PROJECT SUMMARY

OVERVIEW

Video is a uniquely powerful source of information about behavior in context, and it is eminently sharable and reusable. Video provides *data* about what participants did in natural or experimental tasks, *documentation* about what researchers did in their protocols, and *demonstrations* of behavioral phenomena and research findings that enlighten academic communities and the public. Video reveals details of behavior and the surrounding context in formats comparable to the audio-visual experiences of human observers and provides a durable record that allows unlimited opportunities to revisit past events, ask new questions, and make new discoveries. Researchers across the social, behavioral, and economic (SBE) sciences routinely use video as data, documentation, and demonstration. Many share it using Databrary, a "user-friendly, next generation, data resource" that was launched with NSF support a decade ago.

Databrary is a web-based, restricted access digital data repository housed at New York University. It is the *only* large-scale repository specialized for storing and sharing research video. Multi-measure data and metadata accompany and enrich most recordings. Databrary stores 177+ TB of data, documentation, and demonstrations, including 120,647+ hours of video and audio recordings. Some 1,013+ shared datasets—dozens supported by NSF—represent 10,725+ diverse human participants and have garnered 981+ citations to date. A growing global user community of 1,783+ institutionally authorized researchers (and 640+ staff and trainees) from 796+ institutions worldwide have contributed data permissioned by research participants for broad, unspecified future use in research and teaching. The same research community has authorization to reuse shared data for new studies with approval from an ethics board. Indeed, Databrary has *removed the most daunting barriers to video reuse* while *reinforcing core ethical principles* of *informed consent and restricted access to sensitive or identifiable data.*

This project will update and enhance Databrary's infrastructure to make it an even more powerful and essential platform for research on human behavior across SBE and related fields, now and for the future. We will (**Aim 1**) accelerate data reuse through enhanced data discovery tools, (**Aim 2**) ease data reuse with custom collections, (**Aim 3**) expand data sharing through active curation and data workspaces, (**Aim 4**) promote transparency and reproducibility via scriptable access, and (**Aim 5**) enhance existing Databrary data with curation support for search and sharing.

INTELLECTUAL MERIT

Through substantial expansion of and improvements to Databrary, the project will enable novel, innovative, and data-intensive research about the characteristics and consequences of behavior that embrace video as data, documentation, and demonstration. The enhancements will bring powerful, flexible, affordable, and innovative tools to bear on fundamental scientific questions about behavior across SBE, computer science, and related fields while tackling head-on critical challenges of transparency, reproducibility, and ethics.

BROADER IMPACTS

The project will have broad impact across SBE, computer science, and related fields and support substantial numbers of researchers within and outside the U.S. The new tools will expand opportunities for scientists at institutions with limited resources to participate in behavioral research and to engage their students in next-generation research. The proposed enhancements will enrich datasets already shared on Databrary—many funded by NSF and NIH—thereby increasing the value of prior public investments. Project infrastructure will enable more widespread use and transparent sharing of human and AI-derived annotations of video and audio recordings. Together, the proposed activities will elevate the public profile of video-based behavioral research and bolster public interest in SBE fields.

TABLE OF CONTENTS

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

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.

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

Video is a uniquely powerful source of information about behavior in context[1-3], and it is *eminently sharable and reusable*. Video provides *data* about what participants did in natural or experimental tasks, *documentation* about what researchers did when implementing their protocols[2,4], and *demonstrations* of phenomena and findings to enlighten academic communities and the public. Video reveals details of behavior and the surrounding context in formats comparable to the audio-visual experiences of human observers, and digital video is a durable, easily sharable record that allows unlimited opportunities for unlimited observers to revisit and reanalyze past events. Researchers across the social, behavioral, and economic (SBE) and computer sciences routinely use and reuse video as data[5-12], documentation[13-15], and demonstration[1,16], and share it[17,18]. We aim to accelerate the transformative potential of video sharing and reuse by enhancing the NSF-supported Databrary data library—the world's only large-scale repository for research video—to enable previously impossible, data-intensive research in SBE sciences and related fields.

Databrary[18-22] is a web-based, restricted-access data library hosted by New York University, that is specialized for storing and sharing video and audio data from human research participants and non-human animals. Since its launch in 2014, Databrary's corpus of shared video has grown enormously. The library currently stores 177+ TB of data, documentation, and demonstrations, including 120,647+ hours of recordings. Survey, demographic, annotation, and questionnaire data accompany and enrich most recordings. Some 1,013+ openly shared datasets—dozens supported by NSF[23]—represent 10,725+ diverse human participants and have garnered 970+ citations to date[24,25]. A growing user community of 1,783+ authorized researchers (plus 640+ staff and trainees) from 796+ institutions around the globe have contributed data permissioned by participants for broad, unspecified research and educational reuse. The same community of authorized researchers can reuse shared data with approval from their institutional ethics boards. In sum, Databrary has *removed the most daunting barriers* to video reuse while *reinforcing core ethical principles* of *informed consent and restricted access to sensitive or identifiable data*[20,26].

Databrary's access model, ethical policy framework, and ever-growing global community make it a uniquely valuable platform for supporting multidisciplinary research on human behavior—a prime candidate for NSF infrastructure support via HNDS-I. This project will enhance Databrary's infrastructure at every critical point in the research workflow (**Figure 1**). The enhancements—based on lessons learned from a decade of successful operation—will make Databrary an even more powerful and vital resource for behavioral research for the next decade and beyond.

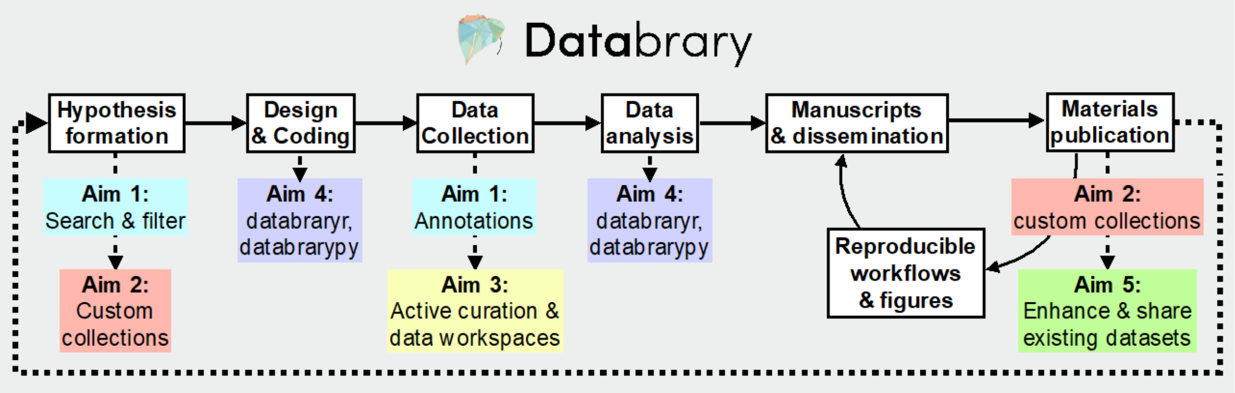

Databrary
Hypothesis formation
Design & Coding
Data Collection
Data analysis
Manuscripts & dissemination
Materials publication
Aim 1: Search & filter
Aim 2: Custom collections
Aim 4: databraryr, databrarypy
Aim 1: Annotations
Aim 3: Active curation & data workspaces
Aim 4: databraryr, databrarypy
Reproducible workflows & figures
Aim 2: custom collections
Aim 5: Enhance & share existing datasets

Figure 1. Enhancements to Databrary at critical points in the research workflow. Aim 1: Accelerate data reuse through enhanced data discovery (search & filter, new annotation layers). **Aim 2**: Ease data reuse with custom collections. **Aim 3**: Expand data sharing via active curation and data workspaces. **Aim 4**: Promote research transparency and reproducibility by expanding scriptable access to Databrary functions (databraryr, databrarypy). **Aim 5**: Enhance the value of existing data on Databrary by adding searchable metadata and by sharing currently unshared but sharable data. These aims will accelerate the power of video-based behavioral research in SBE fields.

PROJECT AIMS

As shown in **Figure 1**, we aim to enhance the Databrary repository to support critical phases of the research workflow. These activities will: (Aim 1) accelerate data reuse, (Aim 2) ease data reuse, (Aim 3) expand data sharing, (Aim 4) promote research transparency and reproducibility, and (Aim 5) enhance the value of existing datasets on Databrary.

Aim 1: Accelerate data reuse through enhanced data discovery.

To reuse videos on Databrary, researchers must be able to find and filter data to suit specific questions and participant characteristics. Databrary's search engine indexes text in dataset descriptions and filters datasets by participant age, filetype, and sharing level. But researchers want to study specific *behaviors* such as "babies laughing," "people playing sports," or "adults falling" in varied *contexts* (e.g., laboratories, homes, or classrooms). Databrary does not currently support this level of search. Doing so requires videos to be accompanied by time-locked annotations in machine-readable formats, a search engine that indexes the annotations, a user interface that returns video segments based on user-supplied search terms, and an interface that enables users to preview videos then select and copy selected segments (with links to accompanying metadata) to custom collections (Aim 2) and data workspaces (Aim 3) for analysis.

Databrary researchers have already generated thousands of time-locked video annotations about infant and adult locomotor, object-related, communicative (speech and gesture), and emotional behaviors[24]. These videos, annotation files, and code books will serve as data for the new search capabilities. The Databrary Application Programming Interface (API) currently supports selection and streaming of specific time-segments from stored videos. Building on this foundation, we will upgrade Databrary to support searching within annotation files linked to videos that tag specific behaviors, utterances, or contexts, starting with the most popular annotation file formats stored on Databrary (Datavyu[25] and CHAT[27]). We will create standard formats for users to share metadata about the meaning and context of variables within annotation files. We will expand the search engine to index protocol documents that describe tasks and instruments and expand the set of user-supplied participant demographic variables available for search and filtering.

Once Databrary's frontend and backend support searching and filtering annotation files and code books, users can then add their own human or machine-generated annotation "layers" to existing shared datasets. The annotation layers will be accompanied by structured metadata about who/what generated the layer and when; variable definitions and associated code books; and other data to support transparent, reproducible analysis workflows (Aim 4). In future work, we can include BIDS[28], ELAN[29], and files derived from computer vision or machine-learning models that can be conformed to the data model supporting annotation indexing on Databrary's backend.

Aim 2: Ease data reuse with custom collections that automatically track provenance across sources.

To capitalize on enhanced search and filtering and ease data reuse, users must be able to create their own custom collections of video files, video segments, annotations, and other data derived from multiple, primary datasets. The custom collections or "virtual datasets" will *link to* but not copy parent datasets and their associated metadata, thereby bolstering the rigor of data provenance and reducing file duplication and storage costs.

Custom collections supplement the enhanced search features implemented in Aim 1. Together, these enhancements will make shared video and audio data on Databrary discoverable to an extent unique among repositories, making Databrary an even more powerful tool for making new discoveries about foundational questions in behavioral science.

Aim 3: Expand data sharing via workspaces that support active curation, thereby reducing the lag between data analysis and open sharing.

Two major hurdles impede widespread adoption of video data sharing—the need for extensive curation prior to sharing[30] and the hassle of moving data from an active workspace where it is analyzed to an accessible, consistently curated, published home in a repository. We learned first-hand about these hurdles in several large-scale, geographically distributed, video-intensive research projects, we lead[13,14,31,32] or consult for[33]. These projects use Databrary as a platform for collecting and analyzing data prior to making it available to the broader Databrary community. We argue that *active curation*—organizing data in a shareable form as it is collected—will accelerate research and reproducibility[21,30,34]. So, we will remove the biggest pain points to using Databrary for active curation. Specifically, as a component of custom collections (Aim 2), we will implement private, flexible, temporary workspaces for datasets that act like folders in cloud storage. Unlike other forms of cloud storage that provide only a temporary home for research data, Databrary's workspaces will provide a permanent and flexible home that is just a button press away from being made accessible to the broader research community. Reducing burdens of *post hoc* curation should make Databrary more attractive to a wider range of researchers and capture data in a sharable form that might otherwise be lost in an electronic file drawer.

Aim 4: Promote research transparency and reproducibility by expanding scriptable access to Databrary functions.

Many scientific fields face profound challenges of reproducibility[35-43]. Video-based research faces additional hurdles. Data collection and video annotation are time-consuming, and personally-identifiable data elements must be altered or handled carefully to protect participant privacy. Tools like Quarto[44], R Markdown[45,] and Jupyter notebooks[46] enable users to generate and document all the steps of quantitative data-processing pipelines and to openly share those workflows. But adoption of such tools is inconsistent. Some video or audio annotation tools support automation via scripting (e.g., Datavyu) but do so only inside the application. This makes it hard to share how annotation files were selected, cleaned, and transformed prior to quantitative analyses. Our solution is to expand access to Databrary's API so that users can write, use, and openly share reproducible workflows that document more steps of their protocol—data collection, cleaning and quality assurance, annotations, visualizations, and analyses.

Imagine a researcher who wants to study why children laugh. They search Databrary for video segments of children laughing and filter by participants' age (using tools developed in Aim 1), collect the segments into their own custom collection and workspace (initially private, but eventually shared, as described in Aims 2 and 3), and then analyze the time window around each bout of laughter for who says and does what by creating time-locked annotations using a tool like Datavyu. The research yields annotation files for each video and a code book that are uploaded to Databrary. The researcher generates data visualizations and conducts quantitative analyses, writes up and submits a manuscript based on the findings, and shares analysis scripts that *document every step of the process* so that any other authorized Databrary researcher (including manuscript reviewers) can run the entire workflow themselves.

To make this possible, we will build on the free, open-source, R package, *databraryr*[47], that PI Gilmore developed with NSF support and openly released to the research community. *Databraryr* wraps Databrary API calls into commands that are useful to researchers who want to download shared data from Databrary. We will add data uploading capabilities to the R package to support Aim 3, develop and publish a parallel Python package, *databrarypy*, sponsor training workshops to bring these tools to the research community, and publish an openly shared web-based how-to manual with examples using Quarto. These enhancements will make fully automated, transparent, and reproducible data workflows involving Databrary's video and related assets much more accessible to the research community.

Aim 5: Enhance the scientific value of existing data on Databrary by adding searchable metadata and sharing unshared data.

Databrary does not initially share newly created datasets outside of the data owner's selected collaborators. As a result, only 33% of datasets are fully shared—even though most uploaded but unshared data could likely be shared. Moreover, many shared datasets would have greater value for reuse if better metadata and descriptive information were provided. To realize the untapped scientific value of these existing data, expert data curators will work with users to enhance existing Databrary datasets with enriched metadata tags and descriptions, and by adding protocols and other documents to make shared datasets more readily discoverable and reusable. In turn, we will encourage users who have sharable but currently unshared or only partially shared datasets to share them, with assistance from a data curator to make the data maximally useful to future research and to discovery via the search engine enhancements described in Aim 1. Some of these enhancements will employ scripting tools we will develop in Aim 4.

INTELLECTUAL MERIT

This project will transform the pace, breadth, depth, transparency, and reproducibility of research in the social and behavioral (SBE) sciences and related fields by creating new, powerful infrastructure for video-based discovery. Through substantial expansion of and improvements to Databrary, the world's only large-scale repository for research video, the project will enable novel, innovative, data-intensive research on multiple facets and dimensions of behavior. Beyond developmental science where Databrary began and has had the most influence, the project will impact substantial numbers of investigators across a wide range of fields that study behavior—psychology, linguistics, anthropology, political science, behavioral economics, education and learning, experimental biology, human-computer interaction, and more. The proposed enhancements to Databrary will bring powerful, flexible, affordable, and innovative tools to bear on central questions in behavioral science while tackling head-on challenges these fields face concerning transparency and reproducibility. The project will thereby create opportunities for new integrative and multidisciplinary research that is at present prohibitively expensive or impossible.

BROADER IMPACTS

The project will have broad impact across SBE fields, and in biological, computer and educational sciences that currently use or could use video, including substantial numbers of researchers outside the U.S. (see **Figure 2A**). The proposed enhancements will enrich datasets already shared on Databrary—many funded by NSF and NIH. The project's tools will expand opportunities for scientists across disciplines at institutions with limited resources to participate in scientific discourse about behavior. Because many under-resourced institutions serve students from under-represented groups, the project will expand research opportunities for them as well. By making video data sharing easier and more attractive to researchers, the project will increase the quantity and quality of shared video datasets available for reuse and the richness of the metadata linked to them. By making coding files, protocols, and coding manuals more readily sharable, the project will accelerate discovery and improve transparency and reproducibility. By making scriptable access to Databrary widely available to everyone, the project will boost transparency, reproducibility, and robustness in behavioral science. Finally, the project will raise the profile of video-based behavioral research and bolster public interest in SBE fields.

RESULTS FROM PRIOR NSF SUPPORT

PIs Adolph and Gilmore received funding from NSF (BCS#1238599, 2012-2014, no cost extension 2014-2016; supplement BCS#1238599, 2015-2016, no cost extension 2016-2017) to support Databrary and Datavyu. These awards built research infrastructure and provided training and technical support to the community. In addition, we published articles that describe Databrary and how it relates to “big data” initiatives in the developmental and behavioral sciences[2,5,9,19,22,48-50]. We developed a policy framework broadly endorsed by NSF and NIH, and we fostered a large community of researchers (currently 1,783+ researchers at 796+ authorizing institutions worldwide, **Figure 2**) committed to ethically sharing identifiable research data based on participant permission. We upgraded the Datavyu video-annotation tool, held workshops to train researchers to annotate video, and wrote about best practices in behavioral video coding[51]. The current proposal builds on and extends these efforts. **Intellectual Merit**. We created infrastructure to enable open sharing and reuse of research video in an open-source web-based repository, Databrary, upgraded and provided user support for the Datavyu video-annotation tool, and fostered a rapidly growing community of researchers committed to video sharing and reuse. Databrary and Datavyu deepen and accelerate the pace of discovery in behavioral science by enabling researchers to view each other’s datasets, reanalyze them to test competing hypotheses, and reuse them to address new questions beyond the scope of the original study. **Broader Impacts**. Databrary empowers behavioral scientists, especially from institutions with limited resources, to conduct high quality research; improves data management practices; and increases transparency and reproducibility. Datavyu brings the power of video-data annotation to any lab with a computer. Databrary's policy framework makes it easy to securely share identifiable video data while upholding ethical principles. Our publications, workshops, and presentations bring this new, collaborative, integrated view of behavioral science to a larger audience.

PI Gilmore received funding from NSF (OAC#2032713, 2020-2022, no cost extension 2022-2023) to expand access to Databrary’s holdings and to bring Databrary into broader compliance with emerging metadata standards[23]. We published the *databraryr* package that interacts with the Databrary API and permits reproducible, scriptable interactions with the system and its data. We use the package to generate weekly reports[24] about Databrary’s holdings, including demographics about the 10,725+ individual participants whose data are currently shared, what funding sources supported the projects, and which investigators are the heaviest users. *Databraryr* supports a project management and quality assurance workflow on the NIH-supported Play & Learning Across a Year (PLAY) Project[13]. A metadata expert compared Databrary’s existing metadata structure to emerging open standards and provided specific guidance about what standards the next version of Databrary should adopt to make the system’s holdings more readily accessible to future research; those recommendations inform the activities proposed here.

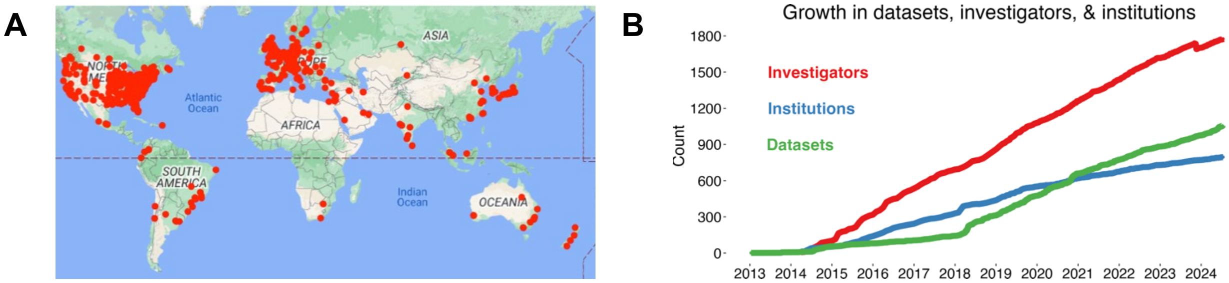


Figure 2. **Databrary status** as of mid-July 2024. **(A)** Map of authorized institutions. **(B)** Growth in datasets, investigators, and institutions from 2013 to current.

BACKGROUND: CONNECTIONS TO ESTABLISHED RESEARCH BASE

Transformative Power of Video

Video has unprecedented power and untapped potential to transform understanding across the social and behavioral sciences. Video documents the microstructure of behavior in real time across most domains of function[1,9,22,50,52,53]. Video uniquely documents interactions between people and their physical and social environment with a richness, detail, and nuance unmatched by any other form of measurement. And it does so with high spatial and temporal resolution. Video provides a permanent record of who did what, and how, when, and where they did it—allowing unlimited revisits and reuse by unlimited numbers of researchers. Video closely mimics the visual and auditory experiences of live human observers, so video collected by one person for a particular purpose can be readily understood and reused by a different person for a different purpose. Indeed, *video data* can make the anatomy of behavior as "tangible as tissue" [54].

Video has untapped value as *documentation*[2,5], regardless of whether video is used as data. The "reproducibility crisis" in science[20,35,36] stems in part from failures to comprehensively report essential details about procedures[55]. Methods sections necessarily omit details about seemingly simple procedures (e.g., recruitment calls, administration of questionnaires, testing environments) and complex procedures (instructions to participants, tasks, video annotations). Words and pictures do not do justice to the subtle interactions and contextual features of typical test situations and computer-based displays. Many research paradigms involve special methods to elicit, test, and record behavior[4,56], such that procedures are like art forms, passed down from mentor to mentee[56]. Video annotation manuals typically refer to tasks and behaviors with quirky, lab-specific labels that make the codes unusable by others[30]. We argue that video documentation of research procedures and codes should be standard practice[2,4,5,20,56].

Video is also a uniquely powerful format for *demonstration*. If a picture is worth 1,000 words, a video is worth 1,000 pictures[1]. Exemplar video clips are the clearest way to illustrate phenomena, and findings, demonstrate how-to's, and clarify annotations for use in teaching and learning. Live links to exemplar videos in journal articles provide readers with instant understanding[1].

Nonetheless, too few behavioral scientists collect video as a primary source of data, use video to document their procedures, or use video to demonstrate their findings. Those who collect video often lack tools and know-how to exploit its full potential; and most researchers who collect or annotate video do not share their videos or annotations. Too many research videos serve only as backup for live human annotation and go unanalyzed and unshared for others to analyze. When videos are annotated, most researchers rely on makeshift spreadsheets or paper and pencil, not powerful annotation tools. Untold hours of video recordings funded by scarce federal research dollars are moldering away in researchers' file cabinets and on defunct hard drives.

The Databrary Video Repository

Databrary is the world's only, large-scale, digital library specialized for research videos of behavior (from fetuses to elderly, in people, non-human animals, and robots), experimental displays, annotation files, and associated metadata. Databrary has overcome the thorniest challenges associated with video sharing while making *video maximally accessible for future reuse*. *Databrary's bedrock principles include open sharing[1], explicit sharing permission from participants, formal institutional authorization, and restricted access[57]*.

Videos often contain personally identifiable information (participants' faces, voices, interiors of their homes, etc.). Ensuring participant privacy is a paramount concern, especially for vulnerable research participants like young children[58]. Altering recordings to obscure faces or change voices renders the videos less usable. Prohibiting human access to the raw video data reduces its value, erodes transparency, and precludes validity checks on computer vision and machine-learning algorithms. Thus, Databrary developed a unique approach to sharing video that keeps the recordings intact but restricts access to researchers who are formally authorized by their institutions via a legally binding agreement[26]. The resulting community is global in scope and

[1] By "open" we mean data that are shared without requiring further permission for use from the original data contributor or co-authorship with the original data contributor, but that cite the parent dataset.

grows weekly (**Figure 2**). In addition, Databrary requires authorized researchers to collect and report explicit sharing permission from participants in a common format[59], using standard language Databrary developed[60] (or an IRB-approved equivalent). Databrary demonstrates that *identifiable (video) data can be openly and widely shared at scale*—not just in proprietary, project- or institution-specific data archives with often idiosyncratic access policies[61,62].

Large file sizes and diverse formats pose technical challenges for video storage, streaming, and sharing. Databrary solves these technical challenges by transcoding all files into a common format (MP4) and by securely storing the original files and transcoded copies on large-capacity servers (see **Technical Plan**).

Video sharing poses practical challenges of data management. Researchers lack time and resources to find, label, organize, link, and convert a typical project's mixed media files into formats that can be used and understood by others[63,64]. Most lack expertise in data curation[21,22,48] and do not fully document workflows or data provenance[48]. When researchers do share, standard practice involves organizing data after a project is finished, perhaps when a paper goes to press[30,53]. Preparing-for-sharing after the fact is a difficult, unrewarding chore that often exceeds the "incremental cost" and "reasonable time frame" envisioned in NSF's Data Sharing Policy[65]. *Post hoc* sharing also makes curation a challenge for repositories[21,30,48]. Databrary reduces *post hoc* sharing burdens by encouraging researchers to self-curate as they collect data, using drag-and-drop file uploading, supporting collections of multiple file types (see **Data Management Plan**) and providing a standardized spreadsheet interface to enter participant and session metadata.

Databrary's unique combination of features has proven itself. In a typical day, Databrary has 1,500 unique visitors to the site (authorized investigators, their lab team members, and the public); authorized users download hundreds of GBs of data for use and reuse, and they upload hundreds of files for future sharing.

Nevertheless, Databrary was developed in 2012-15 using then state-of-the-art technologies. The software stack had not been updated substantially until 2024 (see **Technical Plan**). To address growth in users and stored datasets (**Figure 2**), we are working with application software and data security experts to overhaul and update the system to use modern tools and libraries. The rewrite will make the system more stable and responsive to multiple simultaneous users, patch security vulnerabilities, and ensure a platform for future growth. We are migrating the old system and data as we deploy the new system. In turn, we are creating training materials and documentation to help old users learn the new interfaces, and reintroducing Databrary to the research community via workshops at professional conferences—just as we did for the original Databrary[49]. Note: *The rewrite is separate (and separately funded) from the current proposal*, but essential to it: The new software stack will make implementing new features faster and easier.

A Platform for Sharing and Discovery

In developing and sustaining Databrary, the PIs are among Databrary's most active users, testing Databrary's limits as a tool for active discovery. Our experiences convince us that Databrary can become an even more powerful tool for supporting innovative, multidisciplinary, cross-site, research collaborations across SBE and related fields.

For example, PIs Adolph and Gilmore co-lead (with Catherine Tamis-Lemonda) the NIH-supported Play & Learning Across a Year (PLAY) project[13]. PLAY is advancing discovery about behavioral development in infancy, focusing on advances in language, object interaction, locomotion, and emotion regulation. PLAY is creating the first, cross-domain, large-scale, curated video corpus of human behavior—collected with a common protocol and annotated with common criteria jointly developed by the 73-member launch group. The final corpus will consist of videos of 900+ infant-mother dyads (12-, 18-, and 24-month-olds, 300 per age) from 30 diverse sites across the United States. Videos are transcribed and annotated for infant and mother speech, communicative acts, gestures, object interactions, locomotion, and emotion using the free, open-source video coding tool *Datavyu*[25], developed and maintained by PIs Adolph and Gilmore. The corpus is augmented with video home tours and questionnaire data on infant language, temperament, motor development, gender identity and socialization, home environment and media use, and family health and demographics. PLAY sets new standards for using video as documentation to facilitate discovery and ensure transparency and reproducibility[30,66]. The PIs built on the infrastructure, protocols, and documentation of PLAY for spin-off projects funded by the EPA[14], LEGO[32], and Robin Hood Foundation[67], and we are now taking PLAY global[31].

Databrary is central to PLAY. Each data collection site uses Databrary to curate and upload data into a ready-to-share organizational scheme. Databrary hosts training and sample videos for data collection and coding teams. Custom software applications in Python and R (some using the *databraryr* package) call the Databrary API to assist with project management and tracking[68]. The final PLAY dataset will be shared with the research community in a single PLAY collection drawn from the data collected in 30+ individual sites. In using Databrary so extensively, we became intimately familiar with the system's promise as a platform for discovery. For example, time-locked annotations from human experts provide precisely the information needed to seed a sophisticated search engine. Curating multi-measure data into consistent formats and organization schemes render it (read) accessible and suitable for automation via scripted calls to an API. By incorporating Databrary sharing permissions into our project, we are assured that only data permissioned for sharing will be accessible to the broader Databrary community. In implementing these big-data projects, we also became intimately familiar with Databrary's pain-points. The proposed enhancements in Aims 1-4 reflect that hard-won practical experience.

Remaining Barriers This Project Will Overcome

Databrary has already surmounted barriers that limit the secure, ethical sharing of video. PLAY has overcome barriers that impede progress in the science of early behavioral development, doing so with unprecedented attention to transparency and reproducibility. Here, we will overcome additional barriers that impede the transformative potential of video-based behavioral science:

Aim 1: We will deliver on the potential of search as a tool for discovery by seeding an upgraded search engine with time-locked annotations and code books from PLAY and other big-data projects while expanding filtering to include a broad range of demographic variables.

Aim 2: We will create infrastructure to allow users to create custom "virtual" collections of recordings and annotations that automatically track the provenance of each contributing data element to ensure transparency.

Aim 3: We will make Databrary a more flexible tool for collecting, storing, and analyzing data from in-progress projects—prior to sharing with the research community—by creating private, temporary workspaces that can be easily imported into Databrary proper when it is time to share.

Aim 4: We will make video-based behavioral research a leader in research transparency and reproducibility by enhancing access to Databrary API functions via the publication of free open-source scripting libraries in R and Python, alongside sample scripts and other training materials.

Aim 5: We will make existing shared datasets more readily discoverable by improving searchable metadata. We will expand the number and diversity of shared datasets by providing expert curation services to the owners of private datasets and sharing them.

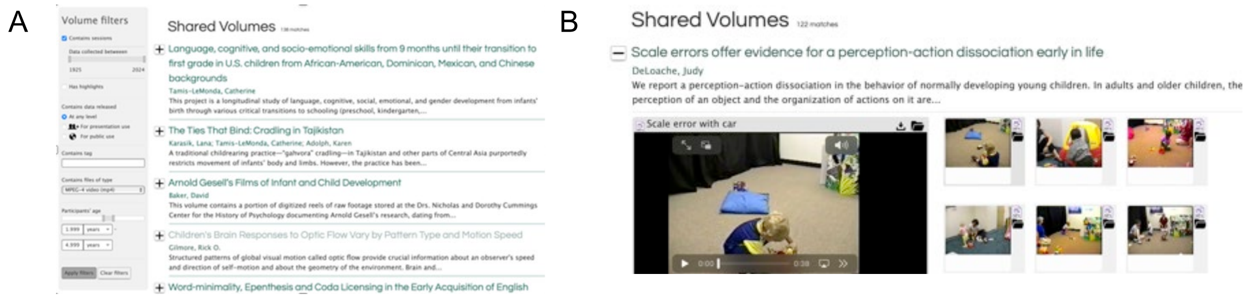
IMPLEMENTATION PLAN

Our implementation plan closely parallels the five aims.

Aim 1: <u>Accelerate data reuse</u> through enhanced data discovery.

To reuse shared data, researchers must be able to find and select videos and segments of videos based on the specific questions they wish to answer. The projects supporting Aim 1 will make data already stored on Databrary more discoverable and easier to reuse.

<u>Project 1.1: Update Databrary's search engine and user interfaces</u>



Shared Volumes 138 matches

+ 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
Tamis-LeMonda, Catherine
This project is a longitudinal study of language, cognitive, social, emotional, and gender development from infants' birth through various critical transitions to schooling (preschool, kindergarten,...

+ The Ties That Bind: Cradling in Tajikistan
Karasik, Lana; Tamis-LeMonda, Catherine; Adolph, Karen
A traditional childrearing practice—"gahvora" cradling—in Tajikistan and other parts of Central Asia purportedly restricts movement of infants' body and limbs. However, the practice has been...

+ Arnold Gesell's Films of Infant and Child Development
Baker, David
This volume contains a portion of digitized reels of raw footage stored at the Drs. Nicholas and Dorothy Cummings Center for the History of Psychology documenting Arnold Gesell's research, dating from...

+ Children's Brain Responses to Optic Flow Vary by Pattern Type and Motion Speed
Gilmore, Rick O.
Structured patterns of global visual motion called optic flow provide crucial information about an observer's speed and direction of self-motion and about the geometry of the environment. Brain and...

+ Word-minimality, Epenthesis and Coda Licensing in the Early Acquisition of English

Shared Volumes 122 matches

− Scale errors offer evidence for a perception-action dissociation early in life
DeLoache, Judy
We report a perception-action dissociation in the behavior of normally developing young children. In adults and older children, the perception of an object and the organization of actions on it are...

Scale error with car

Figure 3: Current Databrary search and filter interfaces. (A) Databrary search page depicting *n* = 138 shared datasets (volumes) with MP4 videos of 2- to 5-year-old children. **(B)** Exemplar dataset with videoclip highlights that are available for reuse in scientific presentations or for teaching.

Preliminary work. Databrary currently allows researchers to search for terms linked in dataset descriptions and titles (**Figure 3A**) and to filter datasets based on participant age, file type, and whether videoclip "highlights" are available for reuse in talks or teaching (**Figure 3B** and **Data Management Plan**). Thus, researchers can search for datasets that include specific participant ages and preview videoclips if available. However, Databrary returns results about *datasets*, not *individual participants*, and not about *segments of behavior within individual recordings*. Furthermore, information in research protocols, annotation manuals, demographic variables, and task descriptions stored on Databrary (as txt, pdf, docx, xls, etc.) is not yet indexed for search. Deficiencies in current search and filter functions make it difficult for users to find appropriate videos for secondary data analyses, for seeing procedures to support review of prior work, or to facilitate replications of procedures, and for using and creating videoclips for teaching.

We will evaluate whether to keep the existing Solr search engine or replace it (**Technical Plan**). The enhanced search engine will expand the set of document types indexed, and we will expand demographic characteristics subject to search. We will update the search engine interfaces to return data at the dataset, individual participant, individual file, and segment-within-file levels.

To test the new search capabilities and interfaces, we will use the extensive, quality-assured data from PLAY and spin-off projects (1,500+ hours of video, all with well-documented demographic data, transcribed video recordings, human-annotated videos of communicative, object-related, locomotor, and emotional behaviors, and complete coding manuals and protocols).

<u>Project 1.2: Update Databrary's backend to support storage of multiple annotation layers and develop interfaces to depict multiple annotation layers</u>

To use annotation files that provide different "lenses" on varied behaviors, such as those provided by PLAY's separate annotation layers, Databrary must parse annotation files and treat them as distinct data streams with their own metadata, including code values and definitions.

Preliminary work. In creating English and Spanish transcripts of the hour-long natural home videos collected in PLAY, we implemented a semi-automated workflow that uses a private-to-NYU-researchers instance[2] of OpenAI's Whisper[69] audio speech-to-text tool. Whisper has dozens

[2] A private instance is essential because many commercial AI tools can result in data leaks to private entities that violate informed consent.

of parameters that can affect the quality of transcribed speech, and transcription passes with identical parameters and model versions can vary widely from run to run. Nonetheless, we find that selecting the best runs from several Whisper passes can save substantial human transcriber time—a huge benefit for large datasets. This work relies on Datavyu, our desktop video coding tool, to capture (human or machine-generated) transcription passes as separately visible layers. We store metadata about model versions and parameters that are critical for transparency and reproducibility in unstructured text files.

Project 1.2 builds on this preliminary work by bringing to Databrary information about annotation layers through files generated by desktop tools like Datavyu that are linked to specific recordings. We will enhance Databrary's backend to support annotation layers and the metadata linked to them (e.g., human- or machine generated, model parameters, etc.). This will require parsing annotation files, each of which can have its own idiosyncratic syntax, and expanding the Databrary database schema to support the time-series structure of annotation files. We will start with Datavyu[25] files, as Databrary has thousands of these already with 1000+ more from PLAY that will be shared by June 1, 2025. Datavyu files are well-known to the Databrary user community, and the format is easy to parse. We will proceed to support the parsing and indexing of CHAT[27] transcripts, as this format can be converted to or from Datavyu, is present in large numbers on Databrary (550+ files) and has an open and easy-to-parse text-based format.

In parallel, we will develop new user interfaces that permit users to select and visualize annotation layers that are available for a given recording, possibly from multiple annotation files. The interface will provide some of the functionality of the Datavyu and CLAN applications that allow the sequence of time-locked annotations to be displayed as the recording is streamed on Databrary. The interface will also provide links to annotation layer metadata—what individual codes mean in a particular layer (e.g., *m*=mouthing vs. *m*=mother), and links to more elaborate narrative descriptions of the coding process stored in materials folders elsewhere on Databrary. We will depict on the frontend and track on the backend the source(s) of annotation layers provided by Databrary users who were not the original contributors of video and annotation files, such as transcripts generated by automated AI tools. We will consider how users who create their own copies of a video (see Project 2.1) and provide their own annotation layer(s) can best make these accessible to the original dataset creators and the rest of the Databrary community. Some of this work naturally overlaps with Projects 3.1 and 3.2.

Aim 2: Ease data reuse with custom collections that automatically track provenance across sources.

Because video captures so many dimensions of behavior, it can readily be reused to answer new questions beyond the original study and support research transparency and reproducibility. Finding shared video data that meet a researcher's needs is one challenge; creating aggregated collections of shared data that can be reused for new purposes is another.

Imagine a researcher interested in whether speech differs if speakers are walking or stationary. Using the enhanced search features developed in Project 1.1, the researcher might find a dozen studies with 100s of videos that meet their criteria—videos of children and adults walking and stationary, either with or without speech transcripts. At present, the researcher must download each dataset to their local computer and examine each dataset one by one to determine whether the data could be recoded to answer their question. The management of found videos, their sources, and citation information is difficult, time-consuming, and error prone. Moreover, 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. Thus, the provenance of the painstakingly collected video data is undermined as is the potential for new insights drawn from linking new sets of annotation layers to previously shared videos (Aim 1).

Project 2.1: Create new custom collections of shared videos for reanalysis

Preliminary work: Databrary already generates persistent identifiers for shared datasets; volume and session interfaces provide vital metadata for reuse; datasets can have multiple links to other web-based resources (including on Databrary); 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 custom collections of videos (and other data) derived from existing datasets in a parent-child structure. The custom collections will be stored and presented using the same volume interface Databrary now uses with the exception that individual sessions consist of videos linked directly to the original dataset plus links to sessions or specific files (e.g. annotation layers in Project 1.2) stored in other datasets. The other datasets could be owned by the new researcher or another authorized researcher. 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 cite the materials from which their custom collection draws. The raw videos and shared coding files would not be copied, but linked, thus saving storage space, and making it easier to track provenance. After a new data collection is shared with the Databrary community, links from the original source to the new collection will be added to the original dataset so that researchers can track how videos are being reused, especially at the level of individual annotations. If a researcher adds one or more annotation 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 future studies can build on both the original and newly applied codes. A new researcher can add new or revised code definitions to the electronic protocol/coding manual associated with the new collection. Those manuals will also link back to the original sources facilitating search.

Implementing custom collections will entail modifications to the Databrary backend and the volume, spreadsheet, and session interfaces. The modifications will distinguish the visual representation of "virtual" (linked components) from those directly associated with the original dataset, and the interface will support logical links for navigation among the elements. We will enable users to save a set of found and filtered data to the new custom collection. We will modify the Databrary backend to track which dataset components are linked from other sources and which are not and add notifications to dataset owners when their existing datasets are cloned or linked to. A transparent, reproducible, and robust chain of knowledge about shared videos that empowers researchers to build on each other's findings and transparently share discoveries will enhance the scientific value of shared video data and accelerate discovery.

Aim 3: <u>Expand data sharing</u> via workspaces that support active curation, thereby reducing the lag between data analysis and open sharing.

Databrary supports and encourages active curation[21,30]—uploading and organizing datasets into publishable forms as data are collected. This reduces barriers to *post hoc* data curation, but poses challenges when users prefer file organization schemes different from the one imposed by Databrary. Repositories like OSF make data curation especially easy by linking projects to a user's cloud storage. But imposing little to no common structure across shared datasets undermines repository-wide searching and filtering. Lack of common structure makes aggregation across datasets much harder (Aims 1-2), and this limits reuse. Work under Aim 3 will strike a balance between extreme flexibility and overly rigid structure for organizing in-progress datasets. This will make Databrary more effective and useful for active curation and more attractive to a broader community of SBE researchers.

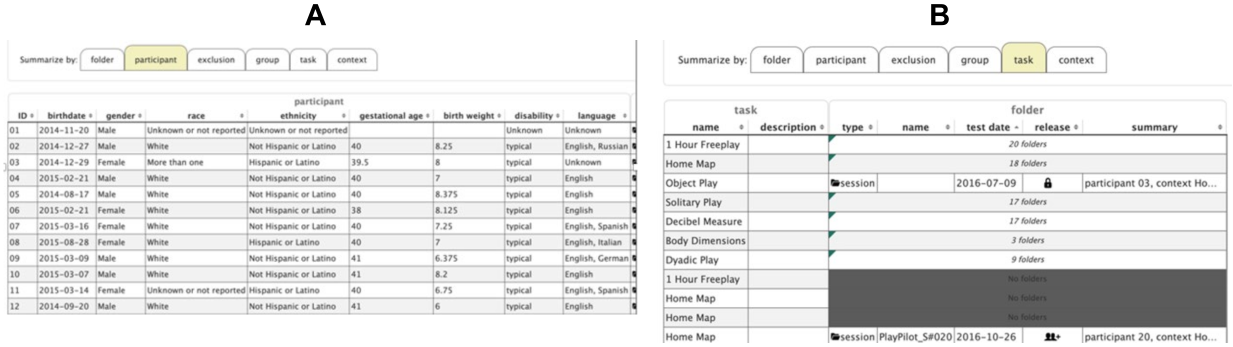
Preliminary work: We extensively tested Databrary's capabilities as a tool for active curation in conducting the PLAY project—a necessity due to 30 geographically distributed data collection sites and 46 video annotation sites. In doing, we recognized the critical need for a more flexible, file-system-like, organizational scheme for in-progress projects whose components could be integrated with Databrary in a later, pre-sharing, data publication step. Indeed, we wrote our own custom data curation application in Python that uses Box cloud storage alongside Databrary to handle the PLAY workflow. Future researchers will be able to do all their curation in Databrary.

Project 3.1: Implement "workspaces" for in-progress research projects

Rather than force data contributors to conform their file organization schemes to the Databrary format, we will implement workspaces where data from in-progress, unshared, research projects can be uploaded. The workspaces can reflect the directory/folder structure of researchers' preferred schemes, and permit them to easily create, modify, or delete directories and move files among them. In a second phase, we will develop user interfaces that empower users to "upload" data to Databrary proper, except that the data are simply moved from one part of Databrary to another. This process will implement a semi-automated curation process where the user answers questions about how the working data are organized (e.g., task/person or person/task) via an "import Genie." Note in initially uploading the in-progress, not-ready-to-be-shared data, the system adds system-unique identifiers, making the files available for subsequent access by the Databrary API. So, quality assurance, data cleaning, and visualization and analysis pipelines can be developed that access files already "known" to Databrary's backend; these need not be altered when the data are ultimately shared with the broader Databrary community. These features strengthen other initiatives to improve transparency and reproducibility (Aim 4).

Project 3.2: Update Databrary's UI to permit flexible views of datasets

The current Databrary application implements a flexible, but complex spreadsheet-like interface for visualizing datasets and participant demographics (**Figure 4A-B**). The spreadsheet requires manual data entry, even when users have created their own well-structured files with task- or participant information. In addition, the UI was written in custom JavaScript and is inflexible and hard to modify or upgrade. We will re-implement the interface to make it more flexible, easily updated, and useful. Researchers who upload structured spreadsheets, CSVs, or data files with task- or participant information to private custom workspaces (Project 3.1) will be able to import them into a searchable and filterable form readable by the Databrary backend.



A

participant								
ID	birthdate	gender	race	ethnicity	gestational age	birth weight	disability	language
01	2014-11-20	Male	Unknown or not reported	Unknown or not reported			Unknown	Unknown
02	2014-12-27	Male	White	Not Hispanic or Latino	40	8.25	typical	English, Russian
03	2014-12-29	Female	More than one	Hispanic or Latino	39.5	8	typical	Unknown
04	2015-02-21	Male	White	Not Hispanic or Latino	40	7	typical	English
05	2014-08-17	Male	White	Not Hispanic or Latino	40	8.375	typical	English
06	2015-02-21	Female	White	Not Hispanic or Latino	38	8.125	typical	English
07	2015-03-16	Female	White	Not Hispanic or Latino	40	7.25	typical	English, Spanish
08	2015-08-28	Female	White	Hispanic or Latino	40	7	typical	English, Italian
09	2015-03-09	Male	White	Not Hispanic or Latino	41	6.375	typical	English, German
10	2015-03-07	Male	White	Not Hispanic or Latino	41	8.2	typical	English
11	2015-03-14	Female	Unknown or not reported	Hispanic or Latino	40	6.75	typical	English, Spanish
12	2014-09-20	Male	White	Not Hispanic or Latino	41	6	typical	English

B

task		folder				
name	description	type	name	test date	release	summary
1 Hour Freeplay		20 folders				
Home Map		18 folders				
Object Play		session		2016-07-09	🔒	participant 03, context Ho...
Solitary Play		17 folders				
Decibel Measure		17 folders				
Body Dimensions		3 folders				
Dyadic Play		9 folders				
1 Hour Freeplay		No folders				
Home Map		No folders				
Home Map		No folders				
Home Map		session	PlayPilot_S#020	2016-10-26	👥+	participant 20, context Ho...

Figure 4. Current Databrary spreadsheet formats. **(A)** Participant-centered view of shared data. **(B)** Task-centered view of the same dataset. Users can toggle between various views via a button click.

Aim 4: <u>Promote research transparency and reproducibility</u> by expanding scriptable access to Databrary functions.

PLAY demonstrates that many aspects of the video annotation and analysis process can be proceduralized and thereby made reproducible[13,30]. Script-based automation improves on manual processes while simultaneously making multi-step data processing procedures more transparent and robust. Work under this aim will enhance and expand the potential for creating and sharing automated workflows that interact with data and metadata stored on Databrary.

Preliminary work: The *databraryr* package[47], developed and published by PI Gilmore with NSF support, enables automated, reproducible, research workflows used in PLAY[68], the Databrary analytics site[24], and other contexts. However, the current R package does not support data upload, and many researchers prefer Python. We created many scripts and functions in Python, but these functions are not yet packaged, documented, made consistent with *databraryr*, or released to the research community in usable ways.

Project 4.1: Polish, document, and promote Databrary's API

We will polish, document, and promote the use of Databrary's API so that Databrary's growing assets can be made more accessible and useful to a wider range of researchers. Polishing and documenting the API will make it easier for researchers to extract Databrary metadata and data in organized ways for subsequent analyses. Promoting use of the API via communication with relevant communities through conference presentations, webinars, workshops, and email lists will expand and diversify the community of researchers who use the system.

Project 4.2: Update databraryr and develop and release a Python package, databrarypy, for accessing Databrary

Building on the refreshed and more thoroughly documented API, we will update *databraryr*, adding file upload abilities, and develop and release a Databrary-specific Python package (*databrarypy*) with parallel functionality. The Python library will integrate code fragments developed for one-off projects and build on the Databrary 2.0 codebase now under development. Since R and Python are common languages in data science, these tools should enable a large audience of users to create transparent and reproducible workflows centered on Databrary.

Project 4.3: Create and publish sample scripts and introduce the new automation tools to the research community

The R code we use for generating the Databrary Analytics report and the PLAY Project survey data pipeline are already published on GitHub[70,71]. In expanding the *databraryr* package and developing the *databrarypy* package, we will create additional exemplar scripts that users can implement in their own data processing workflows. We will hold workshops at professional conferences and webinars to train users to implement these tools—as we have done previously in widely attended trainings for Databrary and Datavyu[49].

Aim 5: Enhance the value of existing data stored on Databrary by adding searchable metadata and sharing unshared data.

Many shared datasets on Databrary are not readily findable because they lack the relevant searchable text in project descriptions, metadata tags, and demographic data. Moreover, datasets already stored on Databrary but not shared provide prime targets for expanding the number and diversity of holdings. So, we will add searchable metadata to shared datasets and encourage owners of partially shared or unshared volumes to share their data, with support from an expert data curator (see **Sustainability Plan**). Note, many shared datasets also include files that were not permissioned for sharing so that researchers can keep their entire datasets intact. But even in the most extreme instances with sensitive data or old data that cannot be grandfathered into Databrary[72] some small portion of the data can be shared as documentation or demonstration.

Project 5.1: Make existing shared datasets more readily findable.

Preliminary work: Based on an internal review of shared datasets on Databrary, we find that many do not contain an informative title or complete descriptions of the dataset with sufficient detail to enable effective search. Others do not contain demographic data about participants that could support filtering. Most do not contain metadata tags or keywords even when the dataset corresponds to a published paper where keywords are given. Still others lack external links to published manuscripts or source code that could be indexed. Study or IRB protocol documents or even grant proposals could be added as materials.

To make shared datasets more findable and reuseable, our expert curators will evaluate each dataset and determine what type of curation help would have the biggest impact on discoverability, thereby supporting Aim 1. We anticipate this curation support to vary by dataset. In some cases, it may be necessary to provide richer and fuller dataset descriptions or more informative titles; in others, participant-specific demographic data may need to be entered; and in still others, project protocols, code books, or other text-based descriptive data or metadata will be uploaded and shared. If Datavyu or CHAT-formatted annotation files are available for individual video or audio recordings, those will also be a priority of the curation effort to support work in Aim 1. If dataset owners have published or unpublished manuscripts that build on the dataset, links to those will be added. If dataset owners created and shared computer code used to analyze data, links to those repositories will be added. Because the curation effort requires substantial human effort, we will develop a process for soliciting interest from the Databrary community, then prioritize projects for assistance based on criteria such as the number or type of video recordings, the availability of informative data and metadata, and related factors.

Project 5.2: Curate unshared datasets and release them for wider use.

Preliminary work: Using *databraryr* (Aim 4), we already evaluate and report[24,70] on Databrary's holdings on a regular basis, including types of stored data, tags applied, funding support, and demographics of participants whose data are shared. These results tell us that although Databrary's shared data have grown substantially (**Figure 2B**), more than 2/3 of the datasets *already uploaded to the system* have data that remain inaccessible to the broader Databrary community. Based on their creation dates, we suspect that many of the not-yet-shared datasets stem from completed projects and few from work still in-progress. Thus, we think Databrary holds a large pool of unshared data that could be shared relatively easily with curation assistance and made available for reuse. Databrary staff have experience providing curation to large-scale projects such as PLAY[13], the Developing Belief Network[33], and others.

Project 5.2 builds on that expertise and on work described in Project 5.1. We will start by identifying unshared datasets with large storage footprints. A data curator will contact the dataset owners and offer to work with them to curate their datasets for eventual sharing. Some level of curation support will be offered without charge, but for large curation efforts, we will ask dataset owners to provide financial support based on an assessment of the workload involved, and the owner's enthusiasm for bringing increased value to the already stored data. Users who share unshared data can avoid data deposit fees (see **Sustainability Plan**).

COORDINATION & MANAGEMENT PLAN

The project will be overseen by PIs Adolph and Gilmore. The PIs hold weekly meetings by phone or video conference to discuss Databrary-related matters. The PIs will meet bi-weekly with the NYU IT project team and the Montrose Software team to formulate long- and short-term plans, get progress updates, and provide input.

Evaluation and Assessment

We will evaluate progress in several ways. The PIs will report progress for each project for each aim relative to the goals outlined in the timetable (see **Technical Plan**) in the annual project report. When a particular feature is released to the larger Databrary community, we will seek feedback from the Databrary User Advisory Board (described below). We will also gather use statistics, such as the numbers and types of searches (Aim 1), number of researchers who create and share custom collections (Aim 2), usage statistics about data workspaces (Aim 3), and downloads and use of *databraryr* and *databrarypy* to access Databrary functions (Aim 4). We will track curation efforts including number of data owners contacted, number who agree to curation support, and numbers curated and shared (Aim 5). At present, we estimate data reuse and citation by processing Google Scholar search results. We will develop new, improved estimates of data reuse based on download statistics, sharing statistics, and the generation and citation of new custom collections. We will administer surveys to Databrary users twice per year to solicit feedback about system operations, focusing on new features, and asking users the extent to which the new features increase their willingness to share data, the ease of doing so, and the attractiveness of reusing others' data. The results of those surveys will be summarized in the annual NSF report and shared publicly on the web.

Program-Specific Criteria

The proposal addresses multiple criteria applicable to the HNDS-I program. Databrary is already a "user friendly, large-scale, next generation data resource" that supports a variety of "analytic techniques to advance fundamental research in SBE areas of study," including developmental science, linguistics, perception, action cognition, cognitive neuroscience, social psychology, sociology, economics, and cultural anthropology. Databrary supports the scholarship of 1,783+ researchers and their affiliates, including dozens of projects funded by NSF[23,24]. Thus, the activities proposed here involve a "substantial expansion [and] revision of an extant database."

With respect to *science*, the enhancements to Databrary will enable new, integrative, interdisciplinary research about the characteristics and consequences of human behavior across ages, domains of function, and contexts. These new questions can be answered with unprecedented depth and impact at scale. The research communities interested in exploring the questions span SBE fields and extend to the education and learning sciences, experimental biology, computer vision/AI, and human-computer interaction. With respect to *information technology*, the project elevates and amplifies the status and importance of video as data, documentation, and demonstration, and it substantially enhances the diversity of metadata provided for and linked to shared video, improves tools for exploiting metadata, and expands the communities that benefit. The proposal builds on new, but established infrastructure with strong institutional backing, led by a pair of dedicated PIs who are committed to the long-term sustainability of the tools. With respect to *governance*, the project will form a Databrary User Advisory Board derived from the 73-member PLAY project group[73], colleagues in the Developing Belief Network (DBN)[33], and the Databrary user community. As PLAY and DBN wind down in the next year, we will solicit volunteers from these groups to provide guidance for the next phase of upgrades to the Databrary system.

Summary

Video uniquely and powerfully captures human behavior in context and naturally serves as the focus of a research ecosystem for behavioral discovery as data, documentation, and demonstration. This project will: (Aim 1) Accelerate data reuse through enhanced data discovery, specifically search and filtering; (Aim 2) Ease data reuse and improve transparency through custom collections; (Aim 3) Expand data sharing via data workspaces that support active curation; (Aim 4) Promote research transparency and reproducibility by making scriptable interactions with Databrary powerful and easy to implement and share widely; and (Aim 5) Enhance the value of existing datasets by adding searchable metadata and by sharing unshared but data. Putting the richness and complexity of behavior into sharper focus by making video recording, annotation, and sharing commonplace will spur discovery in the social and behavioral sciences while making these fields leaders in research transparency, reproducibility, and ethics.

REFERENCES

1. Adolph KE (2020) Oh, behave! *Infancy 25:* 374-392.
2. Gilmore RO, Adolph KE (2017) Video can make behavioral science more reproducible. *Nature Human Behavior 1:* s41562-41017.
3. Adolph KE, Froemke RC (2024) How to get rich quick: Using video to enrich psychology and neuroscience research. Comment on "Beyond simple laboratory studies: Developing sophisticated models to study rich behavior" by Maselli et al. *Physics of Life Reviews 48:* 16-18.
4. Gelman S. (2012) Technology could help: Association for Psychological Science. Available from: https://www.psychologicalscience.org/observer/scientific-rigor#gelman.
5. Adolph KE, Gilmore RO, Kennedy JL (2017) Video data and documentation will improve psychological science. *Psychological Science Agenda*.
6. Derry SJ, Pea RD, Barron B, Engle RA, Erickson F, Goldman R, Hall R, Koschmann T, Lemke JL, Sherin MG, Sherin BL (2010) Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences 19:* 3-53.
7. Goldman R, Pea R, Barron B, Derry SJ (2006). Video research in the learning sciences. Mahwah, NJ: Erlbaum.
8. Alibali MW, Nathan MJ, Wolfgram MS, Church RB, Jacobs SA, Johnson Martinez C, Knuth EJ (2014) How teachers link ideas in mathematics instruction using speech and gesture: A corpus analysis. *Cognition and Instruction 32:* 65-100.
9. Adolph KE (2016) Video as data: From transient behavior to tangible recording. *APS Observer 29:* 23-25.
10. Pasqualino C (2007) Filming emotion: The place of video in anthropology. *Visual Anthropology Review 23:* 84-91.
11. Ossmy O, Adolph KE. Using EEG, Eye-tracking, motion-tracking and video to study human development: Databrary. Available from: http://doi.org/10.17910/B7.707.
12. Ossmy O, Gilmore RO, Adolph KE. AutoViDev: A computer-vision framework to enhance and accelerate research in human development. In: Arai K, Kapoor S, editors. Advances in computer vision: CVC 2019 Advances in Intelligent Systems and Computing. Cham, Switzerland: Springer; 2020. p. 147-156.
13. Play and Learning Across a Year (PLAY) Project. Available from: https://www.play-project.org/.
14. Developmental, behavioral, and environmental determinants of infant dust ingestion [Internet]. Databrary. 2021. Available from: https://doi.org/10.17910/b7.1364.
15. Video Journal of Orthopaedics. JBJS Video Supplements 2011. Available from: https://www.vjortho.com/about/jbjs-video-supplements/.
16. Chaquet JMC, E. J.; Fernandez-Caballero, A. (2013) A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding 117:* 633-659.
17. MacWhinney B. TalkBank [cited 2024]. Available from: https://talkbank.org/.
18. Databrary [cited 2024]. Available from: https://databrary.org.
19. Gilmore RO, Adolph KE, Millman DS (2016) Curating identifiable data for sharing: The Databrary project. *New York Scientific Data Summit*.
20. Gilmore RO, Kennedy JL, Adolph KE (2018) Practical solutions for sharing data and materials from psychological research. *Advances in Methods and Practices in Psychological Science 1:* 121-130.
21. Gordon A, Millman DS, Steiger L, Adolph KE, Gilmore RO (2015) Researcher-library collaborations: Data repositories as a service for researchers. *Journal of Librarianship and Scholarly Communication 3:* eP1238.
22. Gilmore RO, Adolph KE, Millman DS, Gordon AS (2016) Transforming education research through open video data sharing. *Advances in Engineering Education 5:* 1-17.

23. Gilmore RO, Wham B, Clair K, Spies J. EAGER: Expanding public access to restricted research data. Report on NSF OAC 2032713 2023. Available from: https://databrary.github.io/nsf-oac-2032713/.
24. Gilmore RO, Seisler AR. Databrary Analytics Site: Databrary; [cited 2024]. Available from: https://databrary.github.io/analytics/.
25. Datavyu: Video coding and data visualization tool [cited 2024]. Available from: http://datavyu.org.
26. Databrary Policy Framework. Available from: http://www.databrary.org/access/policies.html.
27. MacWhinney B. Tools for Analyzing Talk: Part 1: The CHAT Transcription Format 2024. Available from: https://doi.org/10.21415/3mhn-0z89.
28. Brain Imaging Data Structure (BIDS) [cited 2024]. Available from: https://bids.neuroimaging.io.
29. ELAN [cited 2024]. Available from: https://archive.mpi.nl/tla/elan.
30. Soska KC, Xu M, Gonzalez SL, Herzberg O, Tamis-LeMonda CS, Gilmore RO, Adolph KE (2021) (Hyper)active data curation: A video case study from behavioral science. *Journal of eScience Librarianship 10:* e1208.s.
31. Tamis-LeMonda CS, Adolph KE, Legare C. Global BABIES [cited 2024]. Available from: https://global-babies.org.
32. Tamis-LeMonda C, Adolph KE. The science of everyday play: Databrary. Available from: http://doi.org/10.17910/b7.563.
33. Develping Belief Network [cited 2024]. Available from: https://www.developingbelief.com.
34. Childers B, Davidson J, Graves W, Rous B, Wilkinson D (2016) Active curation of artifacts and experiments is changing the way digital libraries will operate. *CEUR 1686*.
35. Baker M (2016) 1,500 scientists lift the lid on reproducibility. *Nature News 533:* 452.
36. Open Science Collaboration (2015) Estimating the reproducibility of psychological science. *Science 349:* aac4716.
37. Nosek BA, Bar-Anan Y (2012) Scientific utopia I: Opening scientific communication. *Psychological Inquiry 23:* 217-243.
38. Nosek BA, Beck ED, Campbell L, Flake JK, Hardwicke TE, Mellor DT, van 't Veer AE, Vazire S (2019) Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences 23:* 815-818.
39. Nosek BA, Spies JR, Motyl M (2012) Scientific utopia II: Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science 7:* 615-631.
40. Munafo MR, Nosek BA, Bishop DVM, Button KS, Chambers CD, du Sert NP, Simonsohn U (2017) A manifesto for reproducible science. *Nature Human Behavior 1:* 0021.
41. Errington TM, Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021) Investigating the replicability of preclinical cancer biology. *eLife 10*.
42. Oza A (2023) Reproducibility trial: 246 biologists get different results from same data sets. *Nature 622:* 677-678. doi: https://doi.org/10.1038/d41586-023-03177-1.
43. National Academies of Sciences Engineering and Medicine (2019). Reproducibility and replicability in science. Washington, DC: National Academies of Sciences, Engineering, and Medicine.
44. Posit. Welcome to Quarto® 2024. Available from: https://quarto.org/.
45. Allaire JJ, Xie Y, Dervieux C, McPherson J, Lurashi J, Ushey K, Atkins A, Wickham H, Cheng J, Cheng W, Iannone R. rmarkdown. Posit; 2024.
46. Jupyter 2024. Available from: https://jupyter.org/.
47. Gilmore RO, Spies JR. databraryr. 0.6.6 ed: Comprehensive R Archive Network (CRAN); 2024.

48. Gordon AS, Steiger L, Adolph KE. (2016) Losing research data due to lack of curation and preservation. In: Johnston L, editor. Curating research data: A handbook of current practice. Chicago, IL: Association of College and Research Libraries. p. 108-115.
49. Databrary sponsored workshops and events [Internet]. Databrary. 2013. Available from: https://doi.org/10.17910/B7159Q.
50. Adolph KE, Gilmore RO, Freeman C, Sanderson P, Millman DS (2012) Toward open behavioral science. *Psychological Inquiry 23:* 244-247.
51. Adolph KE. Best practices for coding behavioral data from video 2015 [cited 2024]. Available from: http://datavyu.org/user-guide/best-practices.html
52. Adolph KE. (2020) Ecological validity: Mistaking the lab for real life. In: Sternberg RJ, editor. My biggest research mistake: Adventures and misadventures in psychological research. New York: Sage. p. 187-190.
53. Gilmore RO, Adolph KE. (2019) Open sharing of research video: Breaking down the boundaries of the research team. In: Hall KL, Vogel AL, Croyle RT, editors. Strategies for team science success: Handbook of evidence-based principles for cross-disciplinary science and practical lessons learned from health researchers. Cham: Springer. p. 547-583.
54. Curtis S (2011) "Tangible as tissue": Arnold Gesell, infant behavior, and film analysis. *Science in Context 24:* 417-442.
55. Errington TM, Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021) Challenges for assessing replicability in preclinical cancer biology. *eLife 10*.
56. KlavinsLab. Aquarium: Build reproducible experimental protocols and workflows 2018. Available from: http://klavinslab.org/aquaverse/.
57. Gilmore RO, Xu M, Adolph KE. (2021) Data sharing. In: Panicker S, Stanley B, editors. Handbook of research ethics in psychological science. Washington, DC: American Psychological Association. p. 83-97.
58. Begum AJ, Holman R, Goodwin A, Heraty S, Jones EJ (2023) Parent attitudes towards data sharing in developmental science. *PsyArXiv*. doi: https://doi.org/10.31234/osf.io/kv7zw.
59. Databrary Release Levels [cited 2024]. Available from: https://databrary.org/support/IRB/release-levels.html.
60. Data Sharing Release: Participants [cited 2024]. Available from: https://databrary.org/support/irb/release-template.html.
61. Trust W. Wellcome Leap: Unconventional Projects. Funded at Scale [cited 2024]. Available from: https://wellcomeleap.org/1kd/.
62. Hasson U. Research: The First 1,000 Days Project [cited 2024]. Available from: https://hassonlab.princeton.edu/research.
63. Price-Williams D, Gordon W, Ramirea M (1969) Skill and conservation: A study of pottery-making children. *Developmental Psychology 1:* 769.
64. Ascoli GA (2006) The ups and downs of neuroscience shares. *Neuroinformatics 4:* 213–215.
65. NSF. Preparing Your Data Management and Sharing Plan [cited 2024]. Available from: https://new.nsf.gov/funding/data-management-plan#nsfs-data-sharing-policy-1c8.
66. Play & Learning Across a Year (PLAY) project: Quality assurance protocol [cited 2024]. Available from: https://play-project.org/quality.
67. Tamis-LeMonda CS, Adolph KE (2023-2025) Routine math language (MathBABIES). Robinhood Foundation.
68. Gilmore RO. PLAY Project Survey KoBoToolbox Survey Data: PLAY Project; [cited 2024]. Available from: https://play-project.org/KoBoToolbox/.
69. Introducing Whisper: Open AI; [cited 2024]. Available from: https://openai.com/index/whisper/.
70. Gilmore RO, Seisler AR. Databrary Analytics Code [cited 2024]. Available from: https://github.com/databrary/analytics.

71. Gilmore RO. PLAY Project KoBoToolbox Survey Data Report Code: github.com; [cited 2024]. Available from: https://github.com/PLAY-behaviorome/KoBoToolbox.
72. Adolph KE. Pitfall or pratfall? Individual differences in infants' learning from falling: Databrary; [cited 2024]. Available from: http://doi.org/10.17910/b7.1185.
73. PLAY & Learning Across a Year: People [cited 2024]. Available from: https://play-project.org/people.html.
74. Databrary Supported File Formats [cited 2024]. Available from: https://databrary.org/asset/formats.
75. NINDS Common Data Elements [cited 2024]. Available from: https://www.commondataelements.ninds.nih.gov.
76. JSON for Linking Data [cited 2024]. Available from: https://json-ld.org/.
77. Schema.org [cited 2024]. Available from: https://schema.org/.
78. NSF. Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) [cited 2024]. Available from: https://access-ci.org.

FACILITIES, EQUIPMENT AND OTHER RESOURCES

New York University

Laboratory: Not applicable

Clinical: Not applicable

Animal: Not applicable

Computer: Co-PI Adolph has a MacBook Pro laptop that she uses for scholarly activities. Her laboratory is equipped with multiple computers and monitors as described below.

See the **Other Resources** section below for details about resources specifically used by Databrary and the **Technical Plan** for related project-specific information.

Office: Co-PI Adolph's office (120 sq ft) is located adjacent to her lab and is equipped with computer, monitor, and printer. The Adolph lab includes 9 adjoining rooms in a wing of the Psychology Building in the NYU College of Arts and Sciences. The lab rooms have space for ~20 people (postdoctoral fellows, doctoral students, research staff, and Databrary/Datavyu staff) with their own workstations and desks with 32 general-purpose computers (Mac and PC) for research scheduling, behavioral video coding, and data processing, visualization, and analysis. It includes workstations and desktop computers for Databrary and Datavyu technicians and staff and a room dedicated to remote video conferencing for tech support, consulting, and training. The lab also houses one Synology (Linux-based) video data server and an OS X server to provide centralized file storage and user account authentication. Computers are linked via gigabit Ethernet.

The research lab hosts a shared participant database and custom-built scheduling application to update the database and schedule experiments. The laboratory is well equipped for video-based studies of infant and child behavior development, augmented with simultaneously recorded high-speed magnetic motion tracking, electrogoniometers, instrumented gait carpet, head-mounted eye tracking, remote eye tracking, and EEG. The primary laboratory space contains a large (1400 sq ft), calibrated testing room with Unistrut ceiling mounts supporting 16+ fixed high-density video cameras suitable for computer vision analyses, and multiple cameras on tripods of handheld for flexible, panning views. The large testing room can be configured as a playroom with various floor surfaces and climbing toys. The floor space holds two convertible adjustable walkways (a slopes/drop-offs apparatus and a bridges/gaps apparatus) permanently affixed to the floor along opposite walls. Several other adjustable apparatuses (locomotion and reaching apertures, pedestals, hurdles, monkey bars) are stored in a large walk-in closet and assembled as needed.

The space includes a waiting area (changing table, scale, couch, shelves of toys), kitchen/washing area (sink, refridgerator, microwave, etc.), work area with a meeting table and surrounding coding/computer stations for holding project meetings and conducting video coding, data processing, and statistical analyses (e.g., Datavyu, MATLAB, SPSS, R, Adobe Creative Suite including Premiere, Photoshop, Illustrator, and InDesign, Microsoft Office, SigmaPlot). Two large walk-in closets are used to store apparatuses not currently in use, smaller equipment items, and supplies.

Other: Not applicable.

Major Equipment: Not applicable.

Other Resources:

High Performance Computing

NYU's High Performance Computing resources include a powerful HPC cluster, named "Greene", which was deployed in the summer of 2020. Greene includes over 39,000 CPU-cores, 700 GPU cards, and 10PetaBytes of data storage capacity. NYU is in the process of procuring a new HPC cluster, refreshing the Greene cluster incorporating the latest available CPU and GPUs technologies. The upcoming HPC cluster is slated for delivery in the Fall of 2024.

Complementing the HPC cluster resources, the HPC team is actively deploying an on-premises research cloud (RT Cloud) to provide researchers with flexible and secure computing capabilities. This solution leverages a Red Hat OpenShift cluster, offering a unified platform for building, modernizing, and deploying applications at scale.

These HPC and research cloud resources are housed within NYU's Research Computing Data Center (RCDC). The RCDC is a state-of-the-art, private, and energy-efficient facility designed to meet the demand for research computing resources. With 5,000 square feet of raised floor space, housing ten rows of racks, the center currently provides close to 2MW of power.

Furthermore, the on-premises advanced compute resources and key research facilities are interconnected via a dedicated High-Speed Research Network (HSRN). Designed to cater specifically to research community needs, the HSRN operates independently from the academic NYU Campus Ethernet Network (NYU-Net).

NYU entered into a 3-year Cloud Commit Agreement with Google Cloud Platform (GCP) in 2023. This strategic partnership facilitates the bursting of HPC jobs to the public cloud, empowering the HPC team to migrate computational tasks to GCP when faced with extended wait times for GPU-intensive processes.

High Performance Computing Cluster Resources

Greene cluster - Initial Compute Hardware

- 4 login nodes
- 524 "standard" compute nodes with 192GB RAM and dual CPU sockets
- 40 "medium memory" nodes with 384GB RAM and dual CPU sockets,
- 4 "large memory" nodes each with 3TB RAM and quad CPU sockets.

All the above cluster nodes are equipped with 24-core Intel Cascade Lake Platinum 8268 chips. The "standard" and "medium memory" compute nodes (a total of 564 nodes with 27,072 processing cores) are Direct Water Cooled (DWC) nodes by operating two Cooling Distribution Units (CDUs). DWC allows us to run the CPU at Turbo frequency of 3.7GHz nodes while we maintain operation of all processing cores.

All cluster components are interconnected with an Infiniband (IB) fabric in a non-blocking Fat-tree topology, consisting of 20 core switches and 29 leaf switches.

All switches are 200Gbps HDR IB switches while each compute node connects to the fabric using an HDR-100 adapter.

Greene Cluster - GPU Hardware

- The Greene cluster includes 73 compute nodes each equipped with 4 NVIDIA RTX8000 GPUs (a total of 292 RTX8000 GPUs)
- 11 of the nodes are equipped with 4 V100 GPUs for a total of 44 V100 GPUs.
- Each of the above nodes that is equipped with GPUs has 384GB of RAM and two CPU sockets.
- The HPC team expanded the computing power of the Greene cluster in the Spring of 2022 by adding 43 nodes each equipped with 4 A100 GPUs and 2 sockets of Intel Ice Lake CPUs and 1TB of memory (a total of 172 A100 GPUs)
- Another 15 nodes each equipped with 4 of NVIDIA H100 GPUs (a total of 60 H100 GPUs) were added in early 2024.

Greene Cluster - Fall 2021 Cluster Expansion

- The HPC team integrated with the Greene cluster 20 compute nodes provided by AMD and its technology partner Penguin Computing Inc. in the Fall of 2021
- These nodes are part of a larger initiative, the AMD COVID-19 HPC fund, which was established to provide research institutions with computing resources to accelerate medical research on COVID-19 and other diseases.
- The 20 compute nodes, each equipped with an AMD EPYC Rome 7642 processor (having 48 processing cores), 512 GigaBytes (GB) of host memory, 8 MI-50 32GB Graphics Processing Units (GPUs) and 2 TeraByte (TB) of local Solid State Disk (SSD) for data storage.
- All nodes are connected internally using the Infiniband network HDR technology providing a communication bandwidth of 200 Gigabits per second (Gbps).
- The compute nodes are air cooled and were deployed in the Fall of 2020 in a heat-contained area in NYU's Research Computing Data Center (RCDC).
- The added compute nodes can perform one quadrillion (10^15) Floating-point operations per second (a PetaFlop), requiring over 60kW of power to run.
- 5 additional nodes were added in late 2022: 2 equipped with 8 MI-100 AMD GPUs and 3 equipped with 8 MI-250 GPUs.

Databrary transcodes all video and audio files uploaded to the system using HPC resources.

NYU Research Computing Data Center (RCDC)

A private, centralized, colocated Data Center has been designed to meet the growing demand of research computing resources in space, electrical power, and cooling. A state-of-the-art, 5,000 sq feet (50ft x 100ft) of raised floor space, housing ten rows of racks, currently provides 750kW of power and can trivially be expanded to 1.25MW. A modern data center facility with all the electrical power equipment cabling, designed to be power efficient with a Power Utilization Efficiency (PUE) of 1.08. The data center supports energy efficient server cooling methods: direct water cooling to HPC racks, enabling the cooling of dense HPC racks up to 70kW per rack, and heat containment for air cooled racks resulting in improved energy efficiency and contributing to the University's sustainability efforts.

RCDC Network Connectivity

The Data Center is connected to the enterprise NYU network (NYU-Net) and is also linked to a new low-latency, High-Speed Research Network (HSRN), dedicated to research projects and capable of delivering 3.2Tbps to research facilities in the NYU Washington Square campus. The data center has a dedicated fiber connection to 32 Avenue of the Americas, also known as the AT&T building, located in the Tribeca neighborhood of New York City. The building houses Manhattan Landing (MAN LAN) a high-performance exchange point in New York City that supports Layer 2 Ethernet connections to facilitate peering among U.S. and international research and education (R&E) networks. The exchange point is a collaboration between Internet2, NYSERNet (The New York State Education and Research Network), and the Global Research NOC at Indiana University. Through NYSERNet and its peering with the Internet2 Network, we can reach cloud resources, including Microsoft Azure ExpressRoute, Amazon Web Services (AWS) Direct Connect and Google Cloud Platform (GCP) Interconnect. NYU participates in the Internet2 Net+ GCP program and connects to GCP via Internet2 Cloud Connect. The NYU IT Global Command Center (GCC), located within the SDC, provides 24x7 monitoring environmental conditions, UPS/power, physical security, mechanical equipment, all mission-critical administrative and academic systems, data storage, network and connectivity, the processing and scheduling of batch jobs, as well as tape vaulting operations. All of this is made possible using a broad range of monitoring tools: Nagios, ManageEngine, and BMS, to name just a few. GTC in manned 24x7 with system administrators, network engineers, and data center management staff, all co-located in the Command Center. GCC is monitoring of the Syracuse High Availability site (an emergency backup location for many of NYU's crucial software and IT applications) and switch closets. With the addition of Syracuse, the Global Command Center will be monitoring a total of six data centers, including NYU data centers in Abu Dhabi and Shanghai.

Databrary's three servers and storage are housed in the RCDC. The servers are virtual machines hosted on **Cisco HyperFlex HX240c M6** hardware. The storage is hosted on **Dell EMC Isilon A2000** (18 Node) in New York City.

Syracuse Data Center Facility

NYU IT has built an off-site center environment within NYSERNet's Syracuse Data Center facility: The 4,000 sq. ft. data center is maintained and monitored on a 24x7 basis by NYSERNet and is designed to host live systems as well as act as a disaster recovery site for rapid failover of services. To date, NYU IT has deployed 20 racks at the data center, which are fully integrated in the NYU Global Wide Area Network (WAN), enabling the racks to appear as an extension of our NYC data center environment with the same level of security protection mechanisms. The NYU Washington Square campus and Syracuse Data Center are interconnected by dual redundant 10 Gbps links placed along different paths across New York State. The data center has 3 Gbps (expandable to 10 Gbps) of Internet access, and 5 Gbps of access to Internet2 and the global National Research and Education Networks (NRENs).

Backups of Databrary data are hosted in the Syracuse Data Center Facility. All data are replicated to a **Dell EMC Isilon A2000** (18 Node) machine in Syracuse.

The Institute of Human Development and Social Change

The Institute of Human Development and Social Change (IHDSC), a multidisciplinary research institute at New York University (NYU), offers a range of expertise and resources to support research both intellectually and administratively. IHDSC houses over 100 active grants totaling over $70 million and provides access to a disciplinarily diverse network of over 110 faculty and scholarly affiliates. IHDSC specializes in robust and comprehensive grants management support including financial management, subrecipient monitoring, personnel and procurement services. The Institute coordinates closely with NYU's central offices for sponsored programs and financial administration to ensure that research activities take place in compliance with regulatory requirements. IHDSC manages and provides access to research facilities (e.g., workspaces for research personnel, technology-equipped meeting rooms for videoconferencing, etc.) and provides communications support to help researchers disseminate findings.

FACILITIES, EQUIPMENT AND OTHER RESOURCES

The Pennsylvania State University

Laboratory: Not applicable

Clinical: Not applicable

Animal: Not applicable

Computer: The Psychology Department provides computers for faculty. Co-PI Gilmore has a MacBook Pro laptop that he uses for scholarly activities, including software development.

Office: Co-PI Gilmore has office space, human behavioral testing space, and space for video coding and data analysis in the Department of Psychology's Moore Building. His space includes university-provided Mac OS laptops, iMacs, and PCs available for research use. These are networked to the PSU College of Liberal Arts computer system. PSU User accounts are managed by the University Information Technology Administrators. The user's "rights" are set by the PI. Datasets for analyses are de-identified and made available on OneDrive dedicated to the relevant project.

Other: Not applicable.

Major Equipment: Not applicable.

Other Resources: Not applicable.

DATA MANAGEMENT PLAN

OVERVIEW

This document describes plans for managing data and metadata collected as part of the proposed project activities to enhance Databrary as a tool for discovery about human behavior.

DESCRIPTION OF THE DATA

As a restricted access data library, Databrary stores and shares video and audio data, annotations of video and audio data, and metadata about the characteristics of human participants and nonhuman animals along with information about the contexts of data collection. The proposed work will not collect new data but rather enhance Databrary to facilitate storage, sharing, and reuse of research data by others.

STANDARDS FOR DATA AND METADATA FORMAT AND CONTENT

Research Data

Databrary supports storing and sharing multiple file types[74]: video (webm, mpg, mov, mts, avi, wmv, dv, mp4); audio (wav, aac, wma, mp3), text (txt, csv, rtf, cha, eaf); images (png, jpg); and documents (pdf, doc, docx, odf, xls, xlsx, ods, ppt, odp, pptx, opf, sav, its). Video in an unsupported format is automatically transcoded by Databrary into MP4 (H.264 + AAC) using ffmpeg (see **Technical Plan**). Audio is transcoded into MP3. Transcoded duplicates and original files are available for download. Some file types convey time-locked annotations of video (Datavyu: opf; ELAN: eaf), or audio files (CHAT: cha; LENA: interpreted time segments). With few exceptions (e.g., CHAT for speech transcription; BIDS for brain imaging), no standards exist for storing and sharing video and audio data, annotation data, and other data and metadata associated with human and non-human animal research. Some datasets have user supplied text-based tags or keywords that also do not conform to specific community standards.

Databrary encourages researchers to upload and share study protocols, IRB protocols, codebooks, data dictionaries, survey questionnaires, preregistration documents or links to them, and de-identified aggregate files from experimental tasks or surveys administered to research participants. These documents do not currently conform to any widely adopted standards. However, activities described in this proposal will make text information within these files searchable by the Databrary search engine.

In addition, Databrary supports uploading and sharing of participant-, task-, and study-level data and metadata using a spreadsheet-like user interface and stored as a JavaScript Object Notation (JSON) blob. Participant characteristics that can be entered into the spreadsheet include birthdate, test date, sex (categories unspecified), race (OMB categories, but free-text categories permitted), ethnicity (OMB categories, but free-text categories permitted), language, and disability status. Only dates are validated. The demographic variables do not currently conform to metadata documentation standards on the backend, but future versions of Databrary will support NINDS Common Data Elements (CDE)[75], embed machine-readable metadata using JSON-LD type schema components[76] and, follow recommendations consistent with the schema.org framework[77].

Metadata

In addition to primary research data, Databrary generates, stores, and shares metadata about datasets, individual users and their institutions, and project funders. Individual datasets (volumes) have system-wide unique integer identifiers. Dataset owners provide a searchable text-based description or abstract, links (URLs) to material related to a project and identify specific files—video or audio clips or images—as project-wide “highlights” that receive prominent placement on the volume page and are specifically searchable. If users provide funding information about a dataset, that data is also stored and shared alongside the dataset.

In turn, the application generates metadata about the number of participants, summary demographic measures, the level of participant data sharing release associated with specific data elements, file sizes and types, and dates related to when the data were uploaded and shared. Databrary makes public the names and institutional affiliations (but not email addresses) of all users who have institutional approval to access the system.

DATA ACCESS, SHARING AND ARCHIVING

All research data elements uploaded by users to Databrary are assigned a sharing release level[59] that defaults to *Private*. Sharing release levels determine who has (or will eventually have) access to specific files when the dataset owner(s) share a dataset to make it accessible to authorized Databrary investigators beyond the owner's immediate collaborators. Thus, *Private* files can only be read by users chosen by a dataset owner. In turn, dataset owners must specifically choose to make a file or set of files available to the broader Databrary research community, consistent with the provisions of the Databrary Access Agreement (DAA)[26].

The DAA is a formal, legally binding contract between NYU—Databrary's host institution—and an institution that agrees to authorize one or more researchers to access Databrary's restricted materials. Authorizing institutions must have an office of sponsored programs (or equivalent grants/contracts office) and an institutional review board (or equivalent ethical review board). Under the DAA, dataset owners cannot upload identifiable video or audio recordings unless they have sought and secured sharing permission from research participants, or the parents or guardians of minors (as required by their institutional review board). Databrary has published on the Databrary website recommended language and procedures for seeking sharing permission from participants and for documenting the data sharing permissions given by participants. The levels are *Unknown* (treated as Private), *Private*, *Shared* (with other Authorized Investigators on Databrary for research purposes only), *Learning Audiences* (shared with Authorized Investigators for research purposes and also permitting images or short videoclips to be used for teaching or research presentations), and *Public* (shared with anyone for any purpose). Except for the *Public* release level, shared data cannot be used for commercial purposes. These sharing release levels are communicated site-wide using a consistent set of icons, tool tip language, and definitions to ensure that all users understand what can and cannot be done with shared data. In almost all cases, *Shared* or *Learning Audience* release levels permit broad, unspecified research uses, if those uses have been reviewed and approved by the user's institutional ethics board. Metadata Databrary generates about datasets, authorized users, and funders are public.

Virtually all of Databrary's data and metadata can be accessed using API calls, many of which the *databraryr*[47] (and proposed *databrarypy*; see **Aim 4**) support. The application ensures that API calls return only data a user has permission to view. All authorized users must authenticate to Databrary before accessing restricted materials.

Databrary assets will be preserved indefinitely in a secure data center facility at NYU, Databrary's host institution, and mirrored on a server in upstate New York. Central IT staff 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 NYU ITS staff prior to converting formats.

The NYU Washington Square IRB and Office of General Counsel and Penn State Office of General Counsel provide ongoing guidance about ethical and policy-related dimensions of Databrary's data management and sharing practices.

TECHNICAL PLAN

SOFTWARE AND HARDWARE

Databrary 1.0

Databrary is a single-server web application running CentOS 7.0 on a virtual machine operated by NYU Washington Square IT. The Databrary backend uses GHC 8.0.2, Haskell 2010, Solr 6.6, PostreSQL 9.5, and FFmpeg 4.2.4; while the frontend uses AngularJS 1.4, CoffeeScript 1.7.1, JQuery 1.12.3, and Stylus 0.54.5 + nib 1.1.2. Databrary's 177+ TB of data storage reside in on-premise Isilon disk arrays managed by NYU Libraries. NYU's High Performance Computing cluster transcodes all audio and video files at no cost.

Databrary 2.0

With expertise from NYU Langone and NYU Washington Square IT staff and funds contributed by NYU and the PIs, we engaged Montrose Software to evaluate the entire Databrary 1.0 technology stack and estimate the time and cost required to rewrite and migrate the system to new, more permanent storage. Montrose concluded that Databrary's frontend and backend rely on outdated technologies. Bug fixes and new feature development have languished, as it is difficult to find developers willing and capable of working with the specific tools. Montrose recommended that we rewrite most of the Databrary application using Python, Django, PostgreSQL, Solr, FFmpeg/Elastic Transcoder, and Docker for the backend and TypeScript, React, SASS, Webpack, and NPM for the frontend. This would retain components of the technology stack that remain state-of-the-art for comparable systems (PostgreSQL, Solr, FFmpeg) while choosing new tools that should make future feature development, such as the features described in this proposal, cheaper and easier to implement. We began the rewrite in late April 2024 with the goal of replicating existing Databrary 1.0 functions and interfaces but adding only essential security-related new features (two-factor authentication, CAPTCHA on account creation). Montrose estimated that phase would take 6-8 months, meaning a migration in late 2024.

Architecture and Operating System

In a second phase of the Databrary 2.0 rewrite, we will rearchitect the system, adding multiple servers, and load-balancing to serve projected growth in the number of users and number of shared datasets. At the same time, we will migrate the application to a current version of Red Hat Enterprise Linux to bolster system stability and make it easier to administer and maintain.

Storage

NYU Libraries have requested that we migrate data to servers dedicated to Databrary functions as soon as it is practicable. Montrose has provided comparisons about the relative costs and benefits of on premise versus commercial cloud storage. NYU does not currently charge Databrary for data storage or data egress meaning that NYU subsidizes the full cost of data storage and egress for the entire Databrary community. One of the of biggest unknowns with the possible move to commercial cloud storage or compute (AI modeling; transcoding) services concerns the cost of data egress. Given that video downloads and streaming consume substantial bandwidth, it is essential to find solutions that do not result in unexpected and unbudgeted costs for uploading or downloading data. Initially, we expect to migrate Databrary 1.0 data to new on-premise servers. But we will monitor Databrary's current use patterns to make better projections about future growth. Final decisions about on premise vs. commercial cloud vs. mixed storage solutions (including possibly NSF ACCESS[78]) will be made during the rewrite process. We note that the flexibility of cloud storage and the proximity of compute resources in the cloud make hybrid solutions attractive. To ensure flexibility for the activities proposed here, the budget includes funding for storing 100 TB of data over a two-year project period.

UI/UX

The current timeframe and available funding for the Databrary 2.0 rewrite leaves no room or resources for redesigning user interfaces (UI) or improving user experience (UX), or for activities related to documenting the new system's components and introducing them to the research community. Most of these activities will be included in the new feature development described in this proposal.

TECHNICAL EXPERTISE

NYU

NYU Washington Square IT includes staff with decades of professional experience who provide technical support to Databrary 1.0, participate actively in the Databrary 2.0 rewrite process, and will manage the software development activities planned in this proposal. NYU IT experts who provide ongoing support to Databrary include Ken Yelton, Senior Director of Administrative Technology at NYU; Julian Quintero, Research Administration IT Director; Geby Varughese, Director of Cloud Engineering and Operations; and Stratos Efstathiodis, Director of Research Technology Services and the HPC. The project budget includes funds to offset some of the IT and project management costs.

Montrose Software

New Jersey and Poland-based Montrose Software is likely to extend its current consulting engagement with NYU/Databrary. Montrose has an extensive array of developers with decades of industry experience developing web applications for the financial services and medical fields.

TECHNICAL DOCUMENTATION

The proposed activities require the generation of extensive user documentation, both in web/document and video forms. Our team has extensive experience generating and publishing documentation using R Markdown and Quarto. These tools make it easy to generate version-controlled and shareable web pages and PDF documents. Working closely with the software team, PI Gilmore and the Data Curator will generate documents and publish them on the Databrary GitHub organization (databrary.github.io) with input from PI Adolph, NYU IT, and Montrose.

PROJECT SCHEDULE

Databrary application activities (Projects 1-3) are closely related, and the project schedule staggers their start and ending periods, accordingly. The API wrapper (Project 4) and curation (Project 5) projects are largely stand-alone although they use and to some extent extend the other projects. Figure 1 shows the relative timing and estimated durations for the proposed activities and the quarters when we expect to hold in-person or online workshops. The two-year timeframe may seem ambitious, but it reflects recent experience with our current development team.

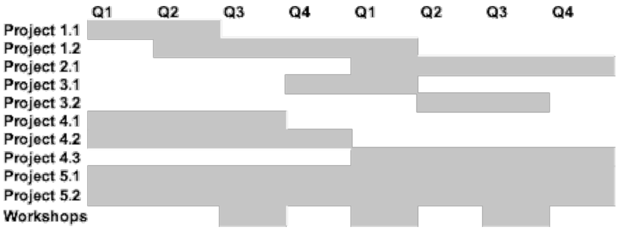


Figure 1: Project Timeline.

SUSTAINABILITY PLAN

Status

Databrary began operation in 2014 and has experienced steady growth in users and datasets, with a marked acceleration in use in 2018-2019 (See **Project Description: Figure 2B**). Databrary currently offers unlimited data storage and streaming services at no charge. NYU Libraries provide in-kind storage that costs at least $40-50K/year to sustain. Databrary runs on virtual machines at no cost; NYU levies no charges for bandwidth (0.1–1 TB/day in data egress); and NYU HPC provides transcoding services at no cost to the project. The PIs have supported software development, dev ops, and support staff salaries via project-specific grants and will continue to seek grant support (e.g., from Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support/ACCESS[78]). Databrary has also curated, stored, and shared some project specific datasets under specific grant subawards or through one-off data curation and deposit agreements. Nevertheless, the PIs and NYU agree that the current *ad hoc*, grant-centric, and F&A-based funding model cannot be sustained in the long term.

Future Plans

Consistent with operating practices in other peer data repositories (e.g., ICPSR, Dryad, FigShare), Databrary developed a business plan to recoup operating costs based on a mix of institutional and individual user subscription fees, data deposit charges, and per-project or hourly curation fees. The plan was developed in cooperation with NYU's Technology Opportunities and Ventures (TOV) unit. We deferred rolling out the subscription and mandatory data deposit fees until the Databrary rewrite project has been completed at the end of 2024 (see **Technical Plan**). Part of the rewrite involves adding hooks to the application that better support administrative accounts for user management and billing.

Beginning in early 2025 in conjunction with the re-introduction of the updated Databrary 2.0 application, we will implement the Databrary membership plan. The plan involves annual institutional membership subscription fees based on institutional size and number of users, accompanied by mandatory one-time data deposit fees based on the current cost of storing data for a 10-to-15-year window. Fees have not yet been set, but some illustrative amounts are $2,500-5,000/year for institutions and $2,500-3,750/TB for data.

Note: Data already stored *and shared* on Databrary as of the membership plan rollout date *will not be charged the data deposit fee*. However, we are evaluating how to handle unshared data which currently makes up the bulk of Databrary's storage footprint. In contacting dataset owners under Projects 5.1 and 5.2, we will start a dialogue about long-term plans for these data. If owners of *Private* datasets share the data, possibly with the curation assistance Databrary provides under this proposal, we would levy no data deposit charge. If the data remain *Private*, then we may charge the deposit fee or move the unshared data to less readily available, offline storage.

Steady-State Projection

Databrary once had and could readily justify now a full-time technical and support staff of 5 (salary + fringe of $700K/year), plus storage, compute, and bandwidth costs of $150K/year (increasing as usage increases). If 50% of the 795 participating institutions paid the minimum annual subscription, Databrary would net nearly $1M/year. If a one-time deposit fee had been charged on Databrary's existing 177 TB, that would have yielded $400-500,000. These projections give us confidence that a thoughtful, carefully implemented subscription and data deposit plan can make Databrary financially self-sustaining in a relatively short time. Moreover, we believe that the new features we describe in the current proposal *will substantially increase the value of Databrary to current and future users* and thus make the value proposition of membership even more attractive.