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-coding tools reveal the layered and multifaceted structure of behavior. Video describes research procedures and essential properties of computer-based tasks in ways text-based descriptions or static images cannot. 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. 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 a robust, transparent, reproducible, integrative, interdisciplinary and insight-generating research ecosystem for behavioral discovery centered on video.

The proposed ecosystem builds on Datavyu.org, a desktop video coding tool, and Databrary.org, a video library the PIs developed and maintain with NSF and NICHD support. This project has three aims. (1) Empower researchers across the behavioral sciences to exploit video through a reinvigorated Datavyu that synchronizes with Databrary. (2) Improve the interoperability, reproducibility, and transparency of video-based research by making it possible to visualize, manipulate, and build on video codes shared on Databrary, share videos of computer-based tasks, and create video-based electronic protocols and coding manuals. (3) Accelerate the reuse of shared video through enhanced features for searching and filtering. For the first time, researchers will be able to find and reuse video segments meeting specific criteria or matching particular codes and to create custom video collections. These features will make Databrary's videos, codes, and metadata more accessible and attractive to data and computer scientists.

Intellectual Merit:

We will create a research ecosystem for discovery centered on video. This ecosystem will transform the behavioral sciences by accelerating the pace of discovery, expanding the breadth and depth of research, and facilitating transparency and reproducibility. It will integrate Datavyu, a widely used video analytic tool, with Databrary, a successful video data-sharing repository. Together, these flexible, innovative - and affordable - tools constitute a powerful platform for discovery. The platform 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, experimental biology, and human-computer interaction. 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 by providing a home for task displays. 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 students from underrepresented groups. The project will increase the quantity and quality of shared video datasets and associated materials. This 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 sciences.

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)	1	
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)**	15	
References Cited	6	
Biographical Sketches (Not to exceed 2 pages each)	6	
Budget (Plus up to 3 pages of budget justification)	6	
Current and Pending Support	4	
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.

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	1	
Facilities, Equipment and Other Resources	1	
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

Behavioral science is at a crossroads. Despite a powerful, growing array of tools to assess behavior, the pace of discovery has slowed. Researchers accumulate a series of findings about how participants behave in this context, that computer task, or that laboratory manipulation, but struggle to assemble the disparate pieces into a cohesive whole. Most behavioral measures reduce the richness and complexity of real-life behavior into small, manageable, but overly simplistic outcome variables. Moreover, research practices that undermine reproducibility [1–4], a literature full of underpowered studies [5–7], and a culture slow to embrace open data sharing have driven behavioral science to a crisis of credibility.

We aim to tackle this crisis head-on by fashioning a new, robust, transparent, reproducible, integrative, interdisciplinary and insight-generating research ecosystem for behavioral discovery. This ecosystem builds on an inexpensive, easy-to-use, readily available, and *uniquely powerful tool*—*video recording*. Video faithfully captures the nuances of behavior in real time and documents changes in behavior across contexts and development. Video records the details of research procedures in ways text-based descriptions, photographs, or schematic illustrations in the methods sections of journal articles cannot [8,9]. 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 developmental and learning sciences [10–18] collect video in home, lab, classroom, museum, and public settings, but few researchers in any field fully exploit the potential of video for scientific discovery. We propose to bring the power of open video data collection, annotation, and sharing to fields across the behavioral sciences.

Our proposed ecosystem builds on existing practices and tools, making our innovative and ambitious vision also feasible and pragmatic. Databrary (databrary.org) is a web-based video library created with support from NSF and NICHD to foster widespread sharing and reuse of research videos, displays, coding files, and metadata among developmental and learning scientists. Launched in 2014, Databrary now houses 7,000+ hours of video from ~750 researchers at ~300 institutions across the globe. Datavyu (datavyu.org) is a free, open source, desktop video-coding tool created and supported by Databrary. It has the largest community of users among comparable analytic tools [19,20]. Datavyu and Databrary provide the foundation of the proposed research ecosystem. By integrating Datavyu's analytic power with Databrary and by transforming Databrary from a passive data repository into a tool for active discovery, we will create a powerful platform for discovery. This will deepen and diversify the analyses researchers can conduct about human behavior across a wide range of fields.

Challenges

Large numbers of researchers who study behavior outside of developmental science—in psychology, linguistics, anthropology, political science, behavioral economics, education and learning sciences, experimental biology, and human-computer interaction—could readily and cheaply enrich their observations with video, but most do not. Among those researchers who collect video, few openly share with other researchers [21]. Similarly, researchers who measure human behavior in controlled tasks using computer-based visual displays or audio recordings could collect and share video-recorded event sequences—metadata essential for interpreting findings and making methods transparent. But few share recordings of stimuli or displays. This failure to exploit the power of video has slowed the pace of discovery in behavioral science and has probably contributed to controversies about replication [22,23].

To make full use of video, researchers must annotate it. This requires painstaking, time-consuming work by trained observers to extract information about what participants did or said and when or how. Researchers document these observations using paper-and-pencil, makeshift

spreadsheets, and free [24–26] or commercial [27–29] coding software. Software tools enable observers to control video playback direction and speed and to apply *codes*—speech transcripts, freeform comments, qualitative annotations, numeric ratings, and categorical indicators—to user-selected time-locked parts of the video. The tools store data in *coding files* that form the basis of subsequent quantitative and qualitative analyses. To guide coder training and bolster inter-observer reliability, most researchers record text-based code definitions in *coding manuals.*

Unfortunately, few researchers convert paper-and-pencil coding forms into shareable electronic files or save spreadsheets in interoperable file formats. Coding tools often store data in incompatible formats. And coding manuals lack standard formats or links to video exemplars that could make code definitions visually concrete. Researchers collect valuable metadata about individual participant characteristics (age, gender, race, ethnicity, etc.), the tasks performed, and the context. But most do so in *ad hoc* ways that involve error-prone manual data entry and in formats unsuitable for sharing. Moreover, far too many researchers who collect and code video, and who produce detailed coding manuals, and who carefully capture metadata do not openly share these priceless materials. This needlessly undermines the reproducibility, reliability, transparency, and robustness of video-based research. It also wastes time and money.

Researchers in computer vision and natural language processing also rely on human observers. Human coders provide "ground truth" for training data used to validate the accuracy of machine-extracted image tags or speech transcripts. Unfortunately, there is minimal interaction between behavioral scientists and computer and data scientists who share a common interest in coded video. Computer and data scientists have no access to the painstakingly annotated videos produced by behavioral scientists, and behavioral scientists have no systematic way to exploit advances in computer vision and language algorithms to make annotating video recordings more time- and cost-efficient.

We argue that videos of participants performing the full range of tasks studied in behavioral research, accompanied by expertly-applied codes, detailed coding manuals, and rich participant metadata constitute a substantial, largely untapped resource for new discovery in the social, behavioral, and economic sciences *and* in computer and data sciences—*if the materials can be widely and openly shared in reusable, easily discoverable formats*. Therefore, making widely shared, expertly annotated video the core of an ecosystem for behavioral discovery has significant potential to transform research and accelerate discovery across a wide range of fields.

Opportunity

Researchers and journals have begun to embrace new tools and practices that promise to bolster the reproducibility and reliability of scientific research [2,22,30]. The unparalleled role of video in documenting procedures and demonstrating phenomena has begun to gain wider recognition [8,9,21]. Researchers are adopting open science tools such as Databrary, Dataverse, and the Open Science Framework (OSF). Increasing numbers of researchers use Datavyu following a series of NSF-sponsored workshops held by the PIs over the past several years. In preliminary work, the PIs demonstrated the feasibility and utility of electronic protocols and coding manuals that contain video clips to illustrate procedures and demonstrate coded behaviors [31]. Databrary's collection of video data sets (280+) [32,33], study displays (50+ datasets), and annotated video segments (1,600+) grows daily.

Researchers hold a common misconception that identifiable video cannot be shared. Databrary has proved this misconception wrong by developing and implementing a practical policy framework for sharing identifiable data. The framework builds on the principle of informed consent, bolstered by a formal institutional agreement that restricts access to authorized

researchers. Hundreds of researchers have gained authorization and adopted Databrary's policy framework; dozens have secured IRB/ethics board approval to store videos using these policies and have uploaded videos in preparation for sharing.

In short, behavioral scientists are eager to embrace more powerful, open, and reproducible video-centered research practices. The time is ripe to make video the focal point for discovery in the behavioral sciences and to create the infrastructure to enable it.

Project Aims

Aim 1: Empower researchers to exploit the richness of video with a reinvigorated state-of-the-art desktop analysis tool—Datavyu—that synchronizes with Databrary

Just as microscopes and telescopes reveal aspects of natural phenomena invisible to the naked eye, video-coding tools allow researchers to unpack the dense, layered, and multifaceted structure of behavior to make it "as tangible as tissue" [34]. Datavyu is the leading desktop video-coding tool used by developmental scientists [19]. It is free, open source, flexible, and scriptable; its interface is optimized to reduce time- and energy-wasting keystrokes and mouse movements and to minimize coders' shifts of attention. We will *upgrade Datavyu to communicate with Databrary*—by synchronizing information about participants, sessions, projects, and videos linked with coding files. Datavyu 2.0 will also *link to electronic coding manuals* hosted on Databrary (see Aim 2) that contain text-based code descriptions and links to specific video exemplars. This will provide human observers an unprecedented ability to encode detailed aspects of behavior with high fidelity and inter-observer reliability and to store those observations for subsequent analysis and sharing in Databrary. In addition, Datavyu is a Java-based application with an aging code base. So, we will rewrite the code in a modern language to improve stability and performance and to keep up with changes in desktop operating systems.

Aim 2: Enhance Databrary with features to improve the interoperability, reproducibility, transparency, and robustness of video-based behavioral research

Databrary allows researchers to upload and download Datavyu coding files and accompanying video files, making the system useful as a "cloud-based" backup system as part of a desktop-centered video-coding workflow. But, Databrary cannot currently extract Datavyu codes or link codes to video segments. We will expand Databrary's ability to work with Datavyu files. Import functionality will *bring the codes into Databrary and make the information they contain available for visualization, searching, and filtering*. Export functionality will allow selected Datavyu codes to be written and exported in the Datavyu format, CSV, and the CHAT transcript format widely used by language researchers.

Similarly, Databrary supports storage of coding manuals in a limited set of standard formats (.docx, PDF, text), but these do not exploit the richness of video exemplars. So, we will develop an *electronic protocol for documenting procedures and an electronic coding manual* with "wiki-like" features that integrates with Databrary and Datavyu. The electronic manuals will allow researchers to fully document their research workflows, from participant recruitment through video coding and data analysis. Researchers can enhance text-based descriptions with embeddable hyperlinks to video (or audio or image) exemplars that demonstrate procedures or illustrate codes. These enhancements will make video-intensive behavioral research more reproducible and transparent while accelerating reuse and new discovery.

Aim 3: Accelerate the reuse of shared research video

Fostering widespread video sharing serves an essential goal of making behavioral science more open and transparent. But shared video must be easy to find, filter, and repurpose to yield new

findings. We will *enhance Databrary's searching and filtering functions*. For the first time, researchers will be able to *find and reuse video segments meeting specific criteria or matching particular codes* and *create custom video collections*. For example, a user might search for all instances of infants mouthing objects, children reading storybooks, or adults performing the Stroop task. 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 a video become part of the searchable record of annotations. Enhanced searching and filtering features will allow researchers to curate custom video collections based on tags or codes that others have applied, either to whole videos or to specific segments.

To *make Databrary the home for sharing task displays, videos of experimental procedures, and de-identified flat-file data*, we will create a streamlined access model for users who do not want access to identifiable video recordings.

Finally, we will enhance and fully document Databrary's application program interface (API) so data and computer scientists can mine the system's thousands of hours of human-annotated recordings through Databrary-specific R and Python packages we will develop and release. This will *make it easier for researchers in computer vision, natural language processing, and machine learning to access Databrary's resources*—to build end-to-end learning models or human-assisted annotation tools—while providing behavioral scientists new opportunities for interdisciplinary collaboration with colleagues in the these fields.

Results from prior NSF support

The project 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. Building research infrastructure and providing training and technical support were the primary focus of the prior awards. In addition, PIs Gilmore, Adolph, and Millman published several articles that describe Databrary [35–40] and how it relates to other "big data" initiatives in developmental [41] and cognitive neuroscience [42]. We developed a policy framework for sharing identifiable research data, now endorsed by ~750 researchers at ~300 authorizing institutions. We upgraded the Datavyu video-coding tool [25], held workshops to train researchers to code video, and wrote about best practices in behavioral video coding [43]. The current proposal builds on and extends these efforts. ***Intellectual Merit***. The investigators created infrastructure to enable sharing and reuse of research videos in an open-source [44] web-based repository, Databrary [45], upgraded and provided user support for the Datavyu video-coding tool [25], and fostered a rapidly growing community of researchers committed to video sharing and reuse [46]. Databrary and Datavyu deepen and accelerate the pace of discovery in developmental 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 developmental scientists, especially from institutions with limited resources, improves data management practices, and increases transparency and reproducibility in behavioral science. Datavyu brings the power of video-data coding to any laboratory with a computer. Databrary's policy framework shows that identifiable video data can be securely shared while upholding ethical principles. Our publications [35–42], workshops, and presentations are bringing this new, collaborative, integrated view of behavioral science to a larger audience.

Background & Rationale

Open data sharing has become a scientific imperative across disciplines and a mandate from research funders [47]. It is common practice in many areas of biomedical [48], physical [49], biological [50] and earth sciences [51], and an emerging priority in neuroscience [42,52]. Despite efforts to make data sharing the norm in the behavioral sciences [3,53], most research

on behavior remains shrouded in a culture of isolation [36]. Researchers share interpretations of distilled, not raw data, through publications and presentations. The path from raw data to findings to conclusions can rarely be traced or validated by others, and lack of access to files and appropriate metadata precludes other researchers from posing new questions using the same raw materials.

The growth of video as data. Developmental researchers have long recognized the power of visual media to capture the richness and complexity of children's behavior [35]. As video replaced film, it became the backbone of research programs for thousands of scientists who study learning and development: Thirty-seven percent of respondents to a recent survey of developmental scientists reported collecting at least 5 hours of video per week [19]. The rate of video use in large collaborative projects is considerably greater [54–57]. The widespread availability of low-cost, high-resolution cameras has made video a large and rapidly growing source of information about human behavior.

Video enables sophisticated behavioral science research, but poses special challenges. Video documents the interactions between people and their physical and social environment unlike any other form of measurement. It captures when, where, and how people look, gesture, move, communicate, and interact [11,34,35]. Video closely mimics the visual and auditory experiences of live human observers, so recordings collected by one person for a particular purpose may be readily understood and reused for different purposes. Furthermore, video is the emerging gold standard for documenting empirical procedures in the behavioral sciences [8,9]. Capitalizing on video's unique potential requires overcoming a unique set of challenges.

Videos contain personally identifiable information; this poses problems for the protection of participant privacy. Most videos of people contain identifiable information—faces, voices, spoken names, or interiors of homes and classrooms. Removing identifiable information from video severely diminishes its value for reuse and puts additional burdens on researchers. Therefore, video sharing requires policies that protect the privacy of research participants while preserving the integrity of raw video for reuse by others.

Large file sizes and diverse formats present technical challenges. Video files are large (one hour of HD video can consume 10 GB of storage) and come in various formats and sources (from cell phones to high-speed video). Many studies require multiple camera views to capture desired behaviors from different angles. Thus, sharing videos requires substantial storage capacity, significant computational resources, and specialized technical expertise for storing and transcoding videos into common formats that can be preserved over the long term.

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 [58]. Most researchers lack training and expertise in standard practices of data curation [37]. Few researchers reliably and reproducibly 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. This "preparing for sharing" *after the fact* presents a difficult and unrewarding chore for investigators, one that often exceeds the incremental cost and reasonable time frame contemplated under NSF's Data Sharing Policy [47]. It also makes curating datasets a challenge for repositories [37].

Extracting behavioral patterns from video involves technical and practical challenges. The rich information in video requires time-consuming work by human observers to extract. The extracted data are represented in ad hoc ways not easily exportable to other tools or statistical analysis software. In principle, researchers could build on the videos and tags generated by others. But in practice, most researchers do not share coding files, and some coding files have proprietary, incompatible data formats not easily shared outside the original research team. As a

result, the hard-won, expensive-to-acquire human insights about behavior contained in research videos remain difficult to analyze and largely hidden from the larger scientific community.

Datavyu allows researchers to extract valuable insights about behavior from video. The Datavyu coding tool was specifically designed to overcome barriers that limit the amount and quality of information researchers can extract from video while addressing limitations of existing commercial and academic tools. Datavyu builds on OpenSHAPA and MacSHAPA developed and refined by Penelope Sanderson and PI Adolph beginning in the early 1990's. Datavyu is free, open source, and written in Java to allow the same core code base to operate on both Windows and Mac OS. Datavyu's interface prioritizes the use of keystrokes over mouse movements to increase speed and decrease user fatigue, and its Ruby language API allows reproducible (scriptable) data analysis workflows, including custom and interoperable (e.g., CSV) export formats. PI Adolph has trained hundreds of researchers to use Datavyu and its predecessors.

Databrary overcomes most barriers to sharing video. Motivated by the scientific promise of video data sharing, the PIs Adolph, Gilmore, and Millman established Databrary, the first digital library for video data, experimental displays, coding files, and associated metadata with support from NSF (BCS-1238599) and NICHD (U01-HD-076595) and additional funding from the Society for Research in Child Development and the LEGO Foundation. Databrary provides a secure platform for storing and sharing videos and associated metadata among authorized researchers. It fosters data reuse and enhances scientific transparency. Databrary has targeted the developmental and learning sciences because this is the PIs' intellectual home and the focus of our current funders. But we specifically designed Databrary for broader use in any domain where video is or could be informative.

Databrary began public operation in 2014 and has since grown to 500+ authorized investigators and 230+ affiliates from ~300 institutions around the world [46,59]. These investigators have contributed more than 7,000 hours of video or audio recordings, representing ~7,000 participants ranging in age from 6 weeks to elderly adults. The system stores 280+ volumes or datasets, of which 70 are currently shared and the rest are studies in progress.

Databrary allows users to store and share data with collaborators, authorized Databrary users, or the public, depending on the level of sharing permission granted by participants. Users may search for, browse, view, and download videos and accompanying coding files. They can view specific characteristics of videos such as participants' ages or recording context (e.g., home, lab, or school). Databrary empowers users to create, view, or download video excerpts that can be shown for educational or research purposes. Thus, Databrary supports sharing, reanalysis, and pre- or non-research/educational uses of video while solving some of the thorniest problems associated with sharing data that contain personally identifiable information.

Databrary's policies enable the secure and ethical sharing of identifiable data. Sharing identifiable research video requires policies to protect participants' privacy. Databrary does not de-identify videos. Instead, Databrary maximizes the potential for reuse by keeping recordings unaltered. To share unaltered videos, Databrary shares identifiable data only with the explicit permission of the participants and restricts access to researchers who secure formal authorization from their institutions [60]. Databrary created template language [61] for seeking participants' permission to share data that researchers may adapt for their own use, and provides video examples of how to obtain permission. An online user guide describes Databrary's policies [62].

Unique among data repositories, the Databrary Access Agreement [60] authorizes both data use and contribution. However, users agree to contribute only identifiable data for which they have ethics board or IRB approval. Data may be stored on Databrary for the contributing

researcher's use regardless of whether it is shared with others. When a researcher chooses to share data beyond a research team, Databrary makes it available to authorized researchers in accord with participants' sharing permissions.

Databrary overcomes technical barriers to video data sharing. To address the problem of diverse video formats, Databrary automatically transcodes each recording into a common format suitable for web-based streaming (currently H.264+AAC in MP4 for video) using NYU's high performance computing services. The system maintains a copy in the original format for long-term preservation. Databrary does not currently place limits on the number or size of files that can be uploaded. Databrary is fully compatible with modern web-browsers, and does not require special software for access. Databrary's current assets total 31.8 TB. These are stored on NYU's central IT storage, which provides one off-site mirror and long-term tape backups.

Databrary overcomes practical barriers to sharing. Databrary has developed a novel *active-curation* framework that eliminates the burden of *post hoc* data sharing [37]. The system empowers researchers to upload and organize data as it is collected. Immediate uploading reduces the workload on investigators, minimizes the risk of data loss and corruption, and accelerates the speed with which materials become available. Databrary employs familiar, easy-to-use spreadsheet and timeline interfaces that allow users to upload videos, add metadata about tasks, settings, and participants, link related coding files and manuals, and assign appropriate permission levels for sharing. To encourage immediate uploading, Databrary provides a complete set of controls so that researchers can restrict access to their own labs or to other users of their choosing prior to sharing. Datasets can be shared with the broader research community at a later point when data collection and ancillary materials are complete, whenever the contributor is comfortable sharing, or when journals or funders require it. Active-curation poses few burdens on researcher's time beyond current practices and offers significant benefits. In effect, Databrary acts as a researcher's lab file server and cloud storage, enabling web-based sharing among research teams and ensuring secure off-site backup.

Furthermore, any de-identified data associated with a dataset, including demographic and study metadata, stimuli or displays, coding manuals, and coding data, may be shared publicly, substantially broadening the availability of these materials.

Remaining barriers this project will overcome. Despite substantial advances in reducing barriers to sharing video, significant hurdles stand in the way of widespread video reuse. Converting, sharing, understanding, retrieving, and building on other researchers' coding files remain a chore. Essential definitions of behavioral tags codes are locked away. Most researchers don't share protocol or coding manuals, raw recordings of behavior, or computer-based displays, and when they do, these can neither be indexed nor easily found. Researchers in other areas of data and computer science could add to the store of information locked away in shared videos if tools for accessing Databrary's data and metadata could make access convenient and easily automated.

Thus, we will enhance Datavyu to allow information contained in coding files to be integrated with Databrary. By electronically linking codes with detailed code definitions both on Databrary and within Datavyu, we will help researchers train human coders to reliably capture subtle nuances in behavior and thereby bolster transparency and reproducibility. By capturing, storing, and indexing codes and definitions, we will help researchers to organize and manage multiple coding passes within and across studies. Moreover, once shared, the coding files, protocols, and coding manuals will provide a foundation for other researchers to build on, and by increasing the consistency of coding, we will facilitate the comparison of results across studies. To foster data reuse, we will create ways for researchers to create custom collections combining their own videos with those shared by others. Researchers can increase their sample sizes to

improve study power, and include diverse populations to improve generalizability. To attract new collaborators in the data and computer sciences, we will streamline, document, and polish Databrary's API and build new data science-friendly tools for interacting with the system. Together, these initiatives create a new ecosystem for behavioral discovery that will enable innovative, integrative, multidisciplinary, transparent, reproducible, and transformative research that is at present difficult, time consuming, and prohibitively expensive.

Implementation Plan

The implementation plan consists of three primary projects aligned with the specific aims. The sections below describe the main phases in general terms, with additional technical details, including a brief timeline, provided in the Technical Plan addendum.

Project 1: Improve the power and utility of Datavyu by synchronizing it with Databrary

Project 1.1 Add Databrary integration features to Datavyu

Researchers find preparing for sharing an onerous demand on their limited time. The need to enter metadata about participants and tasks in multiple places at multiple times adds to the burden, increases input errors, and decreases the likelihood that datasets will be shared.

Preliminary work: Databrary developed spreadsheet and timeline interfaces that make it easy for researchers to enter vital data about participants and tasks. Researchers may upload videos immediately after a testing session and Datavyu or other data files at any time. Databrary's front-end interface interacts with the backend database via a well-structured (but as yet undocumented, see Project 3.4) API, but at present users must enter data by hand. Since researchers also require these metadata for desktop-centered statistical analyses, the current workflows duplicate effort and increase the likelihood of data entry errors.

To ensure the efficient, accurate, and automated transfer of vital data and metadata, we will enable researchers to enter participant-, session-, and study-level metadata into Datavyu and have these data elements automatically synchronize with Databrary. Initially, we will support a limited number of standard participant- and session-related data elements— sharing permission level, participant birthdate, test date, gender, race, ethnicity, testing location, lab/home setting, and so on. We will add features to Datavyu that allow it to *securely connect* to a specific Databrary volume during data synchronization, and to make appropriate and secure Databrary API calls that create new sessions or modify existing ones. We will also develop and enforce specific file naming conventions that will facilitate the exchange of Datavyu coding files with Databrary. We will set up controls in Databrary to avoid data overwriting, and provide data owners with appropriate notifications about session and volume changes such as file uploads and modifications. Data entry of session- and participant-level information, especially data sharing permissions, will be substantially more accurate and error-free by improving integration and interoperability between Datavyu and Databrary. This project also lays the foundation for future API enhancements that will allow users to collect participant metadata electronically (using web- or tablet-based apps or their own spreadsheets) and synchronize it with Databrary.

Project 1.2: Rewrite Datavyu

Researchers need a powerful tool to code video at multiple time scales. For performance reasons (playback at multiple speeds, frame-level accuracy), the tool must run on desktop computers. To support a broad range of users it must be compatible with Windows and Mac OS platforms.

Preliminary work: Databrary has developed and supported Datavyu as a free, open source platform since 2012. The tool is optimized for maximum flexibility in coding video; it is easily extended to incorporate other time-based data streams and to support reproducible workflows.

But, the Java code base is aging and needs to be updated. We tested the utility of a JavaScript-based web browser tool, but found significant time inaccuracies that make it unsuitable for time-sensitive annotation. So, we have decided to create a new Datavyu 2.0 desktop tool.

To support both Windows and Mac OS, we will divide development into a platform-specific user interface component and a common cross-platform backend component. We are considering Swift (Mac OS front-end), C# (Windows front-end), and C++ (backend) coupled with open source libraries like ffmpeg for video playback. Rewriting Datavyu will ensure that the community of researchers who exploit the richness of video recordings continues to grow.

Project 2: Enhance Databrary with features to improve the interoperability, reproducibility, transparency, and robustness of video-based behavioral research

Video-based research has the potential to be among the most transparent, reproducible, and robust forms of behavioral research, but current practices and technologies limit that potential. We propose four initiatives to realize that potential by upgrading Databrary's capabilities.

Project 2.1: Import and export Datavyu and CHAT coding files

Codes and transcripts could provide a valuable foundation for new discovery if the codes were easily visualized alongside the videos they annotate. By upgrading Databrary to store Datavyu codes internally, researchers will be able to view, build on, and extend each other's codes.

Preliminary work: Databrary currently stores Datavyu files linked with the videos in a session, but the codes cannot be visualized or manipulated in any meaningful way. Databrary already supports simple temporal tagging of video segments. We will extend Databrary's data model for tags to include Datavyu codes. This will make Databrary capable of importing codes contained in Datavyu coding files, speed the accumulation of knowledge about the behaviors captured in recordings, make video data reuse more attractive to researchers, and enhance the value of sharing videos and codes with the Databrary community.

In parallel with the backend work needed to import Datavyu codes, we will implement the ability to represent speech transcripts in the CHAT format [63,64]. This format was developed for the Child Language Data Exchange System (CHILDES) Project, housed within TalkBank [57], a collection of language-related databases. TalkBank includes transcripts, audio, and video data from children and adults. Many of TalkBank's audio or video files are linked to text-based transcripts in the CHAT format [63,64]. CHAT files can be used with the Computerized Language ANalysis (CLAN) [24] suite of software tools, a recognized standard in the language community. CHAT files are especially valuable because the format encodes speech transcripts. CHAT is the dominant data format among language researchers, and the format is a leader in interoperability with other language-related analysis tools. TalkBank's founder and director, Dr. Brian MacWhinney (Carnegie Mellon University), serves on the Databrary advisory board.

Preliminary work: CHAT files are well structured; the file specification is open and known; and we have already done preliminary work with CHAT-format transcripts in collaboration with MacWhinney. PIs Adolph and Gilmore serve as consultants on the HomeBank project, an NSF RIDIR-funded data archive project [56] aimed at building a corpus of natural speech using the LENA [65] recording device, linked with CHAT-formatted transcripts, and stored among the TalkBank family of databases. Both projects have made interoperability between HomeBank and Databrary important priorities given the considerable overlap in our research communities and the potential for multidisciplinary research. MacWhinney estimates that 100+ investigators active on TalkBank use video, and many active Databrary users have previously requested support for CHAT-formatted files. In addition to meeting the needs of a substantial existing community of researchers, enabling support for CHAT on Databrary will create an opportunity to enhance existing video datasets already shared with Databrary [66] that have CHAT-formatted

transcript data available on TalkBank [67].

Project 2.2: Develop and deploy user interfaces for visualizing and manipulating coding passes

Improvements to Databrary's interface will make it easier for users to visualize what codes have been applied and by which researchers, to explore these codes in conjunction with the videos, to access and export the codes, and to upload new coding passes to their own sessions. After a user has selected a particular video of interest, Databrary will enable researchers to visualize existing coding passes linked with that video. Databrary will support basic visualization and manipulation of these passes, so that users can preview the codes, select a subset of passes to export in their preferred format, or upload updated versions.

Preliminary work: Databrary's timeline viewer allows users to visualize temporal relations among multiple phases of a data collection (e.g., parallel data streams or multiple camera views). Using a browser, users manually select segments of video and enter codes. We will augment this interface to allow users to visualize different coding passes applied to a particular session. The interface will allow users to select or deselect specific coding passes for display and further manipulation. It will support selective exporting of coding passes along with the associated videos. The existing upload and download functionality on Databrary's timeline session viewer will be expanded to allow more flexible manipulation of coding passes. Users will be able to upload a new coding pass or replace an existing coding pass on sessions they can edit from any coding file on their computer through an enhanced Ruby API and new Python and R APIs. After selecting coding passes of interest, users will be able to export those passes in a single Datavyu or CHAT file or in a CSV format that can be imported into statistical analysis software.

Projects 2.1 and 2.2 will lay the groundwork for future enhancements to Databrary that will allow it to import, visualize, and export coding files created by other video-coding tools [26-29] or from custom software used by computer vision and language researchers who wish to contribute machine-generated tags to Databrary's recordings (see Project 3.3).

Project 2.3: Add video-centered electronic protocols and coding manuals to Databrary

Detailed protocol and coding manuals are essential components of transparent, reproducible behavioral research. The manuals make concrete critical details about procedures and code definitions. Developmental researchers report that they typically create detailed coding manuals that define the behavioral codes associated with a set of analyses [19]. Most use word processing and spreadsheet software for these documents. Consequently, we know that code definitions are widely available in electronic form and that many researchers want to share them. We just need ways to capture and organize the currently disparate information in idiosyncratic formats. Moreover, text-based code definitions and protocol descriptions will become considerably more informative when the definitions are linked to illustrative video excerpts.

Preliminary work: The PIs created a wiki-based electronic protocol in planning for a large-scale video-based study of infant play [31]. The wiki-based format offers many advantages, including flexibility—it supports text, image, and video—and components such as code definitions or procedure descriptions can be linked to specific URLs. These URLs can, in turn, be linked to specific codes or coding passes. If the video clips are stored on Databrary, only authorized users who are signed into the system can view the clips. This protects participant identities, and makes it more attractive for researchers to add video clips from their own studies to illustrate procedures or coding definitions.

Project 2.3 adds electronic coding and protocol manual capabilities to Databrary. These capabilities will enable users who are viewing a coded video on Databrary or coding a video in

Datavyu to click on a code and read its definition from a coding manual associated with the study. In turn, users viewing procedures and codes will be able to read the text-based description of the procedure or code and then watch a linked video clip (stored on Databrary) demonstrating how it is carried out in practice. These features will require both backend and user interface enhancements to Databrary.

We will consult with our Advisory Board and other experts about whether to adapt an existing wiki engine and interface for our use, link to an existing resource (e.g., OSF), or build its equivalent within Databrary. With these proposed enhancements, researchers will be able to view code definitions across shared datasets to evaluate their clarity and consistency. Researchers will be able to initiate conversations about opportunities to achieve consensus around conceptual ontologies in areas where there is currently no consensus. Moreover, all code and protocol definitions will be indexed by the Databrary search engine, and thus become searchable metadata for other researchers.

Project 2.4: Making Databrary's infrastructure scalable and robust

Databrary has shown exponential growth, more than doubling in the last year. The single-server system works well and has minimal downtime. But if the current growth of users continues, proposed large-scale video-based research studies go forward, and the features we propose prove attractive, the technical infrastructure of the system will need to be operationalized as a more robust service. We must ensure that researchers using Databrary suffer no data losses or system outages. Relying on co-I Millman's expertise and working with NYU's Central IT staff, we will redesign the system architecture to include both data and application redundancies to include high availability, load balancing, and disaster recovery for both the application and data.

Project 3: Accelerating video data reuse

The essence of the ecosystem we envision is to create the foundation for new discoveries by making it possible for new researchers to capitalize on the videos, codes, and metadata collected by others. Project 3 consists of three initiatives to accelerate the reuse of video data shared on Databrary.

Project 3.1 Enhance searching and filtering

For researchers to reuse shared data, they must be able to find and select data based on the specific questions they wish to answer. Databrary is among a handful of data repositories that support powerful searching and filtering operations. Searching and filtering are essential to transform a repository from a passive storage facility into an active tool for new discovery.

Preliminary work. Databrary's existing search engine allows researchers to search for terms linked to shared datasets. It allows researchers to filter data based on participant demographic characteristics stored in the Databrary spreadsheet. These powerful capabilities make it possible to search Databrary for datasets that include specific participant ages or other characteristics and to preview video excerpts when they are available.

We aim to take Databrary's searching and filtering capabilities to the next level by empowering researchers to search for *datasets* that coded particular behaviors or involved specific tasks and *video segments tagged with specific codes*. We will build a search interface that returns information about which datasets (or videos) meet specified search requirements—characteristics of the participant (e.g., race/ethnicity, gender, language), session (e.g., time/date, setting, geographic location), dataset (e.g., protocol or coding manuals), and video (e.g., coding files) characteristics. The interface will allow users to select and explore specific datasets, whole videos, or segments of videos returned by the search and then choose what to do with the matching items (see Projects 3.2 and 3.3).

To achieve this, we must enhance or replace Databrary's existing search engine based on Apache Solr. We will need to index and search for entire datasets (and specific segments of videos within those datasets) that match selected criteria. These searches will be based on available metadata and codes in the associated (Datavyu and CHAT) coding files. The design of the search engine poses some complex problems related to matching multiple codes that may overlap based on search terms, linked coding manuals, and other metadata. Because work on Projects 1 and 2 will inform the design of the search engine, we plan to begin the design process early in the project but will implement it based on progress on the other projects. Databrary will rely on the expertise of co-I Millman, and on Databrary Advisory Board Members, Lee Giles and Vasant Honavar, who all have experience in search, big data, and data mining.

Project 3.2 Lowering barriers to sharing displays

Many researchers in cognitive neuroscience and in developmental, cognitive, and social psychology use computer-based tasks that present participants with video display sequences of images or sounds. These displays provide essential information about the tasks participants performed regardless of whether participants' behaviors were video recorded. Some researchers share their displays as supplemental materials linked to publications or on lab websites, but most do not. This means that researchers using similar tasks cannot compare their displays with those of their colleagues, nor can displays used in one study be reused in another.

Preliminary work. Databrary regularly issues calls for researchers to share task displays for others to use in research, talks, and teaching, and as a result already serves as a repository for displays used in developmental science (e.g., [68–75]). A popular and regularly downloaded Databrary volume is the Child Affective Facial Expression (CAFE) stimulus set [76] (200+ downloads). Databrary aims to become the primary home for sharing these sorts of materials across the behavioral sciences.

To lower barriers to sharing displays, we will modify our policies to create a new type of account that requires no formal institutional agreement because access to identifiable data will be prohibited. This will allow researchers to publicly share displays and other de-identified materials while exploiting Databrary's user-friendly interface, searching and filtering operations, video preview, and electronic protocol/coding manuals. Brian Nosek from the Center for Open Science, host of the Open Science Framework (OSF), serves on Databrary's Advisory Board. We will work with OSF to explore how best to link Databrary's display storage feature to leverage the large number of behavioral researchers outside of developmental psychology currently using OSF. We will also consult with Advisory Board member Russell Poldrack on ways to link the Databrary stimulus display library with his Cognitive Atlas database [77] to provide key metadata about task names and types that can facilitate search and discovery of displays. We will also discuss ways to link displays stored on Databrary with other cognitive neuroscience tools that Poldrack's Center for Reproducible Neuroscience is developing.

Project 3.3: Creating new collections of shared videos

Because video captures so many varied dimensions of behavior, it can be reused to answer new questions beyond the focus of the original study [78]. Finding shared videos that meet a researcher's needs is one problem, creating aggregated collections that can be reused for new purposes is another. Project 3.2 solves both of these problems.

Imagine a researcher interested in studying if and how children's speech varies depending on whether they are walking or not. Using enhanced search facilities described in Project 3.1, the researcher can find a dozen studies with several hundred videos that meet her criteria—videos of children 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 currently difficult, time-consuming, and error-prone. The original data providers have no information that their shared data were found and downloaded. Databrary has only limited information about the extent of reuse. The provenance of the painstakingly collected and coded video data is undermined and along with it the potential for new sets of codes to add layers to the accumulated information about each video.

Preliminary work: Databrary already generates persistent identifiers for shared data sets; 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 from other datasets. The virtual datasets will be stored and presented using the same volume interface Databrary now uses with the exception that individual sessions would be a set of (i) videos linked directly to the volume, plus (ii) links to sessions or specific files stored in other volumes. 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 system will automatically link to system-generated citations so that the new researchers can easily cite the materials from which their 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's volume 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 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. Those manuals will get linked back to the original sources. These enhancements naturally extend to supports for user-defined, searchable, and discoverable collections of experimental displays.

Implementing the collections feature will require modifications to the existing Databrary volume, spreadsheet, and session interfaces. The modifications will address how to distinguish the visual representation of "virtual" or linked components from those that are directly associated with a volume. We will need to enable users to save a set of found and filtered data to a new volume. We will also have to modify the Databrary backend to keep track of which volume components are linked from other sources and which are not, notify contributors when their volumes are cloned, and other components. The capacity to create a transparent, reproducible, and robust chain of knowledge about shared videos and to empower researchers to easily build on one another's findings and transparently share discoveries has significant capacity to enhance the scholarly value of shared data and accelerate reuse and discovery.

Project 3.4: Access for data mining, computer vision, and natural language processing

The widespread and open sharing of videos, time-locked codes, and code definitions has considerable untapped value to other data-intensive sciences. In effect, through the course of their research, Databrary users create valuable resources for researchers in data science, machine learning, computer vision, and natural language processing. We aim to make Databrary's resources more readily available to researchers outside developmental science.

Preliminary work. PI Gilmore has applied computer vision analyses to Databrary videos as part of separate work on optic flow processing across the lifespan [78,79]. Databrary Advisory Board

members Chen Yu and James Rehg use computer vision techniques to explore basic questions in perceptual, cognitive, and affective development [81-84]. Yu is also a consultant on this grant. However lack of access to videos and codes has limited the considerable potential of computer-intensive analyses of Databrary's video and flat-file data.

In Project 3.4, we propose three activities that will accelerate the reuse and reanalysis of Databrary's materials by an even wider audience of scientists. The first activity is to document, polish, and refine Databrary's API (see also Project 1.1). This will make it easier for researchers to extract Databrary metadata and data in organized ways for subsequent reanalysis. Next, we will develop Databrary-specific software packages in the R and Python programming languages commonly used in data science to facilitate two-way interaction with Databrary's resources. Finally, we will explore ways to allow new data or metadata extracted from Databrary's videos to be added back to the volumes from which the original materials were derived. PI Gilmore will work with the technical team to test workflows that extract videos from the API, estimate optic flow fields using software his team has developed [80], then return the extracted data back to Databrary. Similarly, consultant Yu will work with the technical team to test how data from the "human hand" recognition algorithm his lab has developed [81–82] can be imported back into Databrary and shared and visualized alongside the raw videos. These demonstration efforts will pave the way for future, more formal collaborations between the behavioral sciences and the computer vision, natural language processing, and data science communities.

Coordination & Management Plan

The project will be overseen by PIs Gilmore and Adolph and co-I Millman. The PIs and co-I currently meet by phone or video conference several times per week to discuss project-related matters. The PIs/co-I also 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, as well as external advisors who bring expertise in data sharing and developmental science [85]. The Databrary Advisory Board meets annually to hear project updates and provide guidance about policy and technical matters.

Evaluation and Assessment

We will evaluate progress in several ways. The PIs/co-I will report progress for each project relative to the goals outlined in the timetable (Technical Plan) in the annual project report. Among other metrics, we will report the number of Datavyu users and Databrary users and the numbers who are sharing coding files and manuals. We will develop estimates of data reuse based on download statistics and the generation of new video collections. The team will send out 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. We will survey the community of video-coding tool users to ask similar questions. The results of those surveys will be summarized in the annual NSF progress report and discussed at the annual Databrary advisory board meetings.

Summary

Video uniquely and powerfully captures human behavior and thus provides the core of an ecosystem for behavioral discovery. Putting the richness and complexity of behavior back into focus by making video recording, coding, and sharing commonplace will reinvigorate discovery in the behavioral sciences and establish these fields as leaders in research transparency and reproducibility. By enhancing Datavyu, an existing analytic tool for coding video, and integrating

it with Databrary, the only data library specifically designed for sharing video and computer-based tasks and displays, this project promises to transform how insights about behavior are captured, characterized, catalogued, communicated, shared, and built upon.

This proposal fulfills multiple criteria for the RIDIR program. With respect to *science*, the enhancements to Databrary and Datavyu will enable new, integrative, and interdisciplinary research questions about the characteristics and consequences of human behavior across ages, domains of function, and contexts, and allow these questions to be answered with unprecedented depth, breadth, and impact. The research communities interested in exploring these questions include developmental science, 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, substantially enhances the diversity of metadata provided for and linked to shared video, improves the tools for exploiting the linkages, and expands the communities that benefit. The proposal builds on new, but established infrastructure with strong institutional backing, and a trio of dedicated PIs/co-I who are committed to seeking continued grant funding for the tools. With respect to *governance*, the project builds on a large, growing, and enthusiastic core user base whose needs and interests have helped to shape the proposed enhancements to the Datavyu and Databrary tools. The PIs/co-I have established, maintain, and regularly consult with a diverse Advisory Board [85] 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— e.g., CHILDES/TalkBank, HomeBank—and its plans for future links with others—e.g., OSF, Cognitive Atlas, OpenFMRI, and WordBank.

Intellectual Merit

We will create a research ecosystem for discovery centered on video. This ecosystem will transform the behavioral sciences by accelerating the pace of discovery, expanding the breadth and depth of research, and facilitating transparency and reproducibility. It will integrate Datavyu, a widely used video analytic tool, with Databrary, a successful video data-sharing repository. Together, these flexible, innovative—and affordable—tools constitute a powerful platform for discovery. The platform 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, experimental biology, and human-computer interaction. 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 the 280+ datasets already stored or 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 by providing a home for task displays. 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 students from underrepresented groups. The project will increase the quantity and quality of shared video datasets and associated materials. This 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 sciences.

REFERENCES

1. Ioannidis JPA, Munafò MR, Fusar-Poli P, Nosek BA, David SP. Publication and other reporting biases in cognitive sciences: Detection, prevalence, and prevention. Trends Cogn Sci. 2014;18: 235–241. doi:10.1016/j.tics.2014.02.010

2. Munafò MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, Sert NP du, et al. A manifesto for reproducible science. Nature Human Behaviour. 2017;1: 0021. doi:10.1038/s41562-016-0021

3. Nosek BA, Bar-Anan Y. Scientific Utopia: I. Opening scientific communication. Psychol Inq. 2012;23: 217–243. doi:10.1080/1047840X.2012.692215

4. Nosek BA, Spies JR, Motyl M. Scientific utopia II. Restructuring incentives and practices to promote truth over publishability. Perspect Psychol Sci. 2012;7: 615–631. doi:10.1177/1745691612459058

5. Button KS, Ioannidis JPA, Mokrysz C, Nosek BA, Flint J, Robinson ESJ, et al. Power failure: Why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci. 2013;14: 365–376. doi:10.1038/nrn3475

6. Szucs D, Ioannidis JPA. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. bioRxiv. 2016; 071530. doi:10.1101/071530

7. Ioannidis JPA. Why most published research findings are false. PLoS Med. 2005;18: 40–47. doi:10.1080/09332480.2005.10722754

8. Gilmore RO, Adolph KE. Video can make science more open, transparent, robust, and reproducible [Internet]. Available: http://osf.io/3kvp7

9. Gilmore RO, Adolph KE. 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. Hall, K, Croyle, R, & Vogel, A, editors. Springer;

10. Derry SJ, Pea RD, Barron B, Engle RA, Erickson F, Goldman R, et al. Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. Journal of the Learning Sciences. 2010;19: 3–53. doi:10.1080/10508400903452884

11. Goldman R, Pea R, Barron B, Derry SJ. Video research in the learning sciences [Internet]. Routledge; 2014. Available: https://books.google.com/books?id=7HZ9AwAAQBAJ

12. Alibali MW, Nathan MJ. Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. Journal of the Learning Sciences. 2012;21: 247–286. doi:10.1080/10508406.2011.611446

13. Masats D, Dooly M. Rethinking the use of video in teacher education: A holistic approach. Teaching and Teacher Education. 2011;27: 1151–1162. doi:10.1016/j.tate.2011.04.004

14. Pasqualino C. Filming emotion: The place of video in anthropology. Vis Anthropol Rev.

2007;23: 84–91. doi:10.1525/var.2007.23.1.84

15. Video sharing, deep tagging and annotation: A scientific archiving and demonstration tool [Internet]. [cited 13 Feb 2016]. Available: http://cmdbase.org/

16. Qualitative Data Repository [Internet]. [cited 13 Feb 2016]. Available: https://qdr.syr.edu/

17. Chaquet JM, Carmona EJ, Fernández-Caballero A. A survey of video datasets for human action and activity recognition. Comput Vis Image Underst. 2013/6;117: 633–659. doi:10.1016/j.cviu.2013.01.013

18. Rautaray SS, Agrawal A. Vision based hand gesture recognition for human computer interaction: a survey. Artificial Intelligence Review. 2012;43: 1–54. doi:10.1007/s10462-012-9356-9

19. Gilmore RO, Adolph KE. Video use survey of ICIS and CDS listserv subscribers. 2012 Feb.

20. Gilmore RO, Adolph KE, Kennedy JL. Survey of Databrary and Datavyu users. Databrary; 2017.

21. Clark TD. Science, lies and video-taped experiments. Nature. 2017;542: 139. doi:10.1038/542139a

22. Collaboration OS. Estimating the reproducibility of psychological science. Science. American Association for the Advancement of Science; 2015;349: aac4716. doi:10.1126/science.aac4716

23. Gilbert DT, King G, Pettigrew S, Wilson TD. Comment on "Estimating the reproducibility of psychological science." Science. 2016;351: 1037–1037. doi:10.1126/science.aad7243

24. Using CLAN [Internet]. [cited 15 Feb 2017]. Available: http://childes.psy.cmu.edu/clan/

25. Datavyu: Video coding and data visualization tool [Internet]. [cited 15 Feb 2017]. Available: http://datavyu.org/

26. ELAN: Language archiving technology [Internet]. Available: http://www.lat-mpi.eu/tools/elan/

27. Behavior Research with Mangold INTERACT [Internet]. [cited 15 Feb 2017]. Available: https://www.mangold-international.com/en/products/software/behavior-research-with-mangold-interact

28. Noldus | Innovative solutions for behavioral research [Internet]. [cited 15 Feb 2017]. Available: http://www.noldus.com/

29. Transana: Qualitative analysis software for video and audio data [Internet]. [cited 10 Feb 2016]. Available: http://www.transana.org/

30. Nosek BA, Alter G, Banks GC, Borsboom D, Bowman SD, Breckler SJ, et al. Promoting an open research culture. Science. 2015;348: 1422–1425. doi:10.1126/science.aab2374

31. The PLAY Project Wiki [Internet]. [cited 17 Feb 2017]. Available: https://dev1.ed-projects.nyu.edu/wikis/docuwiki/doku.php

32. Baker D. Arnold Gesell's films of infant and child development [Internet]. Databrary. 2014. doi:10.17910/B7.70

33. Tamis-LeMonda C. Language, cognitive, and socio-emotional skills from 9 months until their transition to first grade in U.S. children from African-American, Dominican, Mexican, and Chinese backgrounds [Internet]. 2013. doi:10.17910/B7CC74

34. Curtis S. "Tangible as tissue": Arnold Gesell, infant behavior, and film analysis. Sci Context. 2011;24: 417–442. Available: https://www.ncbi.nlm.nih.gov/pubmed/21995223

35. Adolph KE. Video as data. APS Obs. 2016;29. Available: http://www.psychologicalscience.org/observer/video-as-data

36. Adolph KE, Gilmore RO, Freeman C, Sanderson P, Millman D. Toward open behavioral science. Psychol Inq. 2012;23: 244–247. doi:10.1080/1047840X.2012.705133

37. Gordon AS, Millman DS, Steiger L, Adolph KE, Gilmore RO. Researcher-library collaborations: Data repositories as a service for researchers. Journal of Librarianship and Scholarly Communication. 2015;3. doi:10.7710/2162-3309.1238

38. Gilmore, RO, Adolph, KE, & Millman, DS. Curating identifiable data for sharing: The Databrary project. 2016 New York Scientific Data Summit. Available: https://github.com/databrary/presentations/blob/master/nysds-2016/gilmore-adolph-millman-nysds-2016.pdf

39. Gilmore RO, Gordon A., Adolph KE, Millman DS. Transforming education research through open video data sharing. Trans Adv Eng Educ. 2016;5. Available: http://advances.asee.org/publication/transforming-education-research-through-open-video-data-sharing/

40. Gordon, AS, Steiger, L, & Adolph, KE. Losing research data due to lack of curation and preservation. In: Johnston L, editor. Curating research data Volume 2: A handbook of current practice. Chicago, IL: Association of College and Research Libraries;

41. Gilmore RO. From big data to deep insight in developmental science. Wiley Interdisciplinary Reviews: Cognitive Science. 2016; 15. doi:0.1002/wcs.1389

42. Gilmore RO, Diaz MT, Wyble BA, Yarkoni T. Progress toward openness, transparency, and reproducibility in cognitive neuroscience [Internet]. OSF preprint file. Available: https://osf.io/6ybzn/

43. Best practices for coding behavioral data from video || Datavyu: Video coding and data visualization tool [Internet]. [cited 15 Feb 2017]. Available: http://datavyu.org/user-guide/best-practices.html

44. Databrary Repository on GitHub [Internet]. [cited 15 Feb 2017]. Available: https://github.com/databrary

45. Databrary [Internet]. [cited 8 Jul 2015]. Available: https://databrary.org/

46. Authorized Databrary Investigators [Internet]. [cited 10 Feb 2016]. Available: https://nyu.databrary.org/search?volume=false&f.party_authorization=4&f.party_is_institution=false

47. Dissemination and Sharing of Research Results | NSF - National Science Foundation [Internet]. [cited 15 Feb 2017]. Available: https://www.nsf.gov/bfa/dias/policy/dmp.jsp

48. Kaye J, Heeney C, Hawkins N, de Vries J, Boddington P. Data sharing in genomics -- re-shaping scientific practice. Nat Rev Genet. 2009;10: 331–335. doi:10.1038/nrg2573

49. Young JR. Crowd science reaches new heights. The Chronicle of Higher Education. 28 May 2010. Available: http://chronicle.com/article/The-Rise-of-Crowd-Science/65707/. Accessed 10 Feb 2016.

50. Reichman OJ, Jones MB, Schildhauer MP. Challenges and opportunities of open data in ecology. Science. 2011;331: 703–705. doi:10.1126/science.1197962

51. Kleiner K. Data on demand. Nat Clim Chang. 2011;1: 10–12. doi:10.1038/nclimate1057

52. Poldrack RA, Gorgolewski KJ. Making big data open: Data sharing in neuroimaging. Nat Neurosci. 2014;17: 1510–1517. doi:10.1038/nn.3818

53. AERA Code of Ethics: American Educational Research Association Approved by the AERA Council February 2011. Educ Res. SAGE Publications Sage CA: Los Angeles, CA; 2011; doi:10.3102/0013189X11410403

54. (MET) Measures of Effective Teaching Project - K-12 Education. In: K-12 Education [Internet]. [cited 15 Feb 2017]. Available: http://k12education.gatesfoundation.org/teacher-supports/teacher-development/measuring-effective-teaching/

55. Teaching and Learning Exploratory, University of Michigan [Internet]. [cited 10 Feb 2016]. Available: http://soe.mivideo.it.umich.edu/

56. HomeBank [Internet]. [cited 15 Feb 2017]. Available: http://homebank.talkbank.org/

57. TalkBank [Internet]. [cited 15 Feb 2017]. Available: http://talkbank.org/

58. Ascoli GA. The ups and downs of neuroscience shares. Neuroinformatics. 2006;4: 213–216. doi:10.1385/NI:4:3:213

59. Institutions with Authorized Databrary Investigators [Internet]. [cited 10 Feb 2016]. Available: https://nyu.databrary.org/search?offset=0&volume=false&f.party_authorization=5&f.party_is_institution=true

60. Databrary Access Agreement || Databrary: An Open Data Library for Developmental Science [Internet]. [cited 15 Feb 2017]. Available: https://databrary.org/access/policies/agreement.html

61. Participant Release Template || Databrary: An Open Data Library for Developmental Science [Internet]. [cited 15 Feb 2017]. Available: https://databrary.org/access/policies/release-template.html

62. Policies || Databrary: An Open Data Library for Developmental Science [Internet]. [cited 15 Feb 2017]. Available: https://databrary.org/access/policies.html

63. MacWhinney B. The CHILDES project: Tools for analyzing talk. 3rd Edition. Lawrence Erlbaum Associates; 2000.

64. MacWhinney B. Tools for analyzing talk Part 1: The CHAT transcription format [Internet]. Carnegie Mellon University; 2017 Feb. Available: http://talkbank.org/manuals/chat.pdf

65. LENA Research Foundation [Internet]. [cited 10 Feb 2016]. Available: https://www.lena.org/

66. Demuth K. Word-minimality, Epenthesis and coda licensing in the early acquisition of English [Internet]. Databrary. 2014. doi:10.17910/B7B885

67. Phonbank English Providence [Internet]. [cited 14 Feb 2016]. Available: http://childes.talkbank.org/browser/index.php?url=PhonBank-Phon/English-Providence/

68. DeLoache JS. Scale errors offer evidence for a perception-action dissociation early in life [Internet]. Databrary. 2014. doi:10.17910/B7H019

69. Wilkinson K. Preliminary investigation of visual attention to human figures in photographs: Potential considerations for the design of aided AAC visual scene displays [Internet]. Databrary. 2014. doi:10.17910/B7G59R

70. Needham A. Teach to reach: The effects of active vs. passive reaching experiences on action and perception [Internet]. Databrary. 2014. doi:10.17910/B77G6J

71. Frank MC. Measuring the development of social attention using free-viewing [Internet]. Databrary. 2014. doi:10.17910/B79G65

72. Horst JS. The Novel Object and Unusual Name (NOUN) Database: A collection of novel images for use in experimental research [Internet]. Databrary. 2016. doi:10.17910/B7.209

73. Motta-Mena NV, Scherf KS. Pubertal development shapes perception of complex facial expressions [Internet]. Databrary. 2016. doi:10.17910/B7.272

74. Amso D. An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes [Internet]. Databrary. 2014. doi:10.17910/B7C88G

75. Naigles L. Children use syntax to learn verb meanings [Internet]. Databrary. 2014. doi:10.17910/B7J01M

76. LoBue V. The Child Affective Facial Expression (CAFE) set [Internet]. Databrary. 2014. doi:10.17910/B7301K

77. Cognitive Atlas [Internet]. [cited 17 Feb 2017]. Available: http://www.cognitiveatlas.org/

78. Raudies F, Gilmore RO. Visual motion priors differ for infants and mothers. Neural Comput. 2014;26: 2652–2668. doi:10.1162/NECO_a_00645

79. Gilmore RO, Raudies F, Jayaraman S. What accounts for developmental shifts in optic flow sensitivity? 2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob). 2015. pp. 19–25. doi:10.1109/DEVLRN.2015.7345450

80. Gilmore RO, Raudies F. Matlab toolbox for the estimation and analysis of optic flow [Internet]. 2016. Available: https://github.com/opticflow/analysis

81. Bambach S, Lee S, Crandall DJ, Yu C. Lending a hand: Detecting hands and recognizing activities in complex egocentric interactions. Proceedings of the IEEE International Conference on Computer Vision. 2015. pp. 1949–1957. Available: http://vision.soic.indiana.edu/papers/egohands2015iccv.pdf

82. Bambach S, Lee S, Crandall D, Yu C. Lending A Hand: Detecting hands and recognizing activities in complex egocentric interactions [Internet]. [cited 20 Feb 2017]. Available: http://vision.soic.indiana.edu/projects/lending-a-hand/

83. Rehg JM, Abowd GD, Rozga A, Romero M, Clements MA, Sclaroff S, Essa I, Ousley OY, Li Y, Kim C, Rao H, Kim JC, Presti LL, Zhang J, Lantsman D, Bidwell J, Ye Z. Decoding children's social behavior. Proceedings of CVPR.

84. Li Y, Fathi A, Rehg JM. Learning to predict gaze in egocentric video. International Conference on Computer Vision; Sydney, Australia.

85. Databrary Advisory Board Members [Internet]. [cited 16 Feb 2017]. Available: https://databrary.org/community/board.html

PROJECT FACILITIES
NEW YORK UNIVERSITY

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 a 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 have office space provided by the IHDSC.

PROJECT FACILITIES
THE PENNSYLVANIA STATE UNIVERSITY

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.

DATA MANAGEMENT PLAN

1. Types of data produced

The project will collect video and audio recordings of behavior 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. Because some researchers may wish to store non-identifiable displays and not gain access to identifiable video data, we will create a new level of access to permit this. All other researchers will have to 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, Shared, Excerpts, 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*. *Shared* data may be shared with other authorized investigators on Databrary. *Excerpts* 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 the 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.

SUSTAINABILITY PLAN

The first Databrary grant was funded by NSF in 2012, and current NIH funding runs through the spring of 2018. NYU libraries have 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 through the fall of 2020. We are drafting a new grant submission for NICHD for a June 2017 deadline; we have been invited to submit a proposal to the Sloan Foundation; we are in active discussion with Gates Foundation officials, one of whom attended an NICHD-sponsored workshop we held in December 2016 for a new video-centered research initiative; and we have other ideas about enhancements to the library and research projects based on the library's holdings 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.

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 $450,000-500,000.

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 near 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. 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). Finally, we will explore ways for researchers with federally funded research grants to include Databrary storage 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.

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.

TECHNICAL PLAN

Expertise

Technical aspects of the project will be carried out by Databrary's Managing Director, Ahmad Arshad, with the support of a Datavyu Developer, 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 Bachelors degree in Computer Science as well as a Masters degree in Management and Systems concentrating on database technologies, both from NYU. He has over 19 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, will and coordinate daily software development tasks, and will assist with the establishment of best practices in DevOps.

Datavyu Developer (TBD)

The Datavyu Developer will have extensive experience in designing high performance modular desktop applications that manipulate media and data using modern Java technologies, especially JavaFX, SWING, AWT and JNI. The developer must also have familiarity with Swift, C#, or C++ and experience integrating desktop applications with web applications using APIs and web services, as well as with Ruby and Python scripting engines. Familiarity with video and audio codecs and libraries like ffmpeg and avconv will also be important.

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 the 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 Databrary and Datavyu users. Staff will demonstrate new features at training workshops, supported by NSF and SRCD, held at scientific conferences and in targeted regional locations chosen for their high concentrations of video-using researchers.

Schedule and Timeline

Project 1.0: Improve the power and utility of Datavyu. We expect re-writing Datavyu (Project 1.1) to take 12-24 months, beginning in September 2017. The Datavyu developer will take the lead. Importing metadata from Datavyu into Databrary has been allotted a 3-month time period, beginning in September 2017. This will be the responsibility of the Datavyu and Back-end Developers.

Project 2.0: Enhance Databrary. The Back-end and Front-end Developers will cooperate on Projects 2.1 and 2.2, starting in January 2018 and taking 4-6 months. Project 2.3 is expected to take 6-9 months of Back-end and Front-end developer effort, starting in summer 2018. Project 2.4 will take 3-6 months and will be the responsibility of the Managing Director, beginning in September 2017.

Project 3.0: Accelerating video data reuse. Project 3.1 is expected to take 6 months of Back-end developer effort beginning in Spring 2019. The work on Project 3.2 should take 3 months of Back-end effort, scheduled as determined by the Managing Director. Project 3.3 will take 6-9 months of Back-end and Front-end effort, beginning in late 2019. Project 3.4 is projected at 3-6 months and will also be scheduled as determined by the Managing Director.

Annual Databrary Advisory Board meetings occur in May or June.