boxes in Figure 3), each with expertise in a particular domain—locomotion, object interaction, emotion, and communication and gesture. To minimize human error and direct who handles each file when, we streamlined the data handling process with the central PLAY team overseeing the entire workflow (green boxes in Figure 3). Because Databrary does not assess data curation or quality upon upload, the PLAY team assured that final upload was consistent and organized.

The curation workflow and tools are open source and can scale across as many sites as needed (https://github.com/PLAY-behaviorome) and are highlighted and expounded upon throughout Figure 1. Digital parent-report questionnaires (e.g., measures of infant vocabulary, locomotor milestones, infant temperament, mother and infant health, home environment) were built in KoBoToolbox (www.kobotoolbox.org), an open-source, web-based toolkit that provides automatic upload to a central server. To distribute and track progress on annotation across sites, we used Box because of its encryption capabilities, support for storing sensitive data, and ability to automate file transfer from the central PLAY team to and from data annotation sites using APIs. We built a custom Python app to pull metadata from Databrary (https://github.com/PLAY-behaviorome/databraryapi) and push files to Box. Data annotation sites use Datavyu (www.datavyu.org), an open-source video annotation software

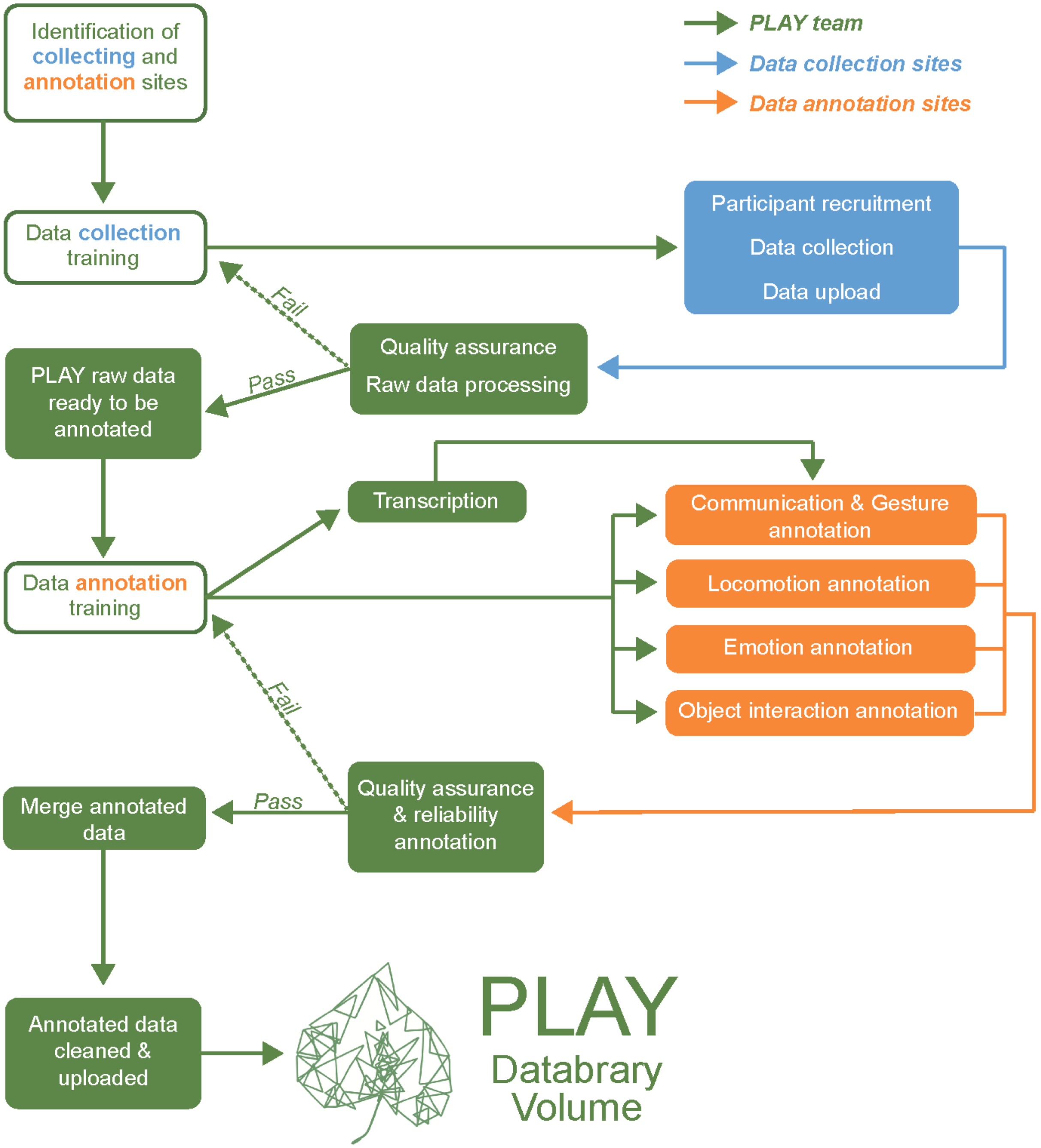
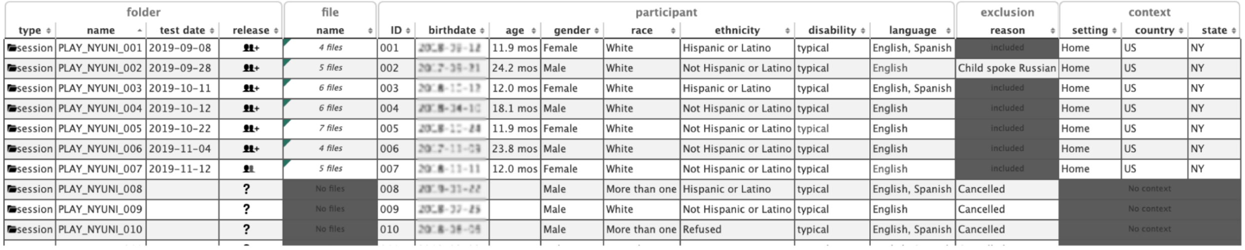


Figure 3: Diagram of the PLAY data collection, annotation, and curation workflow. Steps in green denote tasks performed by the central PLAY staff, including training, quality assurance, transcription of videos, and final sharing of the Databrary volume. Processes in blue are done by researchers at each of the 30 data collection sites, including uploading and curating video data. Processes in orange are completed by the 48 data annotation sites, who use curated videos to complete behavioral annotation of communication and gesture, locomotion, emotion, and object interaction by infants and mothers and return annotation spreadsheets back to the central PLAY team for quality assurance and sharing.

maintained by Databrary. We used a common annotation tool to standardize file formats, annotation columns, and code names, and to support a common set of Ruby scripts to generate annotation spreadsheets, automate checks of data entry errors, and assess inter-observer reliability. Thus, files associated with data collection videos can be actively curated in a consistent manner.

Active Curation for Collection and Quality Assurance of Video Data

A member of the central PLAY team worked with each data collection investigator to create a volume on Databrary, where their video data would be curated, stored, and shared. We developed a common spreadsheet template to ensure consistency in collection of demographics and metadata (Figure 4).

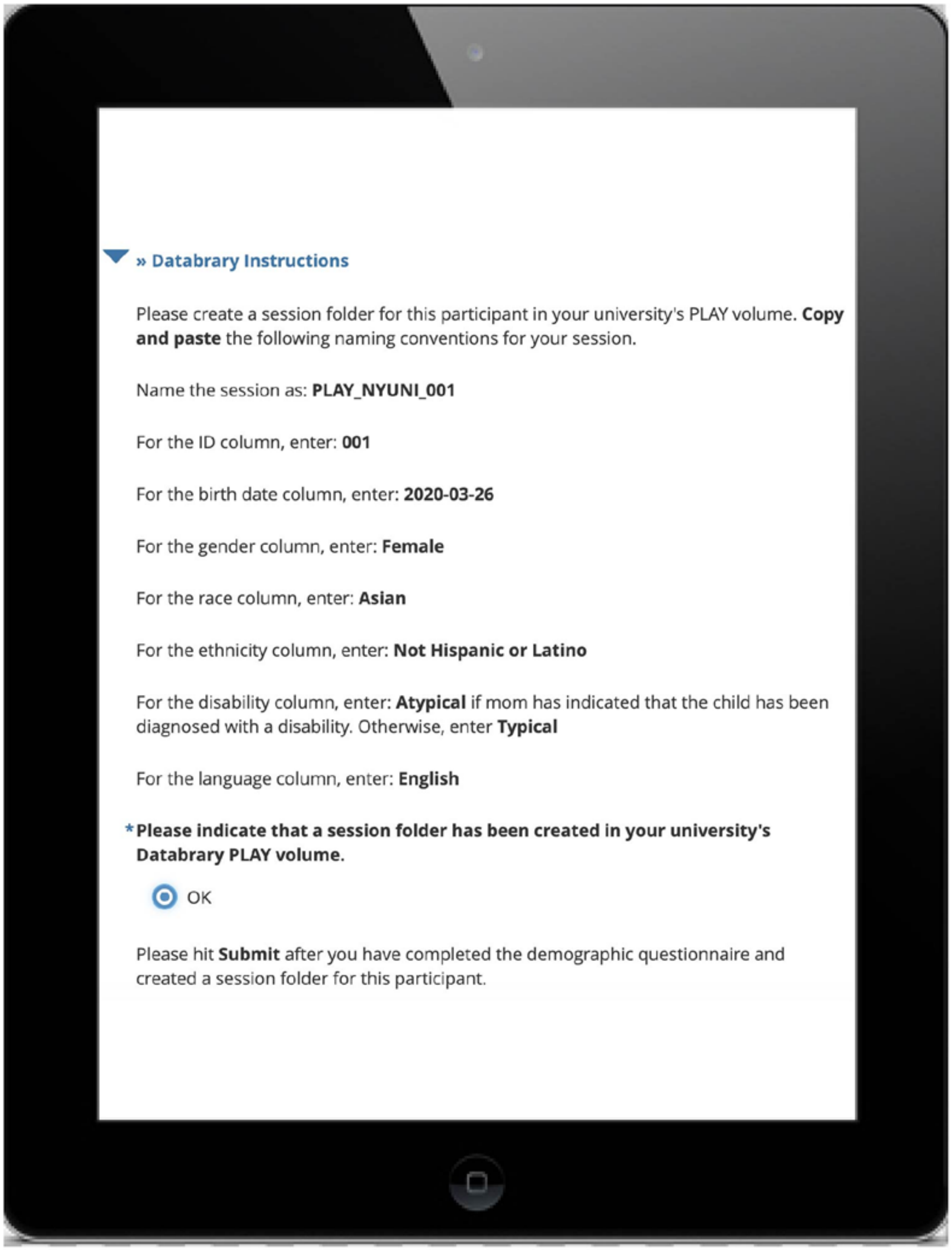


folder				file	participant								exclusion	context		
type	name	test date	release	name	ID	birthdate	age	gender	race	ethnicity	disability	language	reason	setting	country	state
session	PLAY_NYUNI_001	2019-09-08	👥+	4 files	001	[redacted]	11.9 mos	Female	White	Hispanic or Latino	typical	English, Spanish	included	Home	US	NY
session	PLAY_NYUNI_002	2019-09-28	👥+	5 files	002	[redacted]	24.2 mos	Male	White	Not Hispanic or Latino	typical	English	Child spoke Russian	Home	US	NY
session	PLAY_NYUNI_003	2019-10-11	👥+	6 files	003	[redacted]	12.0 mos	Female	White	Hispanic or Latino	typical	English, Spanish	included	Home	US	NY
session	PLAY_NYUNI_004	2019-10-12	👥+	6 files	004	[redacted]	18.1 mos	Male	White	Not Hispanic or Latino	typical	English	included	Home	US	NY
session	PLAY_NYUNI_005	2019-10-22	👥+	7 files	005	[redacted]	11.9 mos	Female	White	Not Hispanic or Latino	typical	English	included	Home	US	NY
session	PLAY_NYUNI_006	2019-11-04	👥+	4 files	006	[redacted]	23.8 mos	Male	White	Not Hispanic or Latino	typical	English	included	Home	US	NY
session	PLAY_NYUNI_007	2019-11-12	👥	5 files	007	[redacted]	12.0 mos	Female	White	Not Hispanic or Latino	typical	English	included	Home	US	NY
session	PLAY_NYUNI_008		?	No files	008	[redacted]		Male	More than one	Hispanic or Latino	typical	English, Spanish	Cancelled	No context		
session	PLAY_NYUNI_009		?	No files	009	[redacted]		Male	White	Not Hispanic or Latino	typical	English	Cancelled	No context		
session	PLAY_NYUNI_010		?	No files	010	[redacted]		Male	More than one	Refused	typical	English, Spanish	Cancelled	No context		

Figure 4: Portion of a template spreadsheet in Databrary used in PLAY for active curation. Each row is one participant session and is created by researchers after participant enrollment. Columns store demographics and metadata (note: birthdates are blurred in this figure and hidden in Databrary except when accessed by authorized researchers). Files are uploaded to each session folder immediately after the data collection.

We relied on researchers at each data collection site to act as data stewards because they have direct contact with their participants and are the first to handle the video files (green rows in Figure 1). After enrollment, researchers collect participant demographics in KoBoToolbox; a folder is created for each participant on Databrary, and demographic information and metadata are generated for entry into the Databrary spreadsheet (see Figure 5). A unique identifier—that follows that session through to final sharing—is automatically created using details from each Databrary session folder.

Participants are informed during recruitment that a primary goal of the research is to share videos of the data collection session with other researchers. If participants are not comfortable with sharing, we do not enroll them. We found that >90% of families are compliant—with no consistent demographic differences between those who agree to share and those who do not. We decouple consent to participate (done before the session starts) from obtaining permission to share the videos (done at the end of the data collection session) so that parents are better informed about what the video will contain after the session than when they consent to participate. After the researcher returns to the lab, videos are uploaded to Databrary. To minimize manual data entry and associated errors, experimenters



» Databrary Instructions

Please create a session folder for this participant in your university's PLAY volume. **Copy and paste** the following naming conventions for your session.

Name the session as: **PLAY_NYUNI_001**

For the ID column, enter: **001**

For the birth date column, enter: **2020-03-26**

For the gender column, enter: **Female**

For the race column, enter: **Asian**

For the ethnicity column, enter: **Not Hispanic or Latino**

For the disability column, enter: **Atypical** if mom has indicated that the child has been diagnosed with a disability. Otherwise, enter **Typical**

For the language column, enter: **English**

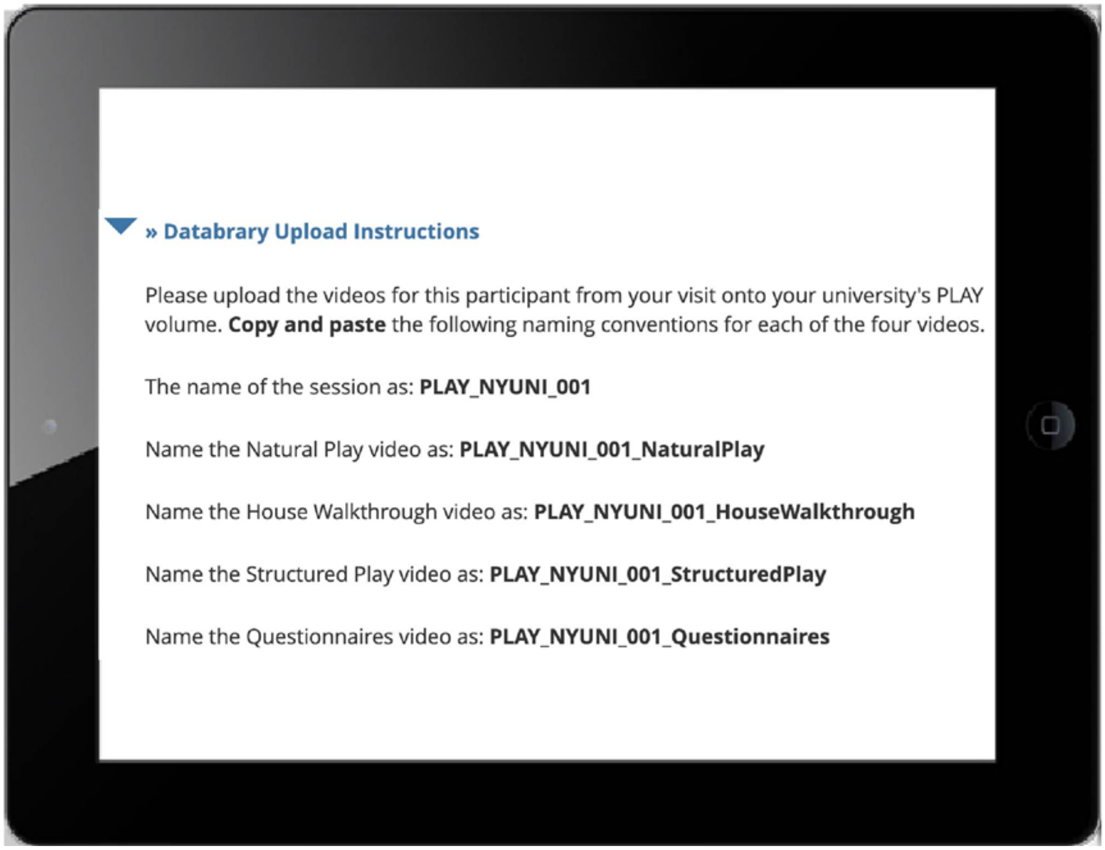
* **Please indicate that a session folder has been created in your university's Databrary PLAY volume.**

● OK

Please hit **Submit** after you have completed the demographic questionnaire and created a session folder for this participant.

Figure 5: Instructions in the KoBoToolbox tablet app used in PLAY to aid data collectors in creating sessions in Databrary, marking demographics, and relaying the information back to the central PLAY team when a new participant family is enrolled.

receive automated prompts that they then copy/paste to label video files consistently (Figure 6).



» Databrary Upload Instructions

Please upload the videos for this participant from your visit onto your university's PLAY volume. **Copy and paste** the following naming conventions for each of the four videos.

The name of the session as: **PLAY_NYUNI_001**

Name the Natural Play video as: **PLAY_NYUNI_001_NaturalPlay**

Name the House Walkthrough video as: **PLAY_NYUNI_001_HouseWalkthrough**

Name the Structured Play video as: **PLAY_NYUNI_001_StructuredPlay**

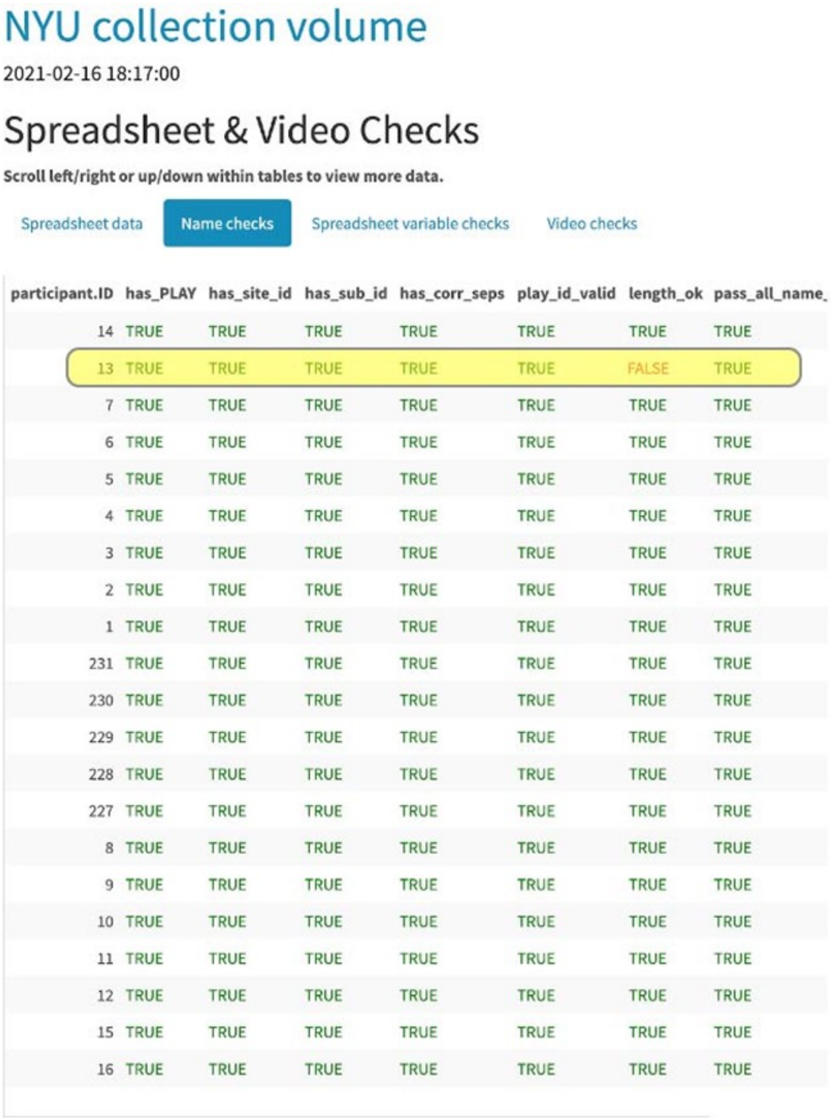
Name the Questionnaires video as: **PLAY_NYUNI_001_Questionnaires**

Figure 6: Instructions in the KoBoToolbox tablet app used by PLAY to aid data collectors in labelling video data to be uploaded to Databrary after a data collection session is completed.

A member of the PLAY team performs quality assurance on each session prior to ingesting it for annotation and transcription (yellow rows in Figure 1). During this process, we use a set of R scripts (https://github.com/PLAY-behaviorome/workflow) to check for data structure issues such as missing videos, incorrect labels, and so on (see Figure 7). We contact the data collection site to fix any issues before ingest is completed. The outcome of the quality assurance is indicated on each Databrary session folder.

Active Curation for Non-Video Data and Annotating Video Data

For video data from PLAY to be fully reusable and shareable, we also curate related questionnaires and annotations (transcriptions and behavioral codes). For all files regardless of whether they pass quality assurance, an R script parses the raw KoBoToolbox questionnaires to facilitate upload to the appropriate session



NYU collection volume

2021-02-16 18:17:00

Spreadsheet & Video Checks

Scroll left/right or up/down within tables to view more data.

participant.ID	has_PLAY	has_site_id	has_sub_id	has_corr_seps	play_id_valid	length_ok	pass_all_name_
14	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
13	TRUE	TRUE	TRUE	TRUE	TRUE	FALSE	TRUE
7	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
6	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
5	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
4	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
3	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
2	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
1	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
231	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
230	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
229	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
228	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
227	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
8	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
9	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
10	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
11	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
12	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
15	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE
16	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE	TRUE

Figure 7: Output of the quality assurance R script run on one exemplar data collection volume in Databrary for PLAY. The displayed section—Name Checks—evaluates each session (notated by participant ID) and looks for compliance with a set of naming conventions, file name lengths, and inclusion of each relevant identifier. Highlighting reflects that participant 13 contains an error in the length of a filename (flagged "False").

folder in Databrary. If the session passes quality assurance, a Python app generates a link to the video on Databrary for each site annotating a target behavior (locomotion, object interaction, emotion, or communication and gesture) along with the associated Datavyu annotation template (orange rows of Figure 1). In the Datavyu files, we document the provenance of who transcribed or annotated a behavior, and when. After annotating a video file, the sites transfer the spreadsheet back to the PLAY team who independently annotate 25% of each session to ensure inter-observer reliability. After the annotations are deemed reliable, the PLAY team merges all transcriptions and annotations into a single file that is stored in the participant's Databrary session folder.

Conclusions

The hyperactive curation workflow we showcase from PLAY can accommodate non-video data, larger or smaller research projects, and repositories other than Databrary. However, a limitation of hyperactive curation is that it entails expanding the scope of research staff attention, skill, and effort beyond collection, annotation, and analysis and into curation and sharing. To defray curation costs, researchers should include funding for curation in grant applications and budget personnel time for curation efforts from the start of the project. Moreover, hyperactive curation of the research project is tacitly assumed from the outset. If researchers begin curation after data collection or annotation begin, the workflows we shared become more difficult and costly to implement. Instead, we propose a shift in curation workflows and in mindsets: start planning for curation and sharing from the get-go.

Hyperactive curation spreads the cost of curation over a longer period and eliminates the headaches and inaccuracies of post hoc curation—making research easier, more streamlined, and less prone to errors. Indeed, we found these workflows improve the quality of collected data by vetting sessions for completeness, protocol adherence, and data formatting as soon as possible. By documenting all aspects of the protocol in shareable formats (website, videos on Databrary, digital manuals), training researchers and new staff is more efficient. Data are easier to organize, find, annotate, and analyze for internal use and data are immediately curated for open sharing. Hyperactive curation, in consultation with library partners, mitigates roadblocks in the curation process and maximizes transparency, reproducibility, and data reuse. In doing, hyperactive curation can create a culture among behavioral scientists where data sharing is standard practice.

Acknowledgements

KCS, MX, OH, and CTL were supported by NICHD R01HD094830, KEA was supported in part by DARPA N66001-19-2-4035, ROG was supported in part by NSF OAS2032713, and SLG was supported in part by NIMH T32-MH019524-28.

Supplemental Content

PDF: Figure 1 and Figure 3
An online supplement to this article can be found at http://dx.doi.org/10.7191/jeslib.2021.1208 under “Additional Files”.

Data Availability

All of the methods are openly shared. All protocols for data collection, coding, and curation are on the PLAY Project website (http://play-project.org). Videos are shared via links to the appropriate volumes on Databrary.org (Planning: https://nyu.databrary.org/volume/254; Implemented Protocol: https://nyu.databrary.org/volume/876). Workflow tools are shared with links to repositories on Github.com (https://github.com/PLAY-behaviorome).

References

Adolph, Karen. E. 2016. “Video as data: From transient behavior to tangible recording.” *APS Observer* 29: 23–25. http://www.apa.org/science/about/psa/2017/10/video-data

———. 2020. "Oh, behave!" Infancy 25. https://doi.org/10.1111/infa.12336

Adolph, Karen E., Whitney G. Cole, Meghana Komati, Jessie S. Garciaguirre, Daryaneh Badaly, Jesse M. Lingeman, Gladys L. Y. Chan, and Rachel B. Sotsky. 2012. “How Do You Learn to Walk? Thousands of Steps and Dozens of Falls per Day.” *Psychological Science* 23(11): 1387–1394. https://doi.org/10.1177/0956797612446346

Akmon, Dharma, Margaret Hedstrom, James D. Myers, Anna Ovchinnikova, and Inna Kouper. 2018. “Building Tools to Support Active Curation: Lessons Learned from SEAD.” *International Journal of Digital Curation* 12(2): 76–85. https://doi.org/10.2218/ijdc.v12i2.552

Alter, George C., and Mary Vardigan. 2015. “Addressing Global Data Sharing Challenges.” *Journal of Empirical Research on Human Research Ethics* 10(3): 317–323. https://doi.org/10.1177/1556264615591561

Friedman, Sarah L., Ellin K. Scholnick, Randall H. Bender, Nathan Vandergrift, Susan Spieker, Kathy Hirsh Pasek, Daniel P. Keating, and Yoonjung Park. 2014. “Planning in Middle Childhood: Early Predictors and Later Outcomes.” *Child Development* 85(4): 1446–1460. https://doi.org/10.1111/cdev.12221

Gesell, Arnold. 1991. “Cinemanalysis: A Method of Behavior Study.” *The Journal of Genetic Psychology* 152(4): 549–562. https://doi.org/10.1080/00221325.1991.9914712

Gewin, Virginia. 2016. “Data Sharing: An Open Mind on Open Data.” *Nature* 529(7584): 117–119. https://doi.org/10.1038/nj7584-117a

Gilmore, Rick O., and Karen E. Adolph. 2017. “Video Can Make Behavioural Science More Reproducible.” *Nature Human Behaviour* 1(7): s41562–017. https://doi.org/10.1038/s41562-017-0128

Gilmore, Rick O., Karen E. Adolph, and David S. Millman. 2016. “Curating identifiable data for sharing: The Databrary project.” In *2016 New York Scientific Data Summit (NYSDS).* IEEE. https://doi.org/10.1109/NYSDS.2016.7747817

Gilmore, Rick O., Karen E. Adolph, David S. Millman, and Andrew S. Gordon. 2016. “Transforming education research through open video data sharing.” *Advances in Engineering Education* 5: 1–17. http://advances.asee.org/wp-content/uploads/vol05/issue02/Papers/AEE-18-Gilmore.pdf

Gilmore, Rick O., Joy Lorenzo Kennedy, and Karen E. Adolph. 2018. “Practical Solutions for Sharing Data and Materials From Psychological Research.” *Advances in Methods and Practices in Psychological Science* 1(1): 121–130. https://doi.org/10.1177/2515245917746500

Gilmore, Rick O., Florian Raudies, and Swapnaa Jayaraman. 2015. “What Accounts for Developmental Shifts in Optic Flow Sensitivity?” In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE. https://doi.org/10.1109/devlrn.2015.7345450

Gordon, Andrew S., Lisa Steiger, and Karen E. Adolph. 2016. “Losing research data due to lack of curation and preservation.” In *Curating research data: A handbook of current practice*, edited by L. Johnston, 108-115. Chicago, IL: Association of College and Research Libraries.

Han, Danyang, and Karen E. Adolph. 2021. “The Impact of Errors in Infant Development: Falling like a Baby.” *Developmental Science* January: e13069. https://doi.org/10.1111/desc.13069

Karasik, Lana B., Karen E. Adolph, Catherine S. Tamis-LeMonda, and Alyssa L. Zuckerman. 2012. “Carry on: Spontaneous Object Carrying in 13-Month-Old Crawling and Walking Infants.” *Developmental Psychology* 48(2): 389–397. https://doi.org/10.1037/a0026040

Karasik, Lana B., Catherine S. Tamis-LeMonda, and Karen E. Adolph. 2011. “Transition From Crawling to Walking and Infants’ Actions With Objects and People.” *Child Development* 82(4): 1199–1209. https://doi.org/10.1111/j.1467-8624.2011.01595.x

———. 2014. “Crawling and Walking Infants Elicit Different Verbal Responses from Mothers.” *Developmental Science* 17(3): 388–395. https://doi.org/10.1111/desc.12129

Krzton, Ali. 2018. “Supporting the Proliferation of Data-Sharing Scholars in the Research Ecosystem.” *Journal of EScience Librarianship* 7(2): e1145. https://doi.org/10.7191/jeslib.2018.1145

Macleod, Malcolm, Andrew M. Collings, Chris Graf, Veronique Kiermer, David Mellor, Sowmya Swaminathan, Deborah Sweet, and Valda Vinson. 2021. “The MDAR (Materials Design Analysis Reporting) Framework for Transparent Reporting in the Life Sciences.” Proceedings of *The National Academy of Sciences* 118(17): e2103238118. https://doi.org/10.1073/pnas.2103238118

McKiernan, Erin C., Philip E. Bourne, C. Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A. Nosek, Karthik Ram, Courtney K. Soderberg, Jeffrey R. Spies, Kaitlin Thaney, Andrew Updegrove, Kara H. Woo, and Tal Yarkoni. 2016. “How Open Science Helps Researchers Succeed.” *eLife* 5: e16800. https://doi.org/10.7554/elife.16800

Messinger, Daniel S., Whitney I. Mattson, James Torrence Todd, Devon N. Gangi, Nicholas D. Myers, and Lorraine E. Bahrick. 2017. “Temporal Dependency and the Structure of Early Looking.” *PLOS ONE* 12(1): e0169458. https://doi.org/10.1371/journal.pone.0169458

Myers, Jim D., and Margaret Hedstrom. 2014. “Active and social curation: Keys to data service sustainability.” National Data Service Consortium Planning Workshop. https://sead-data.net/wp-content/uploads/2014/07/ActiveandSocialCurationKeystoDataServiceSustainability.pdf

Ossmy, Ori, and Karen E. Adolph. 2020. “Real-Time Assembly of Coordination Patterns in Human Infants.” *Current Biology* 30(23): 4553-4562.e4. https://doi.org/10.1016/j.cub.2020.08.073

Soska, Kasey C., Scott R. Robinson, and Karen E. Adolph. 2014. “A New Twist on Old Ideas: How Sitting Reorients Crawlers.” *Developmental Science* 18(2): 206–218. https://doi.org/10.1111/desc.12205

Suls, Jerry. 2013. “Using ‘Cinéma Vérité’ (Truthful Cinema) to Facilitate Replication and Accountability in Psychological Research†.” *Frontiers in Psychology* 4: 872. https://doi.org/10.3389/fpsyg.2013.00872

Vines, Timothy H., Arianne Y.K. Albert, Rose L. Andrew, Florence Débarre, Dan G. Bock, Michelle T. Franklin, Kimberly J. Gilbert, Jean-Sébastien Moore, Sébastien Renaut, and Diana J. Rennison. 2014. "The Availability of Research Data Declines Rapidly with Article Age." *Current Biology* 24(1): 94–97. https://doi.org/10.1016/j.cub.2013.11.014

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(1): 160018. https://doi.org/10.1038/sdata.2016.18