PROJECT SUMMARY

Overview:

Video is a uniquely powerful tool for capturing the nuances of behavior in real time and for documenting changes in behavior across contexts and development. Just as microscopes and telescopes reveal natural phenomena invisible to the naked eye, video reveals the layered and multifaceted structure of behavior. Video documents research procedures and essential properties of computer-based tasks in ways that text-based descriptions and static images cannot. Video data collected for one purpose can be reused to address new questions, illuminate new phenomena, and open up new possibilities never imagined by the original researcher. Thus, videos of behavior accompanied by expertly-applied codes, videos of procedures and displays, and rich participant metadata constitute a substantial, largely untapped resource for new discovery in the social, behavioral, learning, and economic sciences and in computer and data sciences--if the materials can be widely and openly shared in reusable, easily discoverable formats. This project will create the infrastructure to bolster transparent, reproducible, integrative, interdisciplinary and insight-generating research about behavior by exploiting video.

The proposed infrastructure builds on Databrary.org, a video library the PIs developed and maintain with NSF, NIH, and Sloan Foundation support. This project has five aims: (1) Improve the reproducibility, transparency, and scalability of video-based research via LabNanny, a web-based scientific process management (SPM) system; (2) Accelerate data sharing by making self-curation of datasets more reliable, robust, and efficient; (3) Accelerate reuse of shared research video by making videos, metadata, and associated files easier to find, clone, and build upon; (4) Make behavioral research more interoperable, robust, and reliable through video-enhanced, web-based protocols that describe procedures and coding manuals that specify how to annotate the videos; and (5) Expand the use of video as data and documentation beyond the developmental science community. The activities associated with each aim will enhance video-centered discovery across the behavioral sciences and will specifically benefit a 65-member research community that is about to embark on a large-scale study of infant natural activity.

Intellectual Merit:

We will augment an existing, successful research ecosystem for discovery centered on video (Databrary.org), thereby accelerating the pace of discovery, expanding the breadth and depth of research, and facilitating transparency and reproducibility. The enhancements will enable novel, interdisciplinary, big-data research to address central questions about behavior by investigators across a wide range of fields--psychology, linguistics, anthropology, political science, behavioral economics, education and learning sciences, experimental biology, human-computer interaction, machine learning, and computer vision. The project will thereby create new opportunities for integrative and multidisciplinary research that is at present prohibitively expensive or impossible.

Broader Impacts:

The project will have broad impact across fields in the behavioral, social, biological, educational, learning, data and computer sciences that now use or could use video. The proposal will enrich datasets already shared on Databrary--many of them funded by NSF and NIH. The project will benefit fields that do not currently collect or code video by enhancing interoperability with existing repositories and making video data collection and curation more attractive. The project will expand opportunities for researchers at institutions with limited resources, including many outside the U.S., to participate in scientific discourse about behavior. This will expand research opportunities for researchers and students from underrepresented groups. The project will increase the quantity and quality of shared video datasets and associated materials. This will improve research transparency and boost reproducibility. Finally, the project will raise the profile of video-based behavioral research and bolster public interest in and support for the behavioral sciences.

TABLE OF CONTENTS

For font size and page formatting specifications, see PAPPG section II.B.2.

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

TABLE OF CONTENTS

For font size and page formatting specifications, see PAPPG section II.B.2.

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)		
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	0	
References Cited		
Biographical Sketches (Not to exceed 2 pages each)	2	
Budget (Plus up to 3 pages of budget justification)	6	
Current and Pending Support	2	
Facilities, Equipment and Other Resources	2	
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	2	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal. The entire proposal however, must be paginated. Complete both columns only if the proposal is numbered consecutively.

PROJECT DESCRIPTION

Researchers in the behavioral and social sciences aspire to understand human behavior in diverse contexts and from multiple perspectives, but these aspirations face profound challenges. Progress is stymied by research practices that undermine reproducibility and a scientific culture slow to embrace open data sharing[1–4]. Researchers fail to capture the full extent of human behavior with all its richness and complexity despite the availability of a uniquely powerful, inexpensive tool—*video recording*. Video faithfully captures the nuances of behavior in real time and precisely details how behavior changes across contexts and development. Video *documents research procedures* more completely than any text-based description in the methods section of a journal article[5–7]. Video collected for one purpose can be *reused to address new questions*, illuminate new phenomena, and open up new possibilities never imagined by the original researcher. Thousands of researchers in the behavioral and social sciences and related fields already collect video in lab, home, classroom, museum, and public settings[8–17].

Video has such significant untapped potential to transform behavioral and social science that PIs Gilmore and Adolph (along with their collaborator Tamis-LeMonda) have made it the centerpiece of the Play & Learning Across a Year (PLAY) Project (R01-HD094830), which received a 1st percentile ranking following peer review at NICHD in October 2017. The PLAY Project[18] exploits the power of video to catalyze discovery and transform knowledge about behavioral development in infancy. PLAY builds on the infrastructure of the NSF-funded, web-based Databrary video library and the Datavyu video coding tool. PLAY engages the joint expertise of 65 researchers across the United States and Canada who represent diverse domain knowledge, institutions, seniority, and race/ethnicity (Fig. 1A). This PLAY "Launch Group" will cooperatively collect hour-long video recordings in the homes of 900 infant-mother dyads from 30 sites across the U.S. The Launch Group will annotate these videos based on a common coding scheme designed by domain experts in language and communication, locomotion and physical activity, emotional expression, object interaction, gender, home environment, media use, and health. Datavyu—a powerful and flexible, free, open source desktop video coding tool—will be used for annotation. The Launch Group will openly share with the broader research community *all* videos, coding spreadsheets, questionnaires, ambient sound data, and video-documented procedures and code definitions via Databrary—the first repository designed specifically for sharing research video and associated materials from behavioral research. *PLAY will thus create the first, large-scale, fully transcribed, coded, curated, openly shared and readily reusable video corpus of human behavior*.

Here we propose to build research infrastructure—enhancements to Databrary—that will accelerate all kinds of video-intensive behavioral research and make it more efficient and effective while capitalizing on the unique opportunity that PLAY provides. The enhancements will bolster the reliability, robustness, reproducibility, and rigor of the Launch Group's research on the PLAY corpus, thereby contributing immediately to the productivity of a large, clearly identified, diverse research community. Moreover, by the end of the grant period, the same infrastructure will be made openly and freely available to the broader research community, thereby enabling transformative discoveries in many other fields for years to come.

Our proposal builds on existing practices and demonstrated technologies, making it simultaneously ambitious and pragmatic. Databrary is a web-based video library created with support from NSF (BCS#1238599) and NICHD (U01-HD076595) to foster open sharing and reuse of research videos, displays, coding files, and metadata among social and behavioral scientists. Launched in 2014, Databrary now houses 14,500+ hours of video from ~1000 researchers at 380+ institutions across the globe (Fig. 1B). Databrary provides the core infrastructure for PLAY. The proposed enhancements and extensions will maximize the

potential for discovery from all of the data in Databrary, including the PLAY corpus. The enhancements will expand Databrary's reach beyond the developmental science community where it originated into diverse fields across the behavioral and social sciences and beyond.

Project Aims

Aim 1: Improve the reproducibility, transparency, and scalability of video-based research via LabNanny, a web-based scientific process management (SPM) system

Video-based behavioral research involves a sequence of steps: multi-step workflows to collect and process multiple file types; iterative, looping steps to review, code, and conduct quality assurance on the data; tracking task assignments among various researchers/staff; and monitoring progress on project-wide recruitment, testing, coding, and analyses (Fig. 2). Single lab projects and large consortia face similar problems and implement similar workflows. But the unprecedented scale of PLAY and the commitment to openly share all data magnifies the challenges. Most video-based researchers depend on paper-and-pencil checklists or spreadsheets to manage files, people, processes, and tasks. In extensive piloting for PLAY, we determined that no free or open-source scientific process management (SPM) tool meets the needs of video-based research.

So we will develop "LabNanny," a new web-based SPM. LabNanny will be thoroughly vetted by the 65-member PLAY Launch Group as it manages training across geographically dispersed sites, the flow of video and other data files collected across 30 sites, assignment and outcome of QA reviews, training and assignment of coding passes to 48 labs and outcome of reliability checks on coded files, cross-checking of participant metadata, and finally monitoring recruitment, coding, and data-sharing goals. In planning PLAY, PIs Gilmore and Adolph piloted and refined the basic workflow using *ad hoc* tools, thereby specifying the requirements for LabNanny. Moreover, although the number of participating labs in PLAY is many times larger than comparable projects, the sequence of operations required is identical to that used by single labs and smaller collaborative teams. Thus, in developing and deploying LabNanny, we will create, test, and refine infrastructure that will be immediately valuable to the entire community of video-using behavioral and social scientists when openly shared at the end of the grant period.

Aim 2: Accelerate data sharing by making self-curation of datasets more reliable, robust, and efficient

Library scientists and directors of repositories agree: The burden of *post hoc* data curation is a major barrier to data sharing. Databrary encourages researchers to self-curate their data, one session at a time, by immediately uploading videos and associated files and entering participant metadata (e.g., age, sex, race/ethnicity) into a familiar spreadsheet interface. Self-curation while data collection is in progress has many advantages over *post hoc* curation, but the existing process is inefficient, error prone and does not scale well. Our two-part solution is to (1) create a standard project data format that Databrary can import, compatible with emerging standards in related fields; and (2) extend LabNanny to the desktop where it will manage a file storage service that synchronizes with Databrary without the need for manual intervention. Files stored in LabNanny's sync folders in the standard format will be automatically validated and synchronized with Databrary. This will substantially reduce the human cost of curating data for the broader community of Databrary users and for the PLAY Launch Group.

Aim 3: Accelerate reuse of shared research video by making videos, metadata, and associated files easier to find, clone, and build upon

Fostering widespread video sharing makes behavioral and social science more open, transparent, and robust. But to fully exploit the potential of open sharing to spark new

discoveries, videos and other data that meet specific researcher-driven criteria (e.g., participant age, sex, ethnicity, location, task) must be easy to find, filter, and reuse to yield new findings. We will enhance Databrary's indexing, search and filter functions so that investigators can find, select, and "clone" materials into new "virtual" collections meeting user-specified criteria. Databrary will track the provenance of cloned collections so that the original data owner's efforts are recognized and cited and new codes or analyses applied to videos become part of the searchable record of annotations for the original source. These features will also allow researchers to curate custom video collections based on tags or codes that others have applied, either to whole videos or to specific segments. These features will benefit the entire Databrary research community and will be especially useful to PLAY Launch Group members.

Aim 4: Make behavioral research more interoperable, robust, and reliable through video-enhanced, web-based protocols that describe procedures and coding manuals that specify how to annotate the videos

Video-based research typically begins when researchers generate text-based protocols of procedures that describe how a study will take place along with text-based coding manuals that describe the specific behaviors to be annotated by coders. Protocols and coding manuals come in static document formats (MS Word .docx, PDF, text) that do not exploit the richness of video exemplars. In piloting PLAY, the PIs developed a web-based electronic protocol of procedures and a coding manual using a web-based wiki engine that combines text-based descriptions of procedures and code descriptions with links to specific video exemplars stored on Databrary[19]. The PLAY Project wiki is currently stored separately from Databrary. We will develop infrastructure to incorporate this capacity internally. Databrary researchers can then create and share electronic protocols and coding manuals to fully document their research workflows, from participant recruitment through coding and data analysis. These enhancements will make video-based research more reproducible and transparent while accelerating reuse and new discovery.

Aim 5: Expand the use of video as data and documentation beyond the developmental science community

Large numbers of researchers who study behavior outside of developmental and learning sciences—in psychology, cognitive neuroscience, linguistics, animal communication, anthropology, political science, behavioral economics, education and learning, human-computer interaction, and so on—could readily and cheaply enrich their observations with video and benefit from more widespread open sharing of video. Furthermore, Databrary's human-annotated video and audio data constitute an unexploited but potentially invaluable resource for researchers in data science, computer science, computer vision, and machine learning. The PLAY corpus will increase that value many-fold. Project staff will expand the reach of Databrary's video-based research infrastructure by working with an advisory board of leading scholars in cognitive neuroscience, linguistics, animal communication, political science, education and learning sciences, computer science, and artificial intelligence. The board will help us engage these disparate research communities in the design, testing, and implementation of the proposed enhancements to Databrary so that the power of video can be extended to a wider and more diverse community of researchers. We will also expand and more clearly document Databrary's API, create specialized R and Python packages to accelerate use of Databrary's assets by the data science and AI communities, and forge links with other open science data services so that Databrary's materials can seed scholarship in diverse domains of science.

Results from prior NSF support

The PIs received funding from NSF (BCS#1238599, funding period 2012-2014, no cost

extension 2014-2016, $2,443,499; supplement BCS#1238599, funding period 2015-2016, no cost extension 2016-2017) to support Databrary and Datavyu. The primary aims of the prior awards were to build research infrastructure and provide training and technical support. In addition, we published articles that describe Databrary[20–25] and how it relates to other "big data" initiatives in developmental science[26] and the behavioral sciences more broadly[27,28]. We developed the first policy framework broadly endorsed by NSF, NIH, and a large community of researchers (~1000 researchers at 380+ authorizing institutions) for ethically sharing identifiable research data based on participant permission. We upgraded the Datavyu video-coding tool[29], held workshops to train researchers to code video[30], and wrote about best practices in behavioral video coding[31]. The current proposal builds on and extends these efforts. ***Intellectual Merit***. We created infrastructure to enable open sharing and reuse of research video in an open-source[32] web-based repository, Databrary[33], upgraded and provided user support for the Datavyu video-coding tool[29], and fostered a rapidly growing community of researchers committed to video sharing and reuse[34]. Databrary and Datavyu deepen and accelerate the pace of discovery in behavioral science by enabling researchers to view each other's datasets, reanalyze them to test competing hypotheses, and address new questions beyond the scope of the original study. ***Broader Impacts***. Databrary empowers behavioral scientists, especially from institutions with limited resources, to conduct high quality research; improves data management practices; and increases transparency and reproducibility. Datavyu brings the power of video-data coding to any laboratory with a computer. Databrary's policy framework makes it easy to securely share identifiable video data while upholding ethical principles. Our publications[5-7,20–28,31,36], workshops[18,30], and presentations are bringing this new, collaborative, integrated view of behavioral science to a larger audience.

Background & Rationale

Video has unprecedented power and untapped potential to transform understanding across the behavioral and social sciences. Video documents the microstructure of behavior in real time, across domains of function. Video uniquely documents the interactions between people and their physical and social environment with more richness, detail, and nuance than any other form of measurement. It does so with high spatial and temporal resolution. Video chronicles who did what, and how, when, and where they did it[7,9,20,25,35,36]. It closely mimics the visual and auditory experiences of live human observers, so video collected by one person for a particular purpose can be readily understood and reused by a different person for a different purpose. Indeed, video can make the anatomy of behavior as "tangible as tissue"[20].

Video has untapped value as documentation[7,36]. The "reproducibility crisis" in behavioral science[37,38] stems in part from *failure to adequately report and share essential details about procedures and codes*[7,36]. Text descriptions in methods sections necessarily omit details about both seemingly simple and complex procedures (e.g., recruitment calls, instructions to participants, administration of questionnaires, testing environments, and tasks). Words and static images cannot do justice to the subtle interactions and contextual features of typical test situations and computer-based displays. Many research paradigms involve special methods to elicit, test, and record behavior[39,40], such that procedures are like art forms, passed down from mentor to mentee[41]. Video coding manuals typically refer to tasks and behaviors with quirky, lab-specific labels that make the codes uninterpretable and unusable by others. We argue that video documentation of research procedures and codes should become standard practice[7,28,36].

Despite its potential, *too few behavioral scientists use video to document their procedures or collect video as primary data*. Those who collect video often lack tools and know-how to exploit its full potential; and most researchers who collect or code video do not share their videos or annotations. Too many research videos serve only as a backup for live human coding of an

event and go unanalyzed. When videos are coded, most researchers rely on makeshift spreadsheets or paper and pencil, not powerful annotation tools. Untold hours of video recordings paid for by scarce federal research dollars molder away in researchers' offices, file cabinets, and on defunct hard drives[30].

Databrary overcomes the most difficult challenges associated with open video data sharing. Videos contain personally identifiable information, posing problems for participants' privacy. Databrary's policy framework solves these problems. Databrary restricts video sharing to an institutionally authorized community of authorized researchers (Fig. 1B), and Databrary only shares identifiable data when researchers have secured permission from participants, using standard language Databrary developed or an IRB-approved equivalent. Large file sizes and diverse formats present technical challenges for video storage and sharing. Databrary solves these technical challenges by transcoding all files into a common format and by securely storing the original files and transcoded copies on large-capacity servers at NYU. Video sharing poses practical challenges of data management. Researchers lack time and resources to find, label, organize, link, and convert their files into formats that can be used and understood by others[42]; Many lack training and expertise in data curation[22], and thus do not adequately document workflows or data provenance. When researchers do share, standard practice involves organizing data after a project is finished, perhaps when a paper goes to press. "Preparing for sharing" after the fact is a difficult and unrewarding chore, one that often exceeds the incremental cost and reasonable time frame envisioned in NSF's Data Sharing Policy[43]. *Post hoc* sharing also makes curation a challenge for repositories[22]. Databrary solves these *post hoc* sharing problems by encouraging researchers to self-curate as they collect data, using drag and drop to upload files and a spreadsheet interface to enter participant and session metadata.

Databrary demonstrates that open sharing of video data and video documentation can improve scientific transparency and accelerate discovery. The PIs established Databrary with support from NSF and NICHD and additional funding from the Society for Research in Child Development and the Alfred P. Sloan Foundation. Databrary is the first digital library specialized for sharing research videos, experimental displays, coding files, and associated metadata. It provides a secure platform for storing and sharing among authorized researchers while fostering widespread reuse of data and materials and enhancing scientific transparency. Databrary has targeted the developmental and learning science communities because this is the PIs' intellectual home and the focus of our initial funders. But we designed Databrary for broader use in every domain of social and behavioral science where video could be useful. Since launching 4 years ago, the system has grown to serve ~1000 researchers from 380+ institutions around the world[34,44]. These investigators have contributed ~16,400 hours of video or audio, representing 12,190+ participants ranging in age from 6 weeks to elderly adults. The system stores 550+ datasets of which 115 are shared; the rest are studies in progress.

Databrary enables large-scale innovative, multidisciplinary behavioral research collaborations such as the PLAY Project. Inspired by the potential of large-scale open video sharing and the power of rich video annotation, the PIs developed the PLAY (Play & Learning Across a Year) project. PLAY will advance discovery about behavioral development in infancy, focusing on the critical period from 12 to 24 months of age when infants show remarkable advances in language, object interaction, locomotion, and emotion regulation. PLAY leverages the joint and diverse expertise of 65 Launch Group researchers and capitalizes on Databrary's infrastructure and the Datavyu coding tool. Together, PLAY researchers will collect, transcribe, code, share, and analyze a video corpus of infant and mother behavior in the home that is unprecedented in scope, breadth, and depth. The project will advance new ways to use video as documentation to facilitate discovery and ensure transparency and reproducibility.

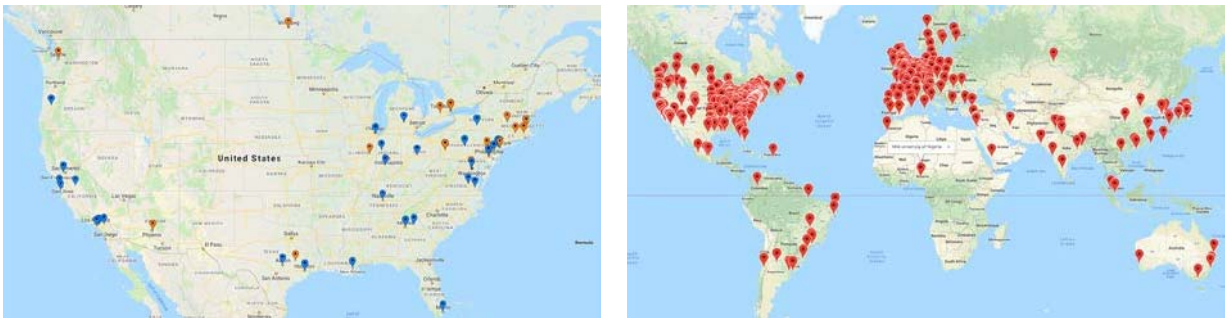

Figure 1. (A) PLAY project data collection (blue) and non-collection (orange) sites. All sites (blue and orange) will code or analyze data. (B) Map of Databrary-authorized institutions as of mid-February 2018.

PLAY has three aims: *Aim 1* is to create the first, cross-domain, large-scale, transcribed, coded, and curated video corpus of human behavior—collected with a common protocol and coded with common criteria jointly developed by the launch group. The corpus will consist of videos of 900 infant-mother dyads (12-, 18-, and 24-month-olds, 300 per age) from 30 diverse sites across the United States (Fig. 1A). Videos will be transcribed and coded for infant and mother speech, communicative acts, gestures, object interactions, locomotion, and emotion. The corpus will be augmented with video home tours and questionnaire data on infant language, temperament, locomotion/fall injuries, gender identity and socialization, home environment and media use, and family health and demographics. *Aim 2* is to leverage the potential of time-locked video codes to test critical questions about behavioral, developmental, and environmental cascades—from one domain to another, between infants and mothers, and from the macro environment (e.g., SES, geographic region, home language) and proximal home environment (e.g., objects for play, home chaos and clutter) to infant and mother behaviors. *Aim 3* is to advance new ways to use video as documentation to ensure scientific transparency and reproducibility. The entire protocol and code definitions are documented in a wiki[19] with exemplar video clips that augment text-based descriptions. The entire corpus and all tools[18] will be openly shared behavioral science community on Databrary and with other data repositories (TalkBank, WordBank, and the Open Science Framework).

PLAY will create a cross-domain, shared video corpus of unprecedented scope and richness. It will provide Launch Group members and the larger research community with the data, tools, and know-how to use time-locked video codes to investigate how natural behavior unfolds in real time. The novel, synergistic approach to "expert crowdsourcing" data collection, coding, and analysis will reduce overall costs while increasing scientific payoffs. PLAY will show how the use of *shared video as data and documentation will accelerate discoveries* in social and behavioral science.

Remaining barriers this project will overcome

Databrary overcomes most barriers to the secure, ethical sharing of video data. PLAY overcomes many barriers that impede progress in the science of early behavioral development. Now, proposed enhancements to Databrary will increase the value, scope, and impact of video-based research across the behavioral and social sciences with PLAY as the scalable test bed.

We must develop a new scientific process management tool, "LabNanny," and integrate it with Databrary in order to manage data collection, quality assurance reviews, coding, inter-observer reliability evaluation, open sharing, and the generation of 65+ scholarly papers describing the PLAY corpus results (*Aim 1*). A new standardized project file format and a facility for synchronizing local files to Databrary is required for efficient, accurate, and consistent curation

of video datasets including PLAY (*Aim 2*). Enhanced means of indexing, searching, and filtering data along with ways to fork and clone data subsets for investigator-specific analyses is required to accelerate scientific discovery from videos shared on Databrary (*Aim 3*). A centralized, readily accessible, web-based electronic protocol and coding manual with video exemplars is required to ensure the implementation of consistent video-based research protocols, whether across campus or across the 30 PLAY data collection and 48+ coding sites (*Aim 4*). With PLAY as the test bed, the infrastructure we develop to support large-scale video-based behavioral research will be made accessible and useful to the broader community of scientists in linguistics, cognitive neuroscience, demography, political science, animal communication, education and learning sciences, computer vision and other fields (*Aim 5*).

Implementation Plan

We will implement five primary projects, each aligned with a specific aim.

Aim 1: Improve the reproducibility, transparency, and scalability of video-based research via LabNanny, a web-based scientific process management (SPM) system

We will develop and deploy a scientific process management (SPM) system to support video-based behavioral research. The LabNanny SPM system must empower project staff to create multi-step workflows and assign people to steps in those workflows (Fig. 2). It must allow files (video, coding spreadsheets, questionnaires, etc.) to be linked to tasks and people and passed along via workflows. The system must allow local (collecting and coding PIs) to monitor their own progress and central staff to monitor progress of the project as a whole. Our plan to deploy LabNanny is divided into three sub-tasks informed by our preliminary work.

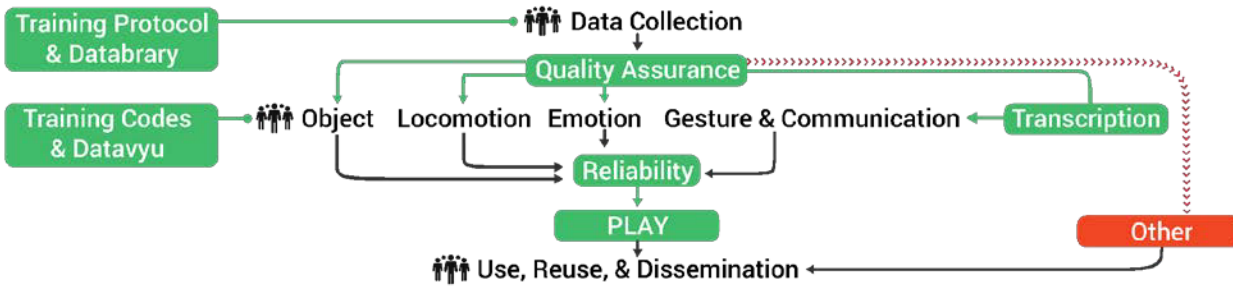

Figure 2. Schematic of workflow for PLAY (and video-based projects in general). Paths in green show flows from PLAY project staff to Launch Group labs. Paths in black show flows from Launch Group Data Collection or Coding labs. Videos that do not pass Quality Assurance will be shared as Other data. Coded sessions that are not reliable will be recoded until they pass checks for inter-observer reliability.

Preliminary work: The PIs interviewed colleagues and Project Advisory Board member Nilam Ram who have successfully conducted similar large-scale, geographically distributed research projects involving multiple data types, including video. We learned that most labs use a combination of checklists and spreadsheets, and sometimes a custom relational database with shared/cloud file storage. We evaluated whether existing project/process management tools (e.g., JIRA, Asana, Trello, ProcessStreet, ProcessMaker, ProcessMate, JBPM, Camunda, Flowable, Aquarium) might be adapted for our use. In so doing, we learned that none of the tools will work without substantial modification, most of the web-based tools lack the integration between tasks and files that we require, and the scale of PLAY pushes the project outside the bounds of what cloud services provide for free.

Project 1.1: Workflow definition; task assignment and queuing

We plan to fork an existing open source system—ProcessMaker is the leading candidate—and adapt and extend it to the needs of video-based research projects like PLAY. In forking, adapting, and extending an existing system, we must allow users to create multi-step workflows and assign tasks within workflows to people. We must also integrate the forked system into Databrary's authentication/permission and user notification system. The result we will call LabNanny to highlight the central (and distinct) role that scientific process management and workflow management plays in robust and reproducible research. LabNanny will require interfaces that allow researchers to create workflows for training, data collection, QA, coding, inter-observer reliability, and data sharing (see Fig. 2). Should forking prove unworkable for some reason, our backup solution is to create our own SQL-style database using PostgreSQL, the core of Databrary's current backend.

Project 1.2: File management

LabNanny must permit files to be attached to specific workflows. For example, once one of the 30 PLAY data collection labs has finished a testing session, that lab will have to upload the video, ambient sound, and questionnaire files to Databrary and assign the newly collected video for QA review by NYU-based PLAY staff (see Fig. 1A and Fig. 2). That assignment must trigger a notification to the PLAY project staff who must subsequently assign the video to a particular person for review. All files uploaded to Databrary get their own URIs (unique links) after they are transcoded into a common video format, and Databrary's API checks file-level access permissions before showing a user a particular file. So, we will build the file management component of LabNanny alongside Databrary's existing framework by attaching file-level URIs to specific workflow steps. This should eliminate the need to send files around like email attachments, a solution unworkable for multiple reasons (privacy, file sizes, etc.). Nevertheless, Databrary's existing back-end is based on the constraint that a file uploaded to one dataset (project) and session (date + time + person) remains there. We will have to implement ways for the same files to progress through multiple, parallel workflows (e.g., multiple coding and review passes), while maintaining the integrity of Databrary's project, session, and file permission structure. Our current plans involve extending Databrary's "tagging" feature to include tags related to a given file's status within the overall workflow (e.g., "Pending QA review", "Transcription Completed", or "Ready to be Shared").

Project 1.3: Progress monitoring dashboard and analytics

For LabNanny to be an effective tool for video-centered research projects, PIs must be able to monitor progress in real time at every phase of the project, from participant recruitment, to data collection, to QA review, to coding, analysis, and ultimately data sharing. In testing LabNanny with PLAY, we will implement dashboards that will show real-time progress reports—in data collection across the 30 sites, in QA review across 900+ videos, in coding across 48+ sites, in achieving inter-observer reliability across 9000+ coding passes, and in transferring all videos, Datavyu coding spreadsheets, and related data files to the appropriate Databrary datasets for sharing. We will also implement dashboards for Launch Group PIs who collect data to monitor their own recruiting progress (including participant demographic targets), and for coding labs to monitor their lab's progress in meeting coding goals. These monitoring and reporting tools will build on Databrary's existing spreadsheet interfaces that capture participant characteristics (age, sex, race/ethnicity) and internal reporting functions that provide PIs with summary information about participants and sessions within a data volume.

Aim 2: Accelerate data sharing by making self-curation of datasets more reliable, robust, and efficient

To accelerate data sharing we will develop and deploy cloud-based project file sync services to enhance self-curation of video datasets. Currently, researchers must manually upload videos and associated data files to Databrary and must hand-enter participant and session-related metadata. This makes self-curation burdensome and error-prone—in the context of PLAY, 30 different labs will be collecting, uploading, and entering 30+ sessions each. Project 2 will make the ingestion and curation of research materials more reliable and efficient by creating a standard, machine-readable project structure and a local file syncing mechanism, thereby making data curation less burdensome and more reliable for all Databrary researchers.

Preliminary work: We have consulted with Russell Poldrack, founder of the Stanford Center for Reproducible Neuroscience[45,46] and Databrary Board member, and with Chris Gorgolewski, leader of the community-developed Brain Imaging Data Structure[47]. BIDS is an emerging standard for representing the task and analysis structure of brain imaging studies, including studies that involve measures of human behavior that have a dense temporal structure similar to video. Indeed, BIDS allows researchers to store participant, task, and measurement data in a format that is easily readable by many brain imaging data repositories. BIDS uses standard, open formats (tab-separated text files, JSON), and our consultations confirmed that the BIDS format can be readily adapted for use by Databrary. Moreover, increasing numbers of cognitive neuroscientists use video as display materials or as behavioral data streams, so cooperating on the creation of an open standard for video-centered behavioral research will pave the way for further integration among research communities and open science repositories (Project 5).

Project 2.1: Adapt BIDS to suit PLAY and video-centered research more broadly

Working with Poldrack and Gorgolewski, we will adapt the BIDS project specification so that it reflects the data and metadata essential for video-based research, keeping in mind the MPEG-7 multimedia annotation format scheme[48] and the Web Annotation Data Model[49]. We will develop API components for Databrary to import session-specific project data using the new format and ensure that the import process proceeds smoothly and reliably. We will also develop and distribute documentation about the format on the Databrary site and more broadly (Project 5).

Project 2.2: Develop a desktop extension of LabNanny for uploading video data and related metadata into Databrary

Rather than require researchers to upload videos and enter metadata by hand, we will develop a desktop extension of LabNanny to manage file synchronization to and from Databrary. LabNanny will link specific desktop file directories or folders to a particular Databrary volume. This will keep responsibility for the control of local video files and related metadata with individual PIs to reflect the reality that security requirements may differ from institution to institution as may ethics board or IRB policies. LabNanny will verify that files stored in the target desktop directory are readable and formatted properly, then copy and enter those files into the Databrary system. Uploads can occur immediately or be scheduled for off hours when network bandwidth demands are lower. Most researchers collect participant metadata electronically using a desktop spreadsheet—PLAY will do so using a custom tablet app that will export participant metadata into a BIDS-compatible format. LabNanny will require researchers to map their individual metadata field names for a particular project to the standard Databrary schema. These stored mappings will allow LabNanny to create new data collection sessions on Databrary and populate them with videos and other data files.

After we have file uploads working well, we will enable LabNanny to pull files from Databrary

and save them locally. For example, when PLAY project staff assign a particular video to a specific lab for a particular coding pass (e.g., locomotion and physical activity), LabNanny (on Databrary) will generate a template Datavyu coding file with all the appropriate codes and links to code definitions (Project 4). The video and coding file template will be tagged for the assigned lab's use, and LabNanny can then manage manual or automatic syncing to the assigned researcher's designated local "download/coding" directory.

Aim 3: Accelerate reuse of shared research video by making videos, metadata, and associated files easier to find, clone, and build upon

Because video captures so many varied dimensions of behavior, it can be reused to answer new questions beyond the focus of the original study. Project 3 implements features that will enhance researchers' ability to find, filter, select, combine, clone, reuse, and extend analyses of video datasets shared on Databrary.

Project 3.1 Enhancing indexing, searching, and filtering

To reuse shared data, researchers must be able to find and select portions based on the specific questions they wish to answer. Databrary has an existing search feature, but it is limited in scope and depth.

Preliminary work. Databrary currently allows researchers to search for keyword terms linked to shared datasets. It also allows researchers to filter datasets based on participant age (if stored in the Databrary spreadsheet) and session-level information (sharing level, file type). These capabilities make it possible to search Databrary for datasets that include specific participant ages and to preview video excerpts when they are available. Powerful search and filtering operations are essential for transforming a repository from a passive storage facility to an active tool for analysis and discovery.

We aim to take Databrary's search and filtering capabilities to the next level by empowering researchers to search for datasets that contain codes for particular behaviors or involve specific tasks or manipulations. We will also expand the number of participant demographic fields that can be used for filtering and search. We will build a search interface that returns information about which datasets (or videos within datasets) meet specific search requirements—whether coding files and coding manuals are available, and whether the coding files or manuals contain specific codes. The interface will allow users to select and explore specific datasets or individual videos returned by the search and then choose what to do with the matching items.

Along the way we must enhance or replace Databrary's existing search engine, currently based on Apache Solr. We will need to index and search for entire datasets (and eventually specific segments of videos within those datasets) that match selected criteria. These searches will be based on available metadata, protocols, and code descriptions in the associated coding manuals (Project 4.1).

Project 3.2: Create new virtual collections of shared videos for reanalysis

Finding shared video data that meet a researcher's needs is an essential first step to creating aggregated collections of shared data that can be reused for new purposes. Imagine a researcher interested in studying whether speech content, quantity, or quality varies depending on whether speakers are walking. Using the enhanced search features described in Project 3.1, the researcher could find a dozen studies with several hundred videos that meet her criteria—videos of children and adults walking and stationary, either with or without speech transcripts. At present, the researcher could download each dataset to her local computer to begin exploring whether the data could be recoded to answer her question. But the management of found videos, their sources, and citation information is difficult, time-consuming, and error-prone. The

original data providers have no information that their shared data were found and downloaded, and Databrary has only limited information about the extent of reuse. The provenance of the painstakingly collected video data is undermined and along with it the potential for new sets of codes to add layers to the accumulated information about the videos. Project 3.2 will provide solutions to these problems.

Preliminary work: Databrary already generates persistent identifiers for shared datasets; volume and session interfaces provide vital metadata for reuse; datasets can have multiple external links to other web-based resources; and videos, coding files, and other stored materials have unique, internally-generated, resolvable uniform resource identifiers (URIs).

We propose enhancements to Databrary that allow researchers to create new virtual datasets that contain collections of videos (and other data) derived or cloned from other datasets. The virtual datasets will be stored and presented using the same dataset interface Databrary now uses, with the exception that individual sessions would consist of a set of (A) videos linked directly to the dataset, plus (B) links to sessions or specific files stored in other datasets. The other volumes could be owned by the researcher or by another authorized user. Databrary will indicate the source(s) of linked videos for transparency (and the timestamp the sources were cloned), and the system will automatically link to system-generated citations so that the new researchers can cite the materials from which their new study draws. The raw videos and shared coding files would not be copied, but linked to, thus saving storage space, and making it easier to track provenance. After a new data collection is shared with the Databrary community, links from the original source to the new collection will be added to the original dataset so that researchers can track how videos are being reused. If a researcher adds one or more coding passes to a video—e.g., adding (human or machine-generated) speech transcript data to videos that lack it—those coding files will be linked back to the original video so that new studies can build on both the original and newly applied codes. A new researcher can also add new or revised code definitions to the electronic protocol/coding manual associated with the new collection (Project 4). Those manuals will get linked back to the original sources.

Implementing the virtual datasets feature will entail modifications to the existing Databrary volume, spreadsheet, and session interfaces. The modifications will distinguish the visual representation of "virtual" or linked components from those that are directly associated with a dataset. We will need to enable users to save a set of found and filtered data to a new dataset. We will also have to modify the Databrary backend to keep track of which dataset components are linked from other sources and which are not, add new notifications when existing datasets are cloned or linked to, and other components. *A transparent, reproducible, and robust chain of knowledge about shared videos* that empowers researchers to easily build on one another's findings and transparently share discoveries *has significant capacity to enhance the scholarly value of shared data*.

Aim 4: Make behavioral research more interoperable, robust, and reliable through video-enhanced, web-based electronic protocols of procedures and coding manuals that specify how to annotate the videos

Detailed protocols and coding manuals make video-based behavioral research maximally transparent and reproducible. These documents make concrete critical details about procedures and the codes used to capture different behaviors. In turn, the information in protocols and coding manuals provide invaluable materials for generating standard ontologies.

Few researchers share their protocols or coding manuals, and limitations in journal space relegate essential details to supplemental material sections. PLAY is committed to sharing all details of the project's research procedures, equipment, coding definitions, and data manipulations openly with the research community. In addition, PLAY will demonstrate the

power of video as documentation by bringing web-based, electronic protocols and coding manuals to Databrary.

Preliminary work: The PIs created a wiki-based electronic protocol of the procedures and coding manual in planning for PLAY[19] (Fig. 3). The wiki-based format offers many advantages, including flexibility—it supports text, images, and video—and components like code or procedure definitions can be linked to specific URLs. In developing the capacity for electronic protocols and coding manuals in Databrary, we will bring this powerful tool to a wider audience.

Project 4.1 Create a wiki-based electronic protocol and coding manual engine within Databrary

For UI/UX consistency, we will build electronic protocol and coding manual features within Databrary by adapting and incorporating an existing wiki framework into our own code base. Building this within Databrary allows us to incorporate version control features that are essential for research transparency, add information in the protocols and coding manuals to Databrary's search index (see Project 3.1), and ideally capture LabNanny (Projects 1) workflows and represent them within the protocol. Also, if linked video clips are stored on Databrary, only authorized users who are signed into the system can view the clips. This protects participants' identities, and makes it more attractive for researchers to add video clips from their own studies.

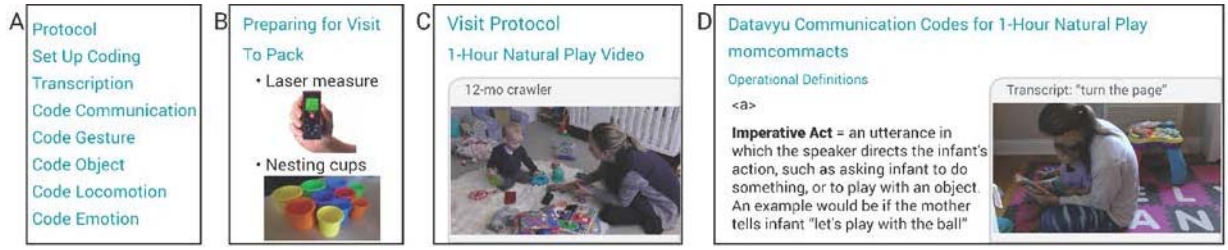

Figure 3. Illustrations from the PLAY wiki[19] that serves as electronic protocol and coding manual. (A) Outline of the wiki's subsections including video-augmented protocol steps and coding definitions. (B) Equipment and apparatus needed. (C) Illustration of a sample visit, including camera angles. (D) Detailed text descriptions of all behavioral codes, augmented with video clips.

In adding these capabilities to Databrary via a dataset-level wiki, we will create a standardized, searchable, web-accessible way for researchers to capture vital details about their procedures. Scientists perusing a paper's methods section can go to a linked Databrary volume to read a full description of the protocol and watch a linked video clip (stored on Databrary) demonstrating how a procedure is carried out in practice. Moreover, for those datasets (like PLAY) where individual videos have been annotated, users who view a video on Databrary can click on a code and read its definition from the manual associated with the study. Researchers can compare code definitions across shared datasets to evaluate their clarity and consistency (Project 3). This information can spark conversations about opportunities to achieve consensus around conceptual ontologies in areas where there is currently no consensus. Furthermore, the new standard (BIDS-compatible) project file standard (Project 2.1) will incorporate a URL to the shared protocol/coding manual. Researchers can also view links to individual code definitions from within Datavyu (or other) video coding files. Thus, when human coders annotate a video file, they will create time-locked indices for video segments that can be used for subsequent search and filtering (Project 3) across a dataset or across the library (e.g., find all instances of people laughing). In turn, these indexable, searchable, video segments, coupled with text-based descriptions of behaviors form invaluable training or test data for machine-learning models that we will make more broadly accessible in Project 5.

Aim 5: Expand the reach of video-based research infrastructure beyond the developmental and learning sciences

Databrary and Datavyu enable behavioral scientists to mine video data and share it openly, securely, and ethically. We propose three activities that will accelerate the reuse and reanalysis of Databrary's materials by an even wider community of scientists and make the system both a data repository and a platform for new discovery.

Preliminary work: We organized a Project Advisory Board consisting of leaders from cognitive neuroscience (Russell Poldrack, Stanford; Chris Gorgolewski, Stanford), demography (Guangqinq Chi, PSU), linguistics (Mike Frank, Stanford; Brian MacWhinney, CMU), animal communication (Emmanuel Chemla, Ecole Normale Supérieure), political science (Diana Kapiszewski, Georgetown and the Qualitative Data Repository), education and learning science (Felice Levine, American Educational Research Association), data science (Todd Gureckis, NYU; Nilam Ram, PSU; Jeffrey Spies, UVa), and computer science/AI (Juliana Freire, NYU; Vasant Honavar, PSU; Florian Metze, CMU; Jim Rehg, Georgia Tech). These leaders will meet annually with the PIs and communicate with and engage their respective research communities in the development of project data standards, testing and deployment of project features, reuse of shared video data, and interoperability with existing data sharing services.

Project 5.1: Polish, document, and promote Databrary's API

We will polish, document, and promote the use of Databrary's API (see also Project 1.2) so that Databrary's growing assets can be made more accessible and useful to a wider range of scientists. Databrary's architecture already involves interaction between the front and back-ends via an API, but the API remains largely undocumented. Polishing and documenting the API will make it easier for researchers to extract Databrary metadata and data in organized ways for subsequent reanalysis. Promoting use of the API via communication with relevant communities through conference presentations, webinars, workshops, and email lists will expand and diversify the community of researchers who can and do use the system.

Project 5.2: Create R and Python packages for accessing Databrary

Building on the refreshed API, we will develop and release Databrary-specific software packages in the R and Python programming languages to facilitate two-way interaction with Databrary's resources. These languages are commonly used in data science. For example, PI Gilmore has created R-based workflows for his own use that involve cross-validating participant metadata stored on Databrary with information stored locally via electronic spreadsheets. These components can be built on and refined to allow other researchers to automate data visualization or analysis workflows, including the application of machine-learning techniques.

Project 5.3: Link with existing repositories and data services

With the foundations laid by Projects 5.1 and 5.2, we will develop mechanisms to synchronize specific data components stored on Databrary with existing data repositories serving other research communities. For example, the child and mother speech transcript data generated by PLAY will be readily exported into the CHAT format used in the TalkBank[50,51] family of databases directed by Project Advisor (and PLAY Launch Group member) MacWhinney. We will contribute the PLAY transcripts to TalkBank and explore other ways to link similar datasets between the systems. Similarly, PLAY will provide parent-report data about child vocabulary using the MacArthur-Bates Communicative Development Inventory (M-CDI) which Project Advisor (and PLAY Launch Group member) Frank collects and openly shares with the research community via WordBank[52,53]. We will export PLAY's M-CDI data in formats that are compatible with WordBank and explore automating the process and providing reference visualizations based on WordBank norms in Databrary. We will explore ways to allow brain imaging

researchers who share data using the OpenNeuro site[46,54] directed by Project Advisors Poldrack and Gorgolewski to store project-related videos on Databrary, as well. Finally, we will work with Project Advisor (and Launch Group member) Chi to bring spatial visualization capabilities (e.g., interactive WebGIS) to Databrary, building on the individual-level Census Block Group codes that will be collected and shared along with other PLAY data elements.

Coordination & Management Plan

The project will be overseen by PIs Gilmore and Adolph. The PIs currently meet by phone or video conference several times weekly to discuss project-related matters. The PIs meet weekly with the development team to formulate long- and short-term plans, get progress updates, and provide input. The Databrary Managing Director, Ahmad Arshad, coordinates daily operations and will hire and supervise the technical team. In addition, the Databrary project has input from an Advisory Board of experts internal to PSU and NYU and external advisers who bring expertise in data sharing and developmental science[55]. Many members of the Project Advisory Board for this proposal are members of the Databrary Advisory Board. The Databrary Advisory Board meets annually to hear project updates and provide guidance about policy and technical matters; we will add a project-specific meeting to that annual gathering to reduce the travel and time burden on our advisors.

Evaluation and Assessment

We will evaluate progress in several ways. The PIs will report progress for each specific project relative to the goals outlined in the timetable (Technical Plan) in the annual project report. Several of the progress metrics will relate directly to the use of the new tools by the PLAY Launch Group. When a particular feature is released to the larger Databrary community, we can then begin to gather broader use statistics, such as the number of researchers using the BIDS-compatible upload/sync, the number of users who are creating electronic protocols and coding manuals, and the numbers and types of searches. We will also develop estimates of data reuse based on download statistics, the generation of new (cloned) video collections, and citation counts. The team will administer surveys to Databrary users twice a year to solicit feedback about system operations, focusing on new features, and asking users the extent to which the new features change their willingness to share data, the ease of doing so, and the attractiveness of reusing others' data. The results of those surveys will be summarized in the annual NSF report and discussed at the annual Databrary advisory board meetings.

Summary

Video uniquely and powerfully captures human behavior and therefore naturally serves as the focal point of a research ecosystem for behavioral discovery. This project will (Aim 1) improve the reproducibility, transparency, and scalability of video-based behavioral research through the development and deployment of the LabNanny web-based scientific process management (SPM) system; (Aim 2) accelerate video data sharing by making self-curation of video-based scientific datasets more reliable, robust, and efficient; (Aim 3) accelerate reuse of shared research videos by making videos and related metadata easier to find, clone, and build upon; (Aim 4) make annotation of video data files and associated metadata more interoperable, robust, and reliable through video-enhanced, web-based electronic protocols and coding manuals; and (Aim 5) expand the reach of video-based research infrastructure beyond developmental science. Putting the richness and complexity of behavior into sharper focus by making video recording, coding, and sharing commonplace will reinvigorate discovery in the behavioral and social sciences while making these fields leaders in research transparency and reproducibility.

The proposal addresses multiple criteria for the RIDIR program. With respect to *science*, the enhancements to Databrary will enable new, integrative, and interdisciplinary research questions about the characteristics and consequences of human behavior across ages, domains of function, and contexts to be answered with unprecedented depth, breadth, and impact. The research communities interested in exploring these questions span psychology, linguistics, anthropology, political science, behavioral economics, education and learning sciences, experimental biology, and human-computer interaction. With respect to *information technology*, the project elevates the status and importance of video, both as data and documentation, and substantially enhances the diversity of metadata provided for and linked to shared video, improves the tools for exploiting the links, and expands the communities that benefit. The proposal builds on new, but established infrastructure with strong institutional backing, and a pair of dedicated PIs who are committed to long-term sustainability for the tools. With respect to *governance*, the project builds on the joint expertise of the 65-member PLAY Launch Group and a newly formed Project Advisory Board. In addition, the PIs have established, maintain, and regularly consult with a diverse Databrary Advisory Board[55] that is broadly representative of expertise in the behavioral and open science communities. The Board's input has influenced this project's emphasis on interoperability and integration with existing resources and its plans for future links with others.

Intellectual Merit

The project will transform the pace, breadth, depth, transparency, and reproducibility of research in the behavioral and social sciences by creating novel, powerful infrastructure for discovery that is centered on video. Through substantial expansion of and improvements to Databrary, a successful video data sharing repository, the project will enable novel, innovative, data-intensive research on behavior, most immediately, the Play & Learning Across a Year (PLAY) project. Beyond developmental science, the project will impact substantial numbers of investigators across a wide range of fields that study behavior—psychology, linguistics, anthropology, political science, behavioral economics, education and learning sciences, experimental biology, and human-computer interaction. The proposed enhancements to Databrary will bring powerful, flexible, affordable, and innovative tools to bear on the central questions in social and behavioral science while tackling head-on the challenges these fields face with regard to transparency and reproducibility. The project will thereby create new opportunities for integrative and multidisciplinary research that is at present prohibitively expensive or impossible.

Broader Impacts

The project will have broad impact across fields in the behavioral, social, biological, and educational sciences that could or now use video, including substantial numbers of researchers outside the U.S. (Fig. 1B). The proposed enhancements will enrich the datasets shared on Databrary—many of them funded by NSF and NIH. The project's tools will expand opportunities for scientists at institutions with limited resources to participate in scientific discourse about behavior across disciplines. Because many of these institutions serve students from underrepresented groups, the project will expand research opportunities for them as well. By making data sharing more attractive to scientists, the project will increase the quantity and quality of shared video datasets and the richness of the metadata linked to them. By making coding files, protocols, and coding manuals more readily sharable, the project will improve transparency and boost reproducibility. Finally, the project will raise the profile of video-based behavioral research and bolster public interest in and support for the behavioral and social sciences.

REFERENCES

1. Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A. & David, S. P. Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. Trends Cogn. Sci. 18, 235–241 (2014).
2. Munafò, M. R. et al. A manifesto for reproducible science. Nature Human Behaviour 1, 0021 (2017).
3. Nosek, B. A. & Bar-Anan, Y. Scientific utopia: I. Opening scientific communication. Psychol. Inq. 23, 217–243 (2012).
4. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. Perspect. Psychol. Sci. 7, 615–631 (2012).
5. Gilmore, R. O. & Adolph, K. E. Video can make science more open, transparent, robust, and reproducible.
6. Gilmore, R. O. & Adolph, K. E. Open sharing of research video: Breaking the boundaries of the research team, in Advancing Social and Behavioral Health Research through Cross-disciplinary Team Science: Principles for Success. (Springer).
7. Adolph, K. E., Gilmore, R. O. & Kennedy, J. L. Video data and documentation will improve psychological science. http://www.apa.org Available at: http://www.apa.org/science/about/psa/2017/10/video-data.aspx. (Accessed: 13th February 2018)
8. Derry, S. J. et al. Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. Journal of the Learning Sciences 19, 3–53 (2010).
9. Goldman, R., Pea, R., Barron, B. & Derry, S. J. Video research in the learning sciences. (Routledge, 2014).
10. Alibali, M. W. & Nathan, M. J. Embodiment in mathematics teaching and learning: Evidence From learners' and teachers' gestures. Journal of the Learning Sciences 21, 247–286 (2012).
11. Masats, D. & Dooly, M. Rethinking the use of video in teacher education: A holistic approach. Teaching and Teacher Education 27, 1151–1162 (2011).
12. Pasqualino, C. Filming emotion: The place of video in anthropology. Vis. Anthropol. Rev. 23, 84–91 (2007).
13. Video sharing, deep tagging and annotation. Available at: http://cmdbase.org/. (Accessed: 13th February 2016)
14. Qualitative data repository. Available at: https://qdr.syr.edu/. (Accessed: 13th February 2016)
15. Chaquet, J. M., Carmona, E. J. & Fernández-Caballero, A. A survey of video datasets for human action and activity recognition. Comput. Vis. Image Underst. 117, 633–659 (2013/6).

16. Rautaray, S. S. & Agrawal, A. Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review 43, 1–54 (2012).

17. Tay, L., Jebb, A. T. & Woo, S. E. Video capture of human behaviors: toward a Big Data approach. Current Opinion in Behavioral Sciences 18, 17–22 (2017).

18. Adolph, K.E., Gilmore, R.O., & Tamis-LeMonda, C.T. Play & Learning Across a Year (PLAY) workshop at NICHD. PLAY Project: NICHD Workshop (2016-12-16) Available at: https://nyu.databrary.org/volume/254. (Accessed: 13th February 2018)

19. The PLAY project Wiki. Available at: https://dev1.ed-projects.nyu.edu/wikis/docuwiki/doku.php. (Accessed: 17th February 2017)

20. Adolph, K. Video as data. APS Obs. 29, (2016).

21. Adolph, K. E., Gilmore, R. O., Freeman, C., Sanderson, P. & Millman, D. Toward open behavioral science. Psychol. Inq. 23, 244–247 (2012).

22. Gordon, A. S., Millman, D. S., Steiger, L., Adolph, K. E. & Gilmore, R. O. Researcher-library collaborations: Data repositories as a service for researchers. Journal of Librarianship and Scholarly Communication 3, (2015).

23. Gilmore, R.O., Adolph, K.E., & Millman, D.S. Curating identifiable data for sharing: The Databrary project. in 2016 New York Scientific Data Summit

24. Gilmore, R. O., Gordon, A., Adolph, K. E. & Millman, D. S. Transforming education research through open video data sharing. WSEAS Trans. Adv. Eng. Educ. 5, (2016).

25. Gordon, A. S., Steiger, L., & Adolph, K. E. Losing research data due to lack of curation and preservation. in Curating research data Volume 2: A handbook of current practice (ed. Johnston, L.) (Association of College and Research Libraries).

26. Gilmore, R. O. From big data to deep insight in developmental science. Wiley Interdisciplinary Reviews: Cognitive Science 15 (2016). doi:0.1002/wcs.1389

27. Gilmore RO, Diaz MT, Wyble BA, Yarkoni T. Progress toward openness,transparency, and reproducibility in cognitive neuroscience. OSF preprint file

28. Gilmore, R. O., Kennedy, J. L. & Adolph, K. E. Practical Solutions for Sharing Data and Materials From Psychological Research. Advances in Methods and Practices in Psychological Science 2515245917746500 (2018). doi:10.1177/2515245917746500

29. Datavyu: Video coding and data visualization tool. Available at: http://datavyu.org/. (Accessed: 15th February 2017)

30. Best practices for coding behavioral data from video. Datavyu: Video coding and data visualization tool Available at: http://datavyu.org/user-guide/best-practices.html. (Accessed: 15th February 2017)

31. Databrary Repository on GitHub. Available at: https://github.com/databrary. (Accessed: 15th February 2017)

32. Databrary. Available at: https://databrary.org/. (Accessed: 8th July 2015)

33. Authorized Databrary investigators. Available at: https://nyu.databrary.org/search?volume=false&f.party_authorization=4&f.party_is_institution=false. (Accessed: 10th February 2016)

34. Curtis, S. 'Tangible as tissue': Arnold Gesell, infant behavior, and film analysis. Sci. Context 24, 417–442 (2011).

35. Gilmore, R. O., Adolph, K. E., Millman, D. S. & Gordon, A. Transforming Education Research Through Open Video Data Sharing. Adv Eng Educ 5, (2016).

36. Gilmore, R. O. & Adolph, K. E. Video can make behavioural science more reproducible. Nature Human Behavior 1, (2017).

37. Open Science Collaboration. Estimating the reproducibility of psychological science. Science 349, aac4716–aac4716 (2015).

38. Harris, R. Rigor Mortis: How Sloppy Science Creates Worthless Cures, Crushes Hope, and Wastes Billions. (Basic Books, 2017).

39. Bornstein, M. H., Arterberry, M. E. & Lamb, M. E. Development in Infancy: A Contemporary Introduction. (Psychology Press, 2013).

40. Miller, S. A. Developmental Research Methods. (SAGE Publications, Inc, 2017).

41. Peterson, D. The Baby Factory Difficult Research Objects, Disciplinary Standards, and the Production of Statistical Significance. Socius: Sociological Research for a Dynamic World 2, 2378023115625071 (2016).

42. Ascoli, G. A. The ups and downs of neuroscience shares. Neuroinformatics 4, 213–216 (2006).

43. Dissemination and sharing of research results. National Science Foundation Available at: https://www.nsf.gov/bfa/dias/policy/dmp.jsp. (Accessed: 15th February 2017)

44. Institutions with authorized Databrary investigators. Available at: https://nyu.databrary.org/search?offset=0&volume=false&f.party_authorization=5&f.party_is_institution=true. (Accessed: 10th February 2016)

45. Stanford Center for Reproducible Neuroscience. Available at: http://reproducibility.stanford.edu/. (Accessed: 10th February 2018)

46. OpenNeuro. Available at: https://openneuro.org/. (Accessed: 13th February 2018)

47. Brain Imaging Data Structure. Available at: http://bids.neuroimaging.io/. (Accessed: 10th February 2018)

48. Wikipedia contributors. MPEG-7. Wikipedia, The Free Encyclopedia (2017). Available at: https://en.wikipedia.org/w/index.php?title=MPEG-7&oldid=807456995. (Accessed: 11th February 2018)

49. Web Annotation Data Model. Available at: https://www.w3.org/TR/annotation-model/. (Accessed: 11th February 2018)

50. TalkBank. Available at: http://talkbank.org/. (Accessed: 15th February 2017)

51. MacWhinney, B. From CHILDES to TalkBank. in Research in Child Language Acquisition (eds. MacWhinney, B., Almgren, M., Barreña, A., Ezeizaberrena, M. & Idiazabal, I.) (Cascadilla, 2001).

52. Frank, M. C., Braginsky, M., Yurovsky, D. & Marchman, V. A. Wordbank: an open repository for developmental vocabulary data. J. Child Lang. 44, 677–694 (2017).

53. Braginsky, M. Wordbank: An open database of children's vocabulary development. Available at: http://wordbank.stanford.edu/. (Accessed: 13th February 2018)

54. Gorgolewski, K., Esteban, O., Schaefer, G., Wandell, B. & Poldrack, R. OpenNeuro—a free online platform for sharing and analysis of neuroimaging data. Organization for Human Brain Mapping. Vancouver, Canada 1677 (2017).

55. Databrary advisory board members. Databrary Available at: https://databrary.org/community/board.html. (Accessed: 16th February 2017)

PROJECT FACILITIES

PSU Office Space

PI Gilmore has office and laboratory space provided by his home department (Psychology).

ICS-ACI Resources

Location: The Institute for CyberScience (ICS) Advanced CyberInfrastructure (ACI) central facility is located on Penn State's University Park Campus and resides in the Data Center located in the Computer Building. The facility serves as the main infrastructure for the Penn State Research Community. The system consists of 37 racks of equipment located in the two main areas of the building.

Computing resources: ICS-ACI operates 16,000 standard- and high-memory cores to support Penn State research. The most recent procurement provides 15 Dell M1000E Blade server enclosures with 240 M620E Blades, 27 Dell R920 servers, 21 Dell R720 servers, 2 Dell R720XD and 6 Dell R620 servers. This equates to 6,000 of the 16,000 cores.

Operating system: The computing environment is operated by Red Hat Enterprise Linux 6.

Power: Nine UPS systems serve the needs of the facility. The UPS systems are allocated to the spaces served based on redundancy requirements. A total of 2,223 KW of UPS exist within the Computer Building facility. Two sets of 360 KW UPS systems are paired as 360 KW redundant systems for the general compute area for a total of 720 KW redundancy. The co-location area has two 144 KW UPS systems to provide a total of 144 KW redundant power. The ICS-ACI area has two 202.5 KW UPS systems utilizing single path distribution, and one 90 KW UPS system utilizing single path distribution.

Storage: ICS-ACI provides over 2.5 PB of network-attached storage (NAS) to support users' Home, Work, and Group storage needs and 2.5 PB of general parallel file system (GPFS) for Scratch storage in addition to 4 PB of tape storage for backup purposes. The NAS storage system has multiple Nexenta NAS pools which service users' Home, Work, and Group storage. The system is deployed on Dell Nexenta reference architecture. The GPFS storage pool is provided on an IBM Parallel Scratch storage system, and the Tape Backup system is an IBM TS3500 Tape system.

Network: The ICS-ACI system has a high-performance Network Fabric built on Brocade VCS fabric technology. It consists of two Brocade 8770 (8 core) switches with 10/40/100GbE network links, two Brocade 8770 (8 core) aggregation switches, and an 80 GbE VLAG between the core and aggregation switches with Mellanox FDR InfiniBand for network connectivity. The IBM GPFS is connected via RDMA.

The entire architecture employs a research-centric software stack available to users. The customized environments allow researchers to deploy software from pre-compiled and tested software catalogs. The software stack supports both commercial and open source software.

PROJECT FACILITIES

The Institute of Human Development and Social Change (IHDSC), a multidisciplinary research institute at New York University, offers a range of administrative, research, and facilities resources. These include intellectual support from a disciplinarily diverse network of over 70 faculty affiliates, administrative support in grants management, communications support in drafting and disseminating findings, and facilities (e.g., offices for senior research personnel, offices or cubicle workstations for postdoctoral associates and graduate/undergraduate research assistants). IHDSC specializes in grants management support, including tracking expenses, hiring and retaining research personnel, and coordinating with NYU's central offices for sponsored programs and financial administration to ensure that research activities take place in compliance with regulatory requirements. The Institute currently manages 40 active grants totaling over $50 million.

NYU Server Facilities

NYU will host the Databrary servers and disk arrays in the South Data Center. South Data Center: NYU's newest data center in downtown Manhattan was designed to accommodate research computing (i.e. HPC-high performance computing), as well as administrative computing equipment. The entire data center has been designed with N+1 capability, so redundancy was planned for power distribution, network, and cooling, consistent with components of the Uptime Institutes Tier 3 standards.

Size: 9,000 square feet of raised floor for Information Technology (IT) equipment, with over 200 racks and cabinets.

Power: 1.2 megawatts of electrical load for IT equipment. Two UPS systems are used to deliver clean power and to back up systems in case of an electrical outage. The battery backup maintains power until shut down or failover to generators. N+1 generator backup capability was completed in Summer 2011. There are two separate Con Edison (public utility) electrical feeds to the data center.

Power Density and Floor Strength: Density of equipment for research and administration used to be quite different, but with the advent of blade server technology (an NYU Standard), the densities are becoming more similar. For this facility, portions of the floor were reinforced to support high-density equipment. Due to constraints in this pre-WWII building we were unable to create a uniform data center for high-density equipment.

Cooling: 600 tons of cooling provided by 2 cooling towers, pumps, heat exchangers, and 30 Computer Room Air Conditioners (CRACs). A hot aisle/cold aisle design was used, and CRACs were placed in the hot aisle to efficiently pull hot air out of the facility. The cooling towers are redundant, and the external air handler allows us to take advantage of cooler ambient air for part of the year.

Networking: Over 375,000 feet of cable, 500 patch panels, 3,000 fiber strands and modules, 2,000 ft. of ladder rack, 500 ft. of cable basket and 1,000 ft. fiber raceway and components throughout the data center. The South Data Center network connects into NYUNET using optical technology, forming a large Manhattan optical ring. NYU connects to NREN networks such as Internet2 and National Lambda Rail, and to the commodity Internet at a "MeetMe" location in the building.

Network and SAN distribution: The MDF is the main connection point for all network and telecom services in the data center and houses Layer 2/3 networking switch equipment. From

the MDF various types of data/telecom backbone and horizontal cables are interconnected to the intermediate distribution frame (IDF) cabinets and onto various server cabinets and IT equipment. A separate SAN distribution frame (SDF) serves the Storage Area Network (SAN) equipment to allow SAN network connections through IDF zone cabinets onto the SAN equipment. Redundant and diverse routes of structured cabling run throughout the space.

Command Center: The Command Center is staffed 24×365. The Building Management System allows the Command Center staff to monitor all power, cooling, network, and security for the facility.

Tape Backups: NYU has an IBM 3584 tape robotic library consisting of 11 frames, 24 IBM TS1130 tape drives, and 2 robots with dual tape grippers. The cartridges each hold ~1 TB of uncompressed data (~2TB compressed), and roughly 3400 cartridges can be stored in the library. The tape backup system/software that is used to store the backup data, is IBM's Tivoli Storage Manager (TSM).

NYU Office Space

PI Adolph has office and laboratory space provided by her home department at NYU. The Databrary staff and PLAY staff have office space provided by the IHDSC.

DATA MANAGEMENT PLAN

1. Types of data produced

The project will collect video and audio recordings of behavior, questionnaire data, and computer based displays and text-based tags of those recordings. These data will be in the form of video and audio files; information and metadata about the recordings in PDF, spreadsheet, word processing, image, and text files (TXT and CSV); and coding files containing annotations in the CHAT or Datavyu file format.

2. Data and metadata standards

Databrary allows video, audio, text, and coding files to be contributed in a variety of formats, as provided by the users who create these data. We will transcode all deposited video and audio data into a standardized format (currently H.264 video codec, AAC audio codec in an MPEG-4 container for video). Access copies of these videos will be provided over the web via the native HTML5 video element. Data from Datavyu software will be exported both in original file formats and converted to open standards such as XML and CSV.

3. Policies for access and sharing

Data will be viewable and downloadable from Databrary only by *authorized investigators* who have been granted password-protected access. Researchers who wish to have access to the data must formally apply. Applicants agree to uphold Databrary's ethical principles and to follow accepted practices concerning the responsible use of sensitive data. All researchers must demonstrate that they are employed by an institution with an Institutional (Human Subjects) Review Board or similar entity. An official from an authorized investigator's institution must sign the Databrary Access Agreement. Full privileges will be granted only to those applicants with independent researcher status at their institutions. Others may be granted privileges if they are affiliated with a researcher who agrees to sponsor their application and to supervise their use of Databrary.

Ethics board or IRB approval is not required by Databrary for non-research uses. IRB approval *is required* to contribute data and for research uses. After they are authorized, users have full access to shared data on the site, and may browse, tag, download for later viewing, and conduct non- or pre-research activities. These policies are spelled out fully in an online user guide.

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

To support contributors in creating research data that may be easily shared, Databrary has extended the principle of informed consent to participate in research to include the act of sharing these data with other researchers. To formalize the process of acquiring these permissions, Databrary developed a *Participant Release Template* with standard language recommended for use with study participants. This language helps participants to understand

what is involved in sharing video data, with whom the data will be shared, and the potential risks of releasing video and other identifiable data to other researchers.

Participants choose from four different levels of release: Private, Authorized Users, Learning Audiences, or Public. ***Private*** implies that identifiable data may be uploaded to Databrary, but shared only with people selected by the data owner, usually members of a research protocol. Data that are missing a release level (e.g., participant wasn't asked, permission level was lost) are treated by default as ***Private***. At the ***Authorized User*** level data may be shared with other authorized investigators on Databrary. ***Learning Audiences*** means that, in addition to sharing with other authorized investigators, photographic images or short audio or video clips may be shown by authorized investigators in public settings for educational or research purposes. ***Public*** data are available to anyone. Databrary automatically makes de-identified data about individual participants and metadata about datasets available to the public when a dataset is shared.

In the event of a breach of data security, the NYU IRB and the IRB at the institution where the breach occurred will be notified.

4. Policies for reuse and redistribution

Data will be made available for educational and research purposes. Access will be provided using the web-based Databrary application whose software is free and open source. Materials generated under the project will be disseminated in accordance with the policies of NSF and participating institutions. Publication of data shared on Databrary by users shall occur during the project, if appropriate, or at the end of the project, consistent with normal scientific practices. Users are provided the tools to cite data sources hosted on Databrary using an automatically-generated persistent identifier. No data may be redistributed outside the principles of the Databrary Access Agreement.

5. Plans for archiving and preservation

Data in Databrary will be preserved indefinitely in a secure data center facility (see Facilities Description) at NYU and mirrored on a server in upstate New York. These facilities are administratively managed by the Information Technology Services (ITS) group, the university's central IT organization. Central IT staff at both sites handle storage, network, and backup systems. Should the current file format for Databrary access copies become obsolete, Databrary would seek guidance and support from the NYU Libraries and ITS staff prior to converting formats.

TECHNICAL PLAN

Expertise

Technical aspects of the project will be overseen by Databrary's Managing Director, Ahmad Arshad, with the support of a Back-end Developer and Front-end Developer.

Managing Director

Mr. Arshad, Databrary's Managing Director, will support project leadership on the technical development of the project, including assistance with the hiring of the development team. Mr. Arshad has a Bachelor's degree in Computer Science and a Master's degree in Management and Systems focused on database technologies, both from NYU. He has 20+ years of experience running complex technical projects and services. He brings a wealth of knowledge in the domains of systems architecture, design, integration and administration, software engineering, and data management. Mr. Arshad will devote 20% of his time to the project. He will provide support in hiring, training, and supervising the team, and will coordinate daily software development tasks, and will assist with the establishment of best practices in DevOps.

Back-end Developer (TBD)

The Back-end Developer must have extensive experience designing, coding and integrating modular MVC web applications in a modern functional programming language such as Haskell. The developer must have a working knowledge of PosgreSQL database and data models, and familiarity with Python and R. The developer is expected to embrace best practices in collaborative software development using test-driven development, design patterns, versioning using git and standard UNIX development tools. An appreciation of security and ethical concerns related to the handling of sensitive data are also important.

Front-end Developer (TBD)

The Front-end Developer must have extensive experience with JavaScript (CoffeeScript), HTML5 (audio/video API), and CSS3 (Stylus), practice with AngularJS or other modern client-side MVC framework, familiarity with git and standard UNIX development tools, and an understanding of security and ethical concerns related to sensitive data.

Databrary Architecture

Databrary is an open-source web application built in Haskell using wai/warp. The back-end is a PostgreSQL relational database. Apache Solr supports search. The user interface is built primarily on the AngularJS JavaScript framework, and all data access is performed through a JSON API.

Databrary stores at least two versions of each item of Databrary video content—a copy for access and the received or originally digitized file. Currently, the access version format is H.264 with AAC audio in an MPEG-4 container, although we expect the appropriate video formats to change over time, as has been the case with many recent digital video formats. The system uses NYU's High Performance Computing (HPC) cluster to transcode videos upon ingest using ffmpeg. Databrary uses NYU central IT to store files in two mirrored and geographically distributed locations and a third copy on offsite tape.

Communication with Target Community

Community engagement has been a critical focus of the Databrary project from the beginning. New Databrary features are announced via the Databrary and Datavyu mailing lists. We also gather feedback from users through electronic surveys to the same audiences. Project staff will continue current practices by providing email, phone, webinar, and in-person support to PLAY project researchers. Project Advisory Board members will assist us in outreach to the cognitive neuroscience, linguistics, animal communication, demography, data science, political science communities, and computer science communities.

Schedule and Timeline

Project 1: Develop LabNanny. We expect that forking an existing open source SPM system and adapting it to the needs of PLAY (Project 1.1) will take 3-5 months. Linking LabNanny's task management system to PLAY related project files (Project 1.2) will take 3-5 months, and implementing reporting, tracking, and monitoring components will also take 3-5 months. Project 1 will take a full year of effort by both the Databrary Front-end and Back-end developers.

Project 2: Local sync to/from Databrary. The Back-end and Front-end Developers will transition to work on Project 2 in the 4th quarter of Project Year 1. We project full implementation of the file sync capability to take 4-6 months.

Project 3: Enhanced search, filtering, and dataset cloning. These features are essential for Year 4 of PLAY when the Launch Group gains full access to the dataset. We expect 4.1 to take 4-6 months, beginning in the second quarter of Project Year 2. Project 4.2 should also take 4-6 months, beginning at the start of Project Year 3. Both projects involve work by the Front-end and Back-end Developers.

Project 4: Electronic protocols and coding manuals. PLAY has an existing workable wiki solution, so this project will be deferred until the last quarter of Year 3 of the current proposal. It is expected to involve 3-4 months' work by the Front-end and Back-end Developers.

Project 5: Integration with other data science/data sharing services. The Back-end developer will take the lead on Projects 5.1 and 5.2, with help from the Scientific Support Specialist and Managing Director in the areas of documentation and outreach. Project 5.1 should take no more than 3 months, beginning in the middle of Project Year 2. Project 5.2 should take 3-4 months, beginning in late Project Year 2. Planning for integration with WordBank, TalkBank, and OpenNeuro (Project 5.3) will begin early in Year 1, with implementation details the focus of activity late in Project Year 2 and early in Project Year 3. We expect these initiatives to take 4-6 months.

Annual Project Advisory Board meetings will be scheduled to coincide with the Databrary Advisory Board meetings in May or June.

SUSTAINABILITY PLAN

The first Databrary grant was funded by NSF in 2012; current NIH funding runs through the spring of 2018, but we expect to receive a no-cost extension through early 2019. NYU libraries committed to preserving the data stored on Databrary indefinitely beyond the end of grant-related funding for the project. We note that TalkBank, one of the most successful data repositories in the behavioral sciences, has been funded by competitive NIH and NSF grants for more than 30 years. ICPSR and the Center for Open Science have similar, grant-dependent funding models. Accordingly, we see that continued grant-seeking remains the most viable and promising means of sustaining and building the Databrary library over the short to medium term.

The current proposal, if funded, would support Databrary and Datavyu through the fall of 2021. The PLAY Project for which the current proposal provides infrastructure support, will run through the summer of 2023. The Sloan Foundation provided support for Databrary and Datavyu through December 2018, and we expect to approach Sloan again. We are in active discussion with officials from the Gates Foundation, the LEGO Foundation, and the McDonnell Foundation, and we are cultivating relationships with other foundation funders. We have other ideas about enhancements to Databrary and Datavyu that we will target to NSF, NIH, and other federal agencies. The PIs have successful track records of seeking NSF and NIH funding, so we are optimistic about these prospects, even in the current funding climate.

In the long term, we are working with NYU development staff to seek a private endowment fund for Databrary that would ensure resources for storage, maintenance, and development staff for an indefinite period. We estimate that Databrary could be made fully self-supporting with an annual budget in the range of $600,000-750,000. We are also discussing core administrative support for one or more technical staff lines to reduce the reliance on external grant seeking.

We continue to work with entities such as ICSPR, TalkBank, and others to advocate for long-term, stable funding sources for data repositories. Databrary is very early in its development, so we think that charging institutional or researcher-specific subscription fees is premature, but it is a possibility in the medium term. We note that the ArXiv PDF preprint repository has a successful model that mixes institutional contributions, foundation grants, and core support from the host institution. For example, a $1,000/year institutional subscription could yield $380,000 year. We will also explore contributions from professional and scientific organizations whose members benefit from Databrary's repository services. For example, a fee of $25/year per member could generate $125,000 from the estimated 5,000 members of the developmental science community (ICIS and SRCD memberships) alone. Finally, we will explore ways for researchers with federally funded research grants to include Databrary storage and technical support fees in their grant budgets based on the projected costs of storage and on staff/support costs needed to maintain the library. For example, a researcher who collects 5 hours of HD video per week over a 3-year NSF grant could generate 7.8 TB of video data. Databrary's current internal cost for storage alone is $450/TB/year. So, a grant of this size could be charged $3,510 for the cost of storage plus an additional amount for the staff and support costs. NSF budgets are always tight, but a $4-5,000 fee for storage across an entire award period seems feasible. Finally, it is possible that customized user support for LabNanny may be offered as subscription service.

We do not think that advertising is appropriate to our mission, but we will continue to monitor the changing landscape of data repository funding, and make adjustments accordingly.