PROJECT SUMMARY

Overview:

The widespread capacity to share digital research data has transformed scholarship, but technical barriers of incompatible and proprietary file formats hinder the free exchange and full use of data. Video is a unique form of research data. For thousands of researchers in the behavioral sciences and related fields, video is the primary means of recording behavior in home, lab, classroom, museum, and public settings (1-9). Video captures the richness and complexity of behavior in real time and depicts how it changes across contexts, learning, and development. With minimal metadata, research video can be reused to answer questions outside the scope of the original project--questions never envisioned by the original investigators. Most researchers enrich video data by manually tagging segments with transcriptions, qualitative annotations, and user-defined categorical codes (10). Novel analyses could build on these tags if the data were openly shared, but many researchers do not yet share videos or coding files, and most coding tools use proprietary and incompatible formats.

Times are changing. Hundreds of researchers now use Databrary (databrary.org), an NSF- and NICHD-funded digital library for video sharing and reuse. The proposed project will enhance Databrary so that: (1) Video tags stored in incompatible software tools can be uploaded, imported, visualized, searched, and downloaded; (2) Researchers can browse codes produced by others, share coding files across geographically distant locations, and build on shared codes; (3) Detailed information about behavioral codes contained in separate manuals can be entered, indexed, visualized, searched, and downloaded; and (4) Researchers can search for specific datasets or video segments based on codes, and browse, select, and download matching datasets for analysis.

Intellectual Merit :

The project will add significant value to Databrary, an existing, next-generation resource for video-based behavioral research, by integrating detailed tags that have been manually applied to research videos by human coders. The new sources of data will be made accessible in a variety of formats for reuse, allowing researchers to choose the tools best suited to their research questions and analyses. Increased interoperability among coding files will enable investigators to address new multidisciplinary research questions based on videos already collected and shared in Databrary. By incorporating tags and definitions from coding manuals, the enhancements will enable Databrary users to readily seek, find, reuse, and build on shared video data. The project will thereby create new opportunities for integrative and multidisciplinary research that is at present prohibitively expensive or impossible.

Broader Impacts :

The project will have broad impact across fields in the behavioral, social, biological, and educational sciences that rely on video data. The proposed enhancements will enrich the datasets shared on Databrary--many of them funded by NSF and NIH--by integrating previously unavailable but valuable and sharable coding files. It will help to make shared video datasets more findable, accessible, interoperable, and reusable (11). The enhancements will expand opportunities for scientists at institutions with limited resources to participate in scientific discourse about behavior and its development. Because many of these institutions serve students from underrepresented groups, the project will expand research opportunities for them as well. By making data sharing more attractive to scientists, the project will increase the quantity and quality of shared video datasets. It will allow researchers to extract more value from existing and future shared video datasets. By making coding files and manuals more readily sharable, the project will improve transparency and boost reproducibility. Finally, the project will raise the profile of video-based research and bolster interest in and support for the behavioral sciences among the public.

TABLE OF CONTENTS

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

	Total No. of Pages	Page No.* (Optional)*
Cover Sheet for Proposal to the National Science Foundation		
Project Summary (not to exceed 1 page)	1	
Table of Contents	1	
Project Description (Including Results from Prior NSF Support) (not to exceed 15 pages) **(Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**	15	
References Cited	3	
Biographical Sketches (Not to exceed 2 pages each)	5	
Budget (Plus up to 3 pages of budget justification)	7	
Current and Pending Support	3	
Facilities, Equipment and Other Resources	2	
Special Information/Supplementary Documents (Data Management Plan, Mentoring Plan and Other Supplementary Documents)	2	
Appendix (List below.) **(Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)**		
Appendix Items:		

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

TABLE OF CONTENTS

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

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

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

PROJECT DESCRIPTION

Video is a primary means of recording behavior in home, lab, classroom, museum, and public settings for researchers in the behavioral and educational sciences and related fields (1–9). Video faithfully records the details and nuances of behavior in real time and depicts how behavior changes across contexts, learning, and development. With minimal metadata, video data collected for one purpose can be reused for a different purpose—to address new questions outside the scope of the original study, to examine phenomena outside the expertise of the original researcher, and to explore new possibilities never envisioned by the original researcher. Most researchers enrich video data by meticulously coding information about a subset of behaviors relevant to their research. Specialized software tools make it easy to apply text-based transcriptions, qualitative annotations, and categorical codes to user-selected segments of the video (10). The resulting coding files contain information about how the data were collected, who the participants were, where they were, what they were doing or saying, and when they did it or said it. Novel analyses could integrate across and build on this information if the videos and coding files were openly shared. But most researchers do not currently share video or coding files, and many coding tools use proprietary and incompatible formats.

Databrary (databrary.org) is a web-based video library funded by NSF and NICHD to enable sharing and reuse of research videos among behavioral scientists. Databrary stores and preserves videos in standard digital file formats. The shared videos are accessible, searchable, and reusable by a rapidly growing group of developmental researchers who are authorized by their institutions with oversight by their ethical review boards. Databrary is housed at New York University, securely protected by the university information technology services, and supported by the university libraries.

We now aim to expand the kind, depth, and diversity of analyses that researchers can conduct by enabling Databrary to store, share, and exchange coding files linked to shared videos.

Challenge

Many behavioral researchers still use paper-and-pencil to score videos or create make-shift coding files in programs such as Excel (10). Others extract information from video using specialized tools such as CLAN (12), Datavyu (13), ELAN (14), Mangold Interact (15), Noldus Observer XT (16), or Transana (17). These tools enable human coders to move backward and forward through digital videos at varied speeds and to apply *tags* to user-selected portions—speech transcripts, comments or other qualitative annotations, numeric ratings, or single words or letters that correspond to categorical codes. Tags are often time-locked to a particular video frame or segment so that information about the sequence and duration of events can be computed. Most tools support the application of a series of tags in separate *coding passes* that focus on different behaviors. The scored data are stored in *coding files*. The information in coding files forms the basis of subsequent quantitative and qualitative analyses. To ensure reliable, reproducible, and robust results, most researchers develop rich descriptions of the behaviors of interest that are captured by the various types of tags. *Coding manuals* saved in word processing or spreadsheet formats capture and organize these definitions. Taken together, the coding files and manuals contain invaluable, expensive-to-acquire, human-validated, text-based information about the contents of raw research videos.

With a few notable exceptions (CHAT, ELAN, Transana), the available coding tools store data in incompatible, proprietary, and non-interoperable formats. Exchange among them can be difficult or impossible. Researchers may need to hire technical experts to support collaboration with colleagues who use an incompatible tool or when an analysis workflow requires moving data to another tool. Those without access to technical expertise or a compatible tool may abandon

attempts to use the best-suited tool or to reuse shared data altogether. Few efficient and secure means exist for sharing videos with linked coding files; those that do exist support only a restricted set of file formats.

Opportunity

The videos linked to coding files and manuals generated by researchers constitute a substantial resource for new discovery with significant potential to catalyze novel research and to capitalize on prior research investments. This potential will remain unrealized until the materials can be made available in interoperable formats. Databrary has demonstrated that researchers will embrace video data sharing when barriers are reduced and appropriate incentives exist (see section below "Databrary overcomes most barriers to sharing video"). We believe this success will extend to coding files and manuals after we make it easy for users to share and exchange these files in whatever format best suits their purposes. The hard-won information contained in the coding files can then be made fully accessible and available for myriad new uses.

Accordingly, we propose to expand Databrary's capabilities to enable sharing of video coding files and manuals that are already part of the researcher's normal workflow, but not easily or readily shared. We propose to make the information currently stored in incompatible coding file formats interchangeable by making file formats interoperable with one another. Where the files contain metadata useful for search, we will import and add it to Databrary's search functionality.

These enhancements will accelerate the pace of video data sharing and reuse in the developmental and learning sciences where hundreds of researchers use video (2, 3, 18). The project will leverage and increase the value of NSF's prior investments in Databrary and create opportunities for new research in other fields that collect and annotate video recordings (4–9). This will enable new, multidisciplinary, multilevel, and integrative research on human behavior that is currently prohibitively expensive or impossible.

Project Aims

The project has four aims.

Aim 1: Enable transcripts, annotations, and codes from targeted video coding tools to be imported into and exported from the Databrary video library.

Databrary currently supports uploading, storing, and sharing some types of coding files but it cannot extract their contents or link them to videos. Working closely with a technical advisory committee (TAC) representing the leading academic and commercial video coding software tools (see Appendix) commonly used by the majority of developmental researchers (10), we will expand Databrary's understanding of and ability to work with the coding files generated by these tools. Databrary will develop ways for users to exchange files between formats and to link files with the original video source. Import functionality will bring the coding files into the Databrary system and make the information they contain available for visualization and search within a dataset and across the library. Databrary will also build the capacity to export codes back to the tools' native formats as well as to more user-consumable and interoperable formats (e.g., CSV).

Aim 2: Expand and extend Databrary's system for visualizing user-defined video tags.

Databrary's existing "timeline" interface for depicting tagged video segments will be modified to allow users to display, filter, and download the tags that have been applied to shared videos. Databrary will be expanded to store and represent information from multiple tags within the same coding file and tags from multiple files. The existing upload and download functionality will be enhanced to support flexible uploading, updating, and exporting of coding files in various

formats. This will make it substantially easier for researchers to distribute coding efforts on shared videos across geographically separate labs and it will maximize the transparency and reproducibility of tags applied to videos.

Aim 3: Allow users to enter, edit, index, visualize, search, and export coding manuals on Databrary.

Databrary currently allows users to upload and share coding manuals in word processing, spreadsheet, and text based file formats, but information in the files cannot be easily accessed or searched. We will build an interface for users to enter and share coding manuals with Databrary that creates electronic equivalents, which capture code definitions and make them searchable. This will help users to understand shared coding files and make it more likely that shared videos can be reused and previous findings built upon. Coding manual information will make Databrary's search engine return more relevant and useful information to users, and it will provide better, more human-readable information for visualizing codes while previewing recordings.

Aim 4: Enhance Databrary's search functionality to allow users to search for videos that meet specific criteria, based on tags, coding manuals, and other metadata.

Databrary currently allows users to search across datasets for selected terms. We will design, implement, and validate back-end technologies and user interfaces that return search results in ways that make it easy for users to find and preview matching video segments. The interface will allow users to discover relevant datasets for browsing or for future reanalysis. These features will make it easy for Databrary users to discover, explore, and reanalyze video datasets that meet the specific requirements of their research questions.

Results from prior NSF support

PIs Adolph and Gilmore and Co-I Millman received funding from NSF (BCS#1238599, funding period 2012-2014, no cost extension 2015-2016, $2,443,499; supplement BCS#1238599, funding period 2015-2016, $222,219) to support the Databrary project for video data sharing and reuse. Building research infrastructure was the primary focus of the prior award. In addition, we published several articles that describe Databrary (18–20) and how it relates to other "big data" initiatives in developmental science (21). We developed a policy framework for sharing identifiable research data, endorsed by more than 167 authorizing institutions. We expanded and maintained the Datavyu video coding tool (13), held workshops to train researchers to code video, and wrote about best practices in behavioral video coding (22). The current proposal builds on and extends Databrary as described in detail below. ***Intellectual Merit***. The investigators created infrastructure to enable sharing and reuse of research videos in an open-source (23) web-based repository, Databrary (24), expanded and maintained the Datavyu video coding tool (13), and fostered a growing community of researchers committed to video sharing and reuse (42). Databrary and Datavyu deepen and accelerate the pace of discovery in developmental science by enabling researchers to view one another's datasets, reanalyze them to test competing hypotheses, and address new questions beyond the scope of the original study. ***Broader Impacts***. Databrary empowers developmental scientists, especially from institutions with limited resources to support research, improves data management practices, and increases transparency in behavioral research. Our publications (18-22), regional and pre-conference workshops, and other venues with visibility are bringing this new, more communal view of developmental science to a larger audience.

Background & Rationale

Open data sharing has become a scientific imperative across disciplines and a mandate from research funders (25). It is common practice in many areas of biomedical (26), physical (27), biological (28) and earth sciences (29), and it is an emerging priority in neuroscience (30). Despite notable efforts to make data sharing a norm in the behavioral sciences (32, 33), most research on behavior remains shrouded in a culture of isolation (19). Researchers share interpretations of distilled, not raw data, almost exclusively through publications and presentations. The path from raw data to findings to conclusions can rarely be traced or validated by others, nor can other researchers easily pose new questions that build on the same raw materials.

The growth of video as data

Developmental researchers have long recognized the power of visual media to capture the richness and complexity of children's behavior (18). As video replaced film, it became the backbone of most developmental research programs for thousands of scientists who study learning and development. More than 100 of the 270 respondents to a recent survey of developmental scientists report collecting more than 5 hours of video per week (10) in their labs. The scale of some large collaborative projects is even larger. The Measures of Effective Teaching Project (34), funded by the Gates Foundation, generated more than 1,000 videos from 3,000 K-12 classrooms over a 3-year period. The data, constituting tens of terabytes of storage, are hosted at the University of Michigan (35) and streamed to registered viewers across the country. The NSF-funded HomeBank project (36), affiliated with the TalkBank/CHILDES archive (37), is collecting and sharing hundreds of hours of naturalistic audio recordings of children's speech, some of which will be accompanied by video. The Autism and Beyond Project at Duke University (38) has deployed an iPhone application that will collect video images of children's facial expressions to evaluate the feasibility of using computer vision techniques to screen children in their homes for developmental disorders and risk of mental illness. Clearly, the widespread availability of low-cost, high-resolution cameras has made video a large and rapidly growing source of information about human behavior.

Video enables behavioral science research, but poses special challenges

Video documents the interactions between people and their physical and social environment unlike any other form of measurement. It captures when, where, and how people look, gesture, move, communicate, and interact (2, 18, 39). Video closely mimics the visual and auditory experiences of live human observers, so recordings collected by one person for a particular purpose may be readily understood and reused for a different purpose. Nevertheless, capitalizing on the unique potential of large-scale video data collections requires overcoming a unique set of challenges.

Videos contain personally identifiable information; this poses problems for the protection of participant privacy. For some time, policies have existed for sharing de-identified text-based data (40). Video inherently contains identifiable information—faces, voices, spoken names, and interiors of homes and classrooms. Removing identifiable information from video severely diminishes its value for reuse and puts additional burdens and costs on researchers. Therefore, video sharing requires new policies that protect the privacy of research participants while preserving the integrity of raw video for reuse by others.

Large file sizes and diverse formats present technical challenges. Video files are large (one hour of HD video can consume 10 GB of storage) and come in various formats and sources (from cell phones to high-speed video). Many studies require multiple camera views to capture desired behaviors from different angles. Thus, sharing videos requires substantial storage

capacity, significant computational resources, and specialized technical expertise for storing and transcoding videos into common formats that can be preserved over the long term.

Video sharing poses practical challenges of data management. Researchers lack time and resources to find, label, clean, organize, link, and convert their files into formats that can be used and understood by others (41). Most researchers lack training and expertise in standard practices of data curation (20). Different coding tools represent the correspondence between video and coding files in tool-specific ways, or not at all. Few researchers reliably or reproducibly document workflows or data provenance. When researchers do share, standard practice involves organizing data after a project is finished, perhaps when a paper goes to press. This "preparing for sharing" after the fact presents a difficult and unrewarding chore for investigators, one that often exceeds the incremental cost and reasonable time frame contemplated under NSF's Data Sharing Policy (25). It also makes curating datasets a challenge for repositories (20).

Technical and practical challenges involved in extracting behavioral patterns from videos present barriers. Videos contain rich and diverse information that requires time-consuming work by human observers to extract. The extracted data are represented in specialized ways and are not easily exportable to other tools or statistical analysis software. In principle, other researchers could build on the videos and tags generated by others. In practice, most researchers do not share coding files, and coding files employ proprietary and incompatible data formats. The files are not easily or readily pushed along the analysis pipeline or shared with other researchers outside the original team. As a result, the hard-won, expensive-to-acquire human insights about behavior contained in research videos remain difficult to analyze and largely hidden to the greater scientific community.

Databrary overcomes most barriers to sharing video

Mindful of these challenges but motivated by the scientific promise of video data sharing, the PIs established Databrary, the first-of-its-kind library for storing and sharing video data and associated metadata. Databrary was created with support from NSF (BCS-1238599) and NICHD (U01-HD-076595) and has garnered additional funding from the Society for Research in Child Development. Databrary is a secure platform for storing and sharing research videos and associated metadata from studies in the developmental and learning sciences. It fosters widespread data reuse and enhances scientific transparency. Databrary has targeted the developmental and learning sciences community that is the PI's intellectual home. But, the team specifically designed Databrary to be adapted for and used by other researchers in the behavioral sciences.

Databrary began public operation in the spring of 2014. The library has since grown to encompass 275 authorized investigators and 161 affiliate investigators from 167 institutions around the world (42, 43). These investigators have contributed more than 3,600 hours of video or audio recordings, representing some 3,400 participants ranging in age from 6 weeks to middle age. The system stores 167 volumes or datasets, of which 56 are currently shared with the community of authorized researchers or with the public.

Databrary permits users to upload, store, organize, and share data with collaborators, the restricted community of authorized Databrary users, or the public, depending on the level of sharing permission granted by participants. Users may also search for, browse, view, and download videos stored on the site. They may view specific characteristics of videos such as participants' ages or recording context (e.g., home, lab, or school) for recoding and reanalysis. Databrary also empowers users to create, view, or download *highlights*—video excerpts that can be shown for educational or research purposes. Thus, Databrary supports sharing,

reanalysis, and pre- or non-research uses of video while solving some of the thorniest problems associated with sharing data that contain personally identifiable information.

Databrary's policies enable the sharing of identifiable data

Sharing research video requires policies to protect participants' privacy while sharing identifiable data. Databrary does not attempt to de-identify videos. Instead, Databrary maximizes the potential for reuse by keeping the content of recordings unaltered. To make unaltered videos available to others, Databrary restricts access to researchers who register and secure formal authorization from their institutions (44) and shares identifiable data only with the explicit permission of the participants. Databrary created template language (45) for seeking participants' permission to share data, which researchers may adapt for their own use. An online user guide fully describes these policies (46).

Unique among data repositories, the Databrary Access Agreement (44) authorizes both data use and contribution. However, users agree to store on Databrary only materials for which they have ethics board or IRB approval. Data may be stored on Databrary for the contributing researcher's use regardless of whether the records are shared with others. When a researcher chooses to share, Databrary makes the data available to the community of authorized researchers. More than 167 institutions in North and South America, Europe, Asia, and Australia have agreed to Databrary's framework, and the number grows daily (43).

Databrary overcomes technical barriers to video data sharing

To address the problem of diverse video formats, Databrary automatically transcodes each recording using NYU's high performance computing services into a common format suitable for web-based streaming (currently H.264+AAC in MP4 for video). The system maintains a copy in the original format for long-term preservation. To address local file storage limitations, Databrary does not currently place limits on the number or size of files that can be uploaded. As a web-based application fully compatible with modern web-browsers, Databrary does not require special software for access. Databrary's current assets total 17TB and are stored on NYU's central IT storage, which provides one off-site mirror and regular long-term tape backups.

Databrary's design overcomes practical barriers to sharing

Video requires little metadata to be useful, and only the participants' data sharing preferences are strictly required. Databrary developed a novel *active-curation* framework that reduces the burden of *post hoc* data sharing (20). The system empowers researchers to upload and organize data as it is collected. Immediate uploading reduces the workload on investigators, minimizes the risk of data loss and corruption, and accelerates the speed with which materials become available. Databrary employs familiar, easy-to-use spreadsheet and timeline-based interfaces that allow users to upload videos, add metadata about tasks, settings, and participants, link related coding files and manuals, and assign appropriate permission levels for sharing. To encourage immediate uploading, Databrary provides a complete set of controls so that researchers can restrict access to their own labs or to other users of their choosing prior to sharing. Datasets can be shared with the broader research community at a later point when data collection and ancillary materials are complete, whenever the contributor is comfortable sharing, or when journals or funders require it.

Active-curation poses few new burdens on a researcher's time beyond current practices while offering significant benefits. In effect, Databrary acts as a researcher's personal lab file server and cloud storage, enabling web-based sharing among research teams and ensuring secure off-site backup.

Furthermore, any de-identified data associated with a dataset, including demographic and study metadata, stimuli or displays, coding manuals, and coding data, may be shared openly, substantially broadening the availability of these materials. But, shared materials must also be made available in findable, accessible, interoperable, and reusable (11) formats in order to be maximally useful to others.

Remaining barriers this project will overcome

Despite Databrary's substantial advances in reducing barriers to sharing video data, significant hurdles stand in the way of widespread video reuse. Converting, sharing, understanding, retrieving, and building upon other researchers' coding files remains a chore. Essential definitions of behavioral tags that coding manuals capture are locked away. Most researchers don't share the manuals, and when they do share, the information is not linked to coding files and not indexed for searching.

So, we will enhance Databrary to allow the information contained in coding files and manuals to be uploaded, integrated with, visualized, shared, and downloaded from the library. This will help researchers make the most of their own coding files, allowing them the flexibility to choose among or move between tools. It will allow researchers to use Databrary to store and share coding files with collaborators, across the hall or across the globe, prior to sharing with the larger research community. By electronically linking tags with detailed code definitions, the enhancements will help researchers to train human coders to more efficiently and reliably capture subtle nuances in behavior. This will bolster transparency and reproducibility. By capturing, storing, and indexing tags and definitions, the system will help researchers to organize and manage multiple coding passes within and across studies. Moreover, once shared, the coding files and manuals will provide a foundation for other researchers to build upon. This will enable new, integrative, multidisciplinary research that is at present difficult, time consuming, and prohibitively expensive to conduct.

We describe implementation details in the next section.

Implementation Plan

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

Project 1: Import, conversion, and export of information in coding files

Project 1 focuses on importing coding files from the leading video coding tools used by developmental scientists, representing that information in Databrary, and exporting coding files in their native formats. Despite significant variation across video-coding tools, most behavioral tags consists of a start time, an optional end time, and an associated value which may be a string or other data structure. Coding files typically consist of one or more arrays of these tags, organized into separate, possibly hierarchical or nested coding passes. Databrary's existing data model includes an internal representation that allows users to apply tags to user-selected segments of specific videos. The tags may be multi-valued record data representing concepts such as depicted individuals or tasks. We will extend this foundation in building the coding-related capabilities that constitute Project 1.

As described in the Technical Plan, Co-I and Databrary Systems Architect, Simon, has extensive experience developing and deploying databases of the sort required to implement these features. Simon also has available to him the resources of NYU's Digital Libraries staff through the contributions of Co-I Millman, including experts in data repositories, databases, and system administration.

Project 1.1: Import and export Datavyu coding files

We will start by making Databrary capable of importing spreadsheets generated by Datavyu (13), the free, Java-based, multi-platform open-source video coding tool maintained by Databrary and supported by PI Adolph over many years. Some 6 volumes of datasets on Databrary representing 230 video sessions contain spreadsheets coded in Datavyu, and dozens of other datasets have Datavyu files that could be added. The Datavyu file format is well known to the Databrary team. Reading Datavyu spreadsheet files will be a demonstration case for projects (1.2-1.6) focused on other data formats. We will also build functionality to export codes in the Datavyu format. At present, users who download Databrary volumes containing videos with linked Datavyu files must manually re-link the video files to the coding files. By developing capabilities within Databrary to read and write information to Datavyu files, we will eliminate this time-consuming step. Datavyu has hundreds of users in the developmental community (10), so enabling Datavyu import and export should have a significant impact.

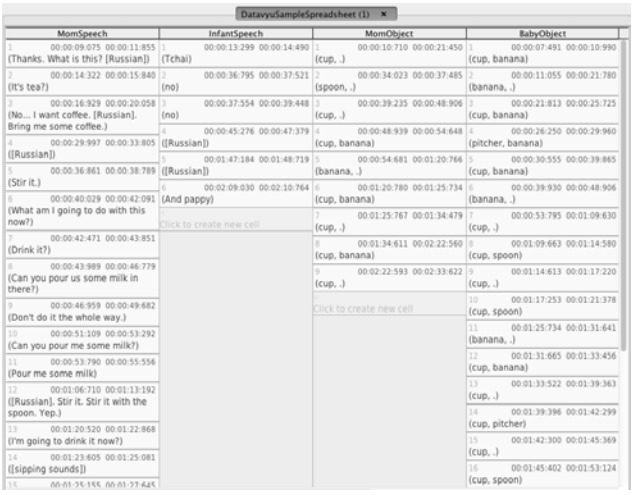

Figure 1: Illustrative Datavyu coding file.

Project 1.2: Import ELAN coding files

ELAN (14) is a free video and audio annotation tool developed by researchers at the Max Planck Institute for Psycholinguistics in the Netherlands. It has special features that enable researchers to encode language features from phonetics to pragmatics. Han Sloetjes, a lead developer on ELAN, has agreed to serve on the TAC and assist the project team to overcome hurdles involved in importing and exporting ELAN files. ELAN, like Datavyu, is built in Java, and the ELAN team has already consulted with the Databrary/Datavyu team on challenges both tools face in importing and playing diverse video formats. Import/export tools currently exist for converting ELAN files to and from the CHAT format used in CHILDES, TalkBank, and HomeBank (see Project 1.3 below). ELAN coding files include significantly more structure than Datavyu, including hierarchical and linked passes (called tiers), but we expect the basic coding information from ELAN files to be extractable, and simple versions to be exportable in the ELAN format. We will build on existing relationships and technical knowledge in developing ELAN import/export features for Databrary. According to our survey (10) and ELAN's internal data, hundreds of researchers in the developmental and language science communities use the tool and thus stand to benefit from the ability to share videos and coding files with Databrary.

Project 1.3: Import and export CHAT files

TalkBank (37) is a collection of language-related databases that include transcripts, audio, and video data from children and adults. Many of TalkBank's audio or video files are linked to text-based transcripts in the CHAT format (47), developed for the Child Language Data Exchange System (CHILDES) Project. CHAT files can be used with the Computerized Language ANalysis (CLAN) (12) suite of software tools, a recognized standard in the language community. CHAT files are especially valuable because the format encodes speech transcripts. CHAT is the

dominant data format among language researchers, and the format is a leader in interoperability with other language-related analysis tools. TalkBank's founder and director, Dr. Brian MacWhinney (Carnegie Mellon University), serves on the Databrary advisory board and has agreed to serve on the TAC. MacWhinney will work with us to enable CHAT-format transcripts linked with video or audio recordings to be imported into Databrary. CHAT files are well structured; the file specification is open and known; and we have already done some preliminary work with CHAT-format transcripts. However, CHAT files are unique in that transcription data are not directly linked to specific temporal codes. Rather, specific points in the transcription text can be associated with points in the video or audio time stream. For this reason, the CHAT format may have unique restrictions on which files can be imported and exported.

Support for CHAT-formatted transcripts within Databrary is essential for interoperability with the NSF-funded HomeBank data archive project (36). Homebank focuses on collecting a large corpus of natural speech using the LENA (48) recording device, linked with CHAT-formatted transcripts. Databrary PIs Adolph and Gilmore serve as consultants on the HomeBank project and work closely with its PIs, Dr. Anne Warlaumont (UC Merced), Dr. Mark VanDam (Washington State University), and Dr. Brian MacWhinney. The leadership of both projects has made interoperability among the datasets an important priority, given the considerable overlap in our research communities and the potential for multi-level, multidisciplinary research. We will build on our colleagues' expertise and contacts in developing CHAT-compatible features for Databrary. MacWhinney estimates that more than 100 investigators active on TalkBank use video, and we have several active Databrary users who have previously requested support for CHAT-formatted files. In addition to meeting the needs of a substantial existing community of researchers, enabling support for CHAT on Databrary will create an opportunity to enhance existing video datasets already shared with Databrary (49), which have CHAT-formatted transcript data available on TalkBank (50).

We will also pursue importing, linking, and storing LENA analysis files, which are automatically-generated analyses of the audio and language. LENA files are stored in an XML format that our HomeBank colleagues have begun to work with. These files are a critical part of the workflow of many language researchers prior to the production of CHAT files.

Project 1.4: Import and export files from Transana

Transana (17) is an open-source software package used by researchers in the educational, learning, and developmental sciences to analyze digital video and audio. Earlier phases of the Transana project received support from NSF and TalkBank. Dr. David Woods, Transana's lead developer, has agreed to serve on the TAC. Transana already enables the full export of project codes and related metadata into an open XML format. Interoperability with CHAT, ELAN and other language-specific file formats can be achieved via a third-party tool (51). Transana offers multi-user versions for lab groups who want to share and collaborate on coding audio or video files and a fee-for-service cloud storage feature. Its capabilities extend well beyond video coding features, but we expect that restricted import and export functions for Transana files can be added to Databrary. Transana's popularity among a large number of education researchers, including existing Databrary users, gives the project team confidence that the proposed enhancements will have an impact on a substantial portion of the educational research community.

Project 1.5: Import Noldus Observer XT files

Noldus Observer XT (16) is a commercial software package produced by Noldus Information Technology that is specialized for the collection, coding, and analysis of audio, video, eye

tracking, and physiological data streams. Noldus develops, sells, and supports complete integrated hardware and software packages for multi-measure behavioral analyses. Niek Wilmink, Product Portfolio Manager at Noldus, has agreed to serve on the TAC. Noldus Observer XT files are in a proprietary format, but the software supports exporting codes and short code definitions into a Microsoft Excel spreadsheet. We will work with Wilmink and the Noldus team to build support within Databrary for uploading and downloading coding files and templates. With support from Noldus, we will investigate the feasibility of importing Noldus files from their native formats and exporting them in native formats. In the worst case, Databrary will build functionality to import files that are exported by the Noldus software into an open or easy-to-read format, like the Excel spreadsheet format used by the coding template export function. Hundreds of researchers in developmental science, human factors research, ethology, and market research use Noldus, so the proposed Databrary enhancements will have widespread impact.

Project 1.6: Import and export Mangold Interact files

Mangold Interact (15) is a commercial video annotation tool used by more than 20% of respondents in our recent survey (10). Alongside its video collection, coding, and analysis software, Mangold develops, sells, and supports integrated hardware and software laboratories for video and audio-based behavioral analyses. Mangold has not yet agreed to join the TAC, but discussions are ongoing. Mangold Interact has several features the project team hopes to exploit in developing import and export functionality for Databrary. The software has a user-scripting function that enables coding files to be exported in a variety of formats for subsequent data analysis. It stores study/project-level metadata separately from session-specific data, and allows users to create coding templates that contain structured representations of coding definitions. We will first enable uploading and downloading of Mangold Interact files. We will then develop tools that import and export these coding files in an open, text-based (CSV or XML) format that the software currently supports. Finally, with anticipated support from the company, we will develop native file format import and export functions.

Beyond the project team's desire to serve an important segment of the research community, we believe that import/export functionality for Mangold Interact is important for another reason. One of the largest (N = 344 participants, 1,344 sessions) and most diverse video datasets currently stored and shared on Databrary comes from an NSF-funded longitudinal study led by Dr. Catherine Tamis-LeMonda (52), a member of the Databrary advisory board. Tamis-LeMonda has Mangold Interact coding files that can be shared on Databrary. Successful implementation of Mangold Interact import functionality should make it possible to store and share these files with Databrary alongside the already-shared videos, thus augmenting the current value of the data.

Project 2: Timeline interface for visualizing/manipulating coding passes

Project 2 focuses on improvements to Databrary that make it easier for users to visualize what tags have been applied to shared sessions and by which researchers, explore these tags in conjunction with the videos, access and export them, and upload new coding passes to their own sessions. After a user has selected a particular session of interest, Databrary will enable visualization of existing coding passes linked with that session. Databrary will support basic visualization and manipulation of these passes, so that users can preview the tags, select a subset of passes to export in their preferred format, or upload updated versions.

Project 2.1 User interface for displaying data about coding pass sources

Databrary has a timeline session viewer (Figure 2) that allows users to visualize temporal relations among multiple phases of a data collection (e.g., parallel data streams or multiple camera views). We will augment this interface to allow users to visualize the different coding passes that have been applied to a particular session. The interface will allow users to select or deselect specific coding passes for display and further manipulation. The interface will also support selective exporting of coding passes in the available formats along with the associated videos. We will adopt an iterative UI/UX design approach for this project and other UI/UX related efforts. We will generate requirements for the interface. We will develop a functional specification document with wireframes, have it reviewed by users, and then we will implement and continue to iterate the design. We will seek input from users, the TAC and from the Databrary Advisory Board about the implemented design, and then revise it as necessary.

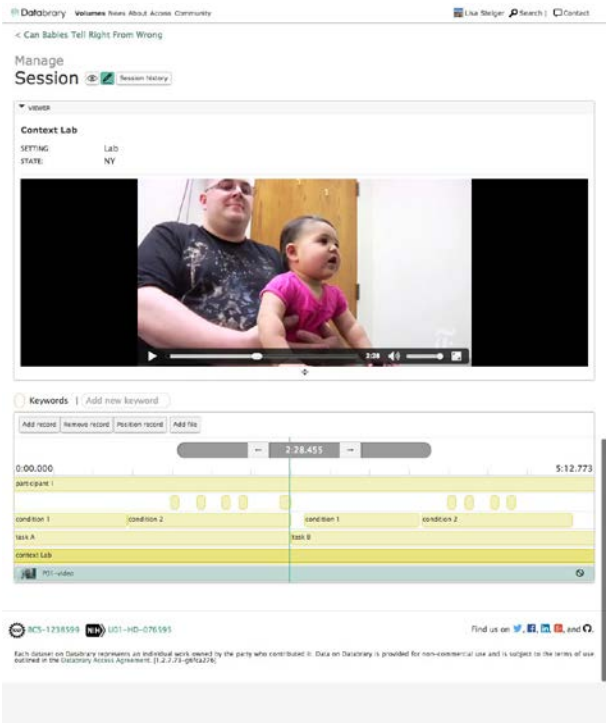

Figure 2. Illustration of current Databrary timeline session viewer.

Project 2.2 User interface for importing and exporting coding files

The existing upload and download functionality on Databrary's timeline session viewer will be expanded to allow more flexible manipulation of coding files. Users will be able to upload a new coding pass or replace an existing coding pass on sessions they can edit from any coding file on their computer. After selecting coding passes of interest, any user will be able to export those passes in a single coding file in the format of their choice, including various options for different representations in CSV format that may be imported to statistical analysis software.

Project 2.3 Improved full-volume download/export functionality

Users can already download packages from Databrary containing all the shared files in a volume. The download interface will need to be improved to allow exporting existing coding files in the users' preferred format rather than the originally uploaded format.

Project 3: Rich, text-based definitions of tags

Project 3 involves building Databrary's capacity to edit, represent, and export definitions of various types of tags in the form of coding manuals, and base these manuals on imports from a subset of coding tools that store code definitions internally. The definitions will be indexed by the Databrary search engine, and thus become searchable metadata. Sixty-three percent of developmental researchers who responded to our survey (10) report they "always" create detailed coding manuals that define behavioral codes associated with a particular study's set of

analyses. Respondents most often create these manuals using word processing or spreadsheet software. Consequently, we know that code definitions are widely available in electronic form and that many researchers want to share them. We need ways to capture and organize the current idiosyncratically formatted information.

Project 3.1 Make existing Databrary functionality for storing coding manuals more salient to users

Databrary currently allows users to upload materials that apply across a study (such as coding manuals, data files, and sample images) in word processing, spreadsheet, PDF, and other formats. Relatively few researchers use this capability now to upload coding manuals, but some may not understand how doing so enhances the value of a shared study to others. The team will make changes to the Databrary website text and design so that users are encouraged to upload coding manuals. We will also stress in our training and technical support, including the regional workshop series currently supported by NSF, the importance, value, and ease of sharing coding manuals with Databrary.

Project 3.2: Design back-end representation of coding manual information

Although coding manuals are useful, unless the information contained in them can be extracted and linked to videos, their value is reduced. Similarly, coding tools vary in the extent to which they capture and store code definitions within coding files or templates. For example, Datavyu does not currently store code definition information. CHAT-formatted files employ a structured set of well-defined keywords that describe the content of speech transcripts and other metadata such as the speaker and context and analyst comments. Transana, Mangold Interact, and Noldus Observer XT allow simple code definition information (e.g., “o = object”) to be stored within the coding files or in linked files, but the tools limit the amount of text stored in code description fields. Many researchers require much more detailed code descriptions, including elaborated definitions (e.g., *objects* must be detached from the surrounding surface and of a size that infants can hold in their hands), examples and exceptions, required speed for viewing the behavior, and precision required for the particular tag. As a result, even when a tool stores brief code definitions, researchers regularly augment them with more elaborate definitions contained in coding manuals.

Project 3.2 will enhance Databrary’s back-end data model to allow tag definitions and other information contained in coding files and coding manuals to be stored, shared, and indexed. We will modify Databrary’s data model to allow the additional coding manual text to be associated with particular datasets. For coding file formats that contain definitions, we will build on the coding file import functions developed as part of Project 1. Because parsing coding manual documents in varied and unstructured formats poses a serious challenge, we will start with a simpler approach. Tag definitions will be entered manually into a structured set of fields by means of a user interface we will develop in Project 3.3. Definitions contained in coding files will be imported and mapped to these fields internally.

Project 3.3: Design user interfaces for uploading and visualizing definitions contained in coding manuals

Developing an interface for users to enter code-specific information stored in coding manuals poses a challenge. We are collecting sample coding manuals from respondents to our community survey. These samples will shape the design of interfaces for uploading code definitions to Databrary. After code-specific information has been entered into Databrary, we will need to build interfaces that allow users to view and edit the information. In the best case, a

subset of this coding information will be exported along with tags to the data packages of the coding tools we support. This will depend on the extent to which we are able to read from and write to native file formats. The same iterative design and implementation process employed elsewhere will be used here.

With the proposed enhancements, researchers will be able to view tag definitions across shared datasets to evaluate their clarity and consistency and initiate conversations among colleagues about opportunities to achieve consensus around conceptual ontologies.

Project 4: Search interface and functionality

Project 4 will focus on designing, implementing, and validating the user interface that returns video search results in powerful, flexible, and informative ways. The bulk of the work for this project will revolve around ways to index and retrieve clips of sessions based on their associated metadata and coding data. This will require enhancing the existing Apache Solr-based search engine, or developing new indexing methods appropriate for overlapping video segments. The project will also develop the back-end and user interface components that allow users to view, explore, and download clips of the videos they find through search.

Project 4.1: Implement back-end search engine

Databrary's existing search engine indexes text-based information about datasets. It is based on the Apache Solr search engine by indexing entire volumes as documents. This engine will need to be enhanced or replaced. We need to index and search for entire datasets and also specific segments of videos within those datasets that match the selected criteria, based on available metadata and coding data in the associated coding files. The design of the search engine poses some complex problems related to matching multiple codes that may be overlapping based on search terms, linked coding manuals, and other metadata that will be informed by work on Projects 1-3. As a result, we plan to begin the design process early in the project but expect implementation to depend on progress in the other projects. We note that the Databrary team has at its disposal the expertise of Lee Giles, who sits on the Databrary Advisory Board, and Vasant Honavar who has agreed to serve on the project's TAC. Both of these computer scientists have extensive experience in search, big data, and data mining.

Project 4.2: Design search interface

Databrary's existing search interface returns and displays information about entire datasets: title, authors, description, and selected video highlights from that dataset. A search interface (see Figure 3) will need to be designed that returns information about which datasets or videos meet specific search requirements, including whether coding files are available and in what formats, whether coding manuals are available, and possibly specific codes within coding files. Ideally, the interface will allow users to preview videos, possibly by clicking on thumbnail images. The interface will allow a user to select and explore specific datasets or individual videos returned by the search.

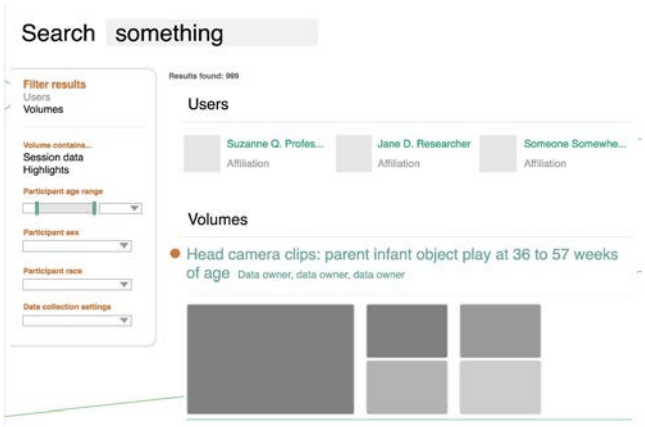

Figure 3: Wireframe illustrating possible user interface for search.

Coordination & Management Plan

The project will be overseen by PIs Adolph and Gilmore with guidance from Co-Is Millman and Simon. Simon also serves as Technical Director (see Technical Plan). Adolph and Gilmore currently meet by phone or video conference several times each week to discuss Databrary project-related business matters with the Co-Is and with project staff. Adolph and Gilmore collaborate on most decisions and regularly seek input from the Co-Is, but as overall Project Director, Adolph has the final say. Adolph and Gilmore also meet weekly with various members of the development team to formulate long- and short-term plans, get progress updates, and provide input. A project manager coordinates day-to-day operations.

In addition, the Databrary project as a whole has input from an advisory board. The board consists of experts internal to PSU and NYU and external advisers who bring expertise in data sharing and developmental science. The Databrary advisory board meets annually to hear project updates and provide guidance about policy and technical matters.

Because of the unique nature of this proposal, we have assembled a technical advisory committee (TAC) consisting of representatives from the leading video data coding software tools used by the developmental and learning sciences research communities (Appendix). The members of the TAC have agreed to provide technical assistance to the project by email, phone, or video conference for 1 hour per month. They have also agreed to participate in a half-day webinar meeting held annually in each of the three years of the project period.

Evaluation and Assessment

We will evaluate progress on the project in several ways. The PIs will report progress for each of the specific projects relative to the goals outlined in the timetable (Technical Plan) in the required annual project report. Among other metrics, we will report the number of Databrary users who are sharing coding files and manuals by coding tool used. We will also develop estimates of data reuse based on download statistics. The team will send out surveys to Databrary users once a year to solicit feedback about system operations, focusing on new features, and asking users the extent to which the new features change their willingness to share data, the ease of doing so, and the attractiveness of reusing others' data. With the cooperation of TAC members, we will survey the community of video coding tool users to ask similar questions. The results of those surveys will be summarized in the annual NSF report and discussed at the annual Databrary advisory board meetings.

Summary

Multiple barriers limit the widespread adoption of video data sharing and reuse in the developmental and learning sciences. In only a few years of operation, the large-scale, next-generation Databrary digital library has overcome many of these barriers, becoming an invaluable research resource for hundreds of developmental scientists. The current proposal will enhance Databrary, making it possible for contributors to upload, store, and share the results of human-generated annotations of video segments along with text-based code definitions. The data and metadata will accelerate the opportunities for users to reuse and build on the detailed information and analyses provided by other researchers. The enhancements will catalyze new types of data-intensive, crosscutting research across multiple fields in the social, behavioral, and economic sciences and allied disciplines where video collection is widespread.

For example, in the near future, researchers who study the quality of mother-infant interactions could augment their analyses with information provided by experts in movement science who code body posture and gesture and with transcripts provided by language experts about the kind and quality of mothers' speech in different emotional contexts. This will yield richer, more

integrative and complete understandings about the multiple dimensions of and influences on human behavior than are currently possible. By making transparent, visible, and easily shared the information hidden in coding files and manuals, this project will accelerate innovation and advance discovery across the behavioral sciences.

Intellectual Merit

The project will add significant value to Databrary, an existing, next-generation resource for video-based behavioral research by integrating detailed tags that have been manually applied to research videos by human coders. The new sources of data will be made accessible in a variety of formats for reuse, allowing researchers to choose the tools best suited to their research questions and analyses. Increased interoperability among coding files will enable investigators to address new multidisciplinary research questions based on videos already collected and shared in Databrary. By incorporating tags and definitions from coding manuals, the enhancements will enable Databrary users to readily seek, find, reuse, and build on shared video data. The project will thereby create new opportunities for integrative and multidisciplinary research that is at present prohibitively expensive or impossible.

Broader Impacts

The project will have broad impact across fields in the behavioral, social, biological, and educational sciences that rely on video data. The proposed enhancements will enrich the datasets shared on Databrary—many of them funded by NSF and NIH—by integrating previously unavailable but valuable and sharable coding files. It will help to make shared video datasets more findable, accessible, interoperable, and reusable (11). The enhancements will expand opportunities for scientists at institutions with limited resources to participate in scientific discourse about behavior and its development. Because many of these institutions serve students from underrepresented groups, the project will expand research opportunities for them as well. By making data sharing more attractive to scientists, the project will increase the quantity and quality of shared video datasets. It will allow researchers to extract more value from existing and future shared video datasets. By making coding files and manuals more readily sharable, the project will improve transparency and boost reproducibility. Finally, the project will raise the profile of video-based research and bolster interest in and support for the behavioral sciences among the public.

REFERENCES

1. Derry SJ, et al. (2010) Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences* 19(1):3–53.

2. Goldman R, Pea R, Barron B, Derry SJ (2014) *Video research in the learning sciences* (Routledge).

3. Alibali MW, Nathan MJ (2012) Embodiment in mathematics teaching and learning: Evidence from learners' and teachers' gestures. *Journal of the Learning Sciences* 21(2):247–286.

4. Masats D, Dooly M (2011) Rethinking the use of video in teacher education: A holistic approach. *Teaching and Teacher Education* 27(7):1152–1162.

5. Pasqualino C (2007) Filming emotion: The place of video in anthropology. *Visual Anthropology Review* 23(1):84–91.

6. Qualitative data repository. Available at: https://qdr.syr.edu/ [Accessed February 13, 2016].

7. Video sharing, deep tagging and annotation: A scientific archiving and demonstration tool. Available at: http://cmdbase.org/ [Accessed February 13, 2016].

8. Chaquet JM, Carmona EJ, Fernández-Caballero A (2013) A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding* 117(6):633–659.

9. Rautaray SS, Agrawal A (2012) Vision based hand gesture recognition for human computer interaction: A survey. *Artificial Intelligence Review* 44(1):1–54.

10. Gilmore RO, Adolph KE (2016) *Video use survey of ICIS and CDS listserv subscribers and Datavyu and Databrary users*.

11. The FAIR Data Principles - FOR COMMENT (2014) *FORCE11*. Available at: https://www.force11.org/group/fairgroup/fairprinciples [Accessed February 10, 2016].

12. CLAN. Available at: http://childes.psy.cmu.edu/clan/ [Accessed February 10, 2016].

13. Datavyu. Available at: http://datavyu.org/ [Accessed February 10, 2016].

14. ELAN. Available at: http://tla.mpi.nl/tools/tla-tools/elan/ [Accessed February 10, 2016].

15. Mangold Interact: The professional software for observational research. Available at: http://www.mangold-international.com/en/software/interact [Accessed February 10, 2016].

16. Noldus The Observer XT. Available at: http://www.noldus.com/ [Accessed February 10, 2016].

17. Transana: Qualitative analysis software for video and audio data. Available at: http://www.transana.org/ [Accessed February 10, 2016].

18. Adolph KE (in press) Video as data: From transient behavior to tangible recording. *APS Observer*.

19. Adolph KE, Gilmore RO, Freeman C, Sanderson P, Millman D (2012) Toward open behavioral science. *Psychological Inquiry* 23(3):244–247. doi: 10.1080/1047840X.2012.705233.

20. Gordon AS, 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(2). doi:10.7710/2162-3309.1238.

21. Gilmore RO (2016) From big data to deep insight in developmental science. *Wiley Interdisciplinary Reviews: Cognitive Science.* doi: 0.1002/wcs.1389.

22. Best practices for coding behavioral data from video. Available at: http://datavyu.org/user-guide/best-practices.html [Accessed February 19, 2016].

23. Databrary on GitHub. Available at: http://github.com/databrary [Accessed February 10, 2016].

24. Databrary.org. Available at: http://databrary.org [Accessed February 17, 2016].

25. National Science Foundation Dissemination and Sharing of Research Results. Available at: http://www.nsf.gov/bfa/dias/policy/dmp.jsp [Accessed February 14, 2016].

26. Kaye J, Heeney C, Hawkins N, Vries J de, Boddington P (2009) Data sharing in genomics — re-shaping scientific practice. *Nature Reviews Genetics* 10(5):331–336.

27. Young JR (2010) Crowd science reaches new heights. *The Chronicle of Higher Education*. Available at: http://chronicle.com/article/The-Rise-of-Crowd-Science/65707/ [Accessed February 10, 2016].

28. Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and opportunities of open data in ecology. *Science* 331(6018):703–705.

29. Kleiner K (2011) Data on demand. *Nature Climate Change* 1(1):10–12.

30. Poldrack RA, Gorgolewski KJ (2014) Making big data open: Data sharing in neuroimaging. *Nature Neuroscience* 17(11):1520–1527.

31. Poline J-B, et al. (2012) Data sharing in neuroimaging research. *Frontiers in Neuroinformatics* 6:9.

32. AERA Code of Ethics: American Educational Research Association Approved by the AERA Council. February 2011 (2011) *Educational Researcher* 40(3):146–156.

33. Nosek BA, Bar-Anan Y (2012) Scientific utopia: I. Opening scientific communication. *Psychological Inquiry* 23(3):217–244.

34. Measures of Effective Teaching Project. Available at: http://www.metproject.org/ [Accessed February 10, 2016].

35. Teaching and learning exploratory, University of Michigan. Available at: http://soe.mivideo.it.umich.edu/ [Accessed February 10, 2016].

36. HomeBank. Available at: http://homebank.talkbank.org/ [Accessed February 10, 2016].

37. TalkBank. Available at: http://talkbank.org/ [Accessed February 10, 2016].

38. Autism & Beyond. Available at: https://autismandbeyond.researchkit.duke.edu/ [Accessed February 10, 2016].

39. Curtis S (2011) “Tangible as tissue”: Arnold Gesell, infant behavior, and film analysis. *Science in context* 24(3):417–443.

40. U.S. Department of Health and Human Services (HHS) guidance regarding methods for de-identification of protected health information in accordance with the Health Insurance Portability and Accountability Act (HIPAA) privacy rule. Available at: http://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html [Accessed February 15, 2016].

41. Ascoli GA (2006) The ups and downs of neuroscience shares. *Neuroinformatics* 4(3):213–215.

42. Authorized Databrary Investigators. Available at: https://nyu.databrary.org/search?volume=false&f.party_authorization=4&f.party_is_institution=false [Accessed February 10, 2016].

43. Institutions with Authorized Databrary Investigators. Available at: https://nyu.databrary.org/search?offset=0&volume=false&f.party_authorization=5&f.party_is_institution=true [Accessed February 10, 2016].

44. Databrary Access Agreement. Available at: https://databrary.org/access/policies/agreement.html [Accessed February 10, 2016].

45. Databrary Participant Release Template. Available at: https://databrary.org/access/policies/release-template.html [Accessed February 10, 2016].

46. Databrary policies. Available at: https://databrary.org/access/policies.html [Accessed February 10, 2016].

47. MacWhinney B (2015) *Tools for Analyzing Talk – Part 1: The CHAT Transcription format.* Available at: http://childes.psy.cmu.edu/manuals/CHAT.pdf.

48. LENA Research Foundation. Available at: http://www.lenafoundation.org [Accessed February 10, 2016].

49. Demuth K (2014) Word-minimality, epenthesis and coda licensing in the early acquisition of English. doi:10.17910/B7B885.

50. Phonbank English Providence. Corpus Available at: http://childes.talkbank.org/browser/index.php?url=PhonBank-Phon/English-Providence/ [Accessed February 14, 2016].

51. Transformer. Available at: http://www.oliverehmer.de/transformer/ [Accessed February 14, 2016].

52. Tamis-LeMonda C (2013) Language, cognitive, and socio-emotional skills across the first 6 years in U.S. children from African-American, Dominican, Mexican, and Chinese backgrounds. doi:10.17910/B7CC74

53. ArXiv Business Support and Governance Model. Available at http://arxiv.org/help/support [Accessed February 22, 2016].

NEW YORK UNIVERSITY
FACILITIES, EQUIPMENT, AND OTHER RESOURCES

New York University and 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 administrative, research, and facilities resources. These include intellectual support from over 70 faculty affiliates in four working groups, administrative support in grants management, the hiring and retention of research personnel, communications support in drafting and disseminating findings, and facilities (e.g., offices for senior research personnel, offices or cubicle workstations for postdoctoral associates and graduate/undergraduate research assistants). IHDSC specializes in grants management support, including budgeting, reviewing and approving expenses, hiring personnel, and working with central and school-level university offices to ensure that each grant is managed in a responsible and timely manner. The Institute currently manages 45 active grants totaling over $45 million.

NYU Server Facilities

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

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

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

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

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

Networking: Over 375,000 feet of cable, 500 patch panels, 3,000 fiber strands and modules, 2,000 ft of ladder rack, 500 ft of cable basket and 1,000 ft fiber raceway and components throughout the data center. The South Data Center network connects into NYUNET using optical technology, forming a large Manhattan optical ring. NYU connects to NREN networks

such as Internet2 and National Lambda Rail, and to the commodity Internet at a “MeetMe” location in the building.

Network and SAN distribution: The MDF is the main connection point for all network and telecom services in the data center and houses Layer 2/3 networking switch equipment. From the MDF various types of data/telecom backbone and horizontal cables are interconnected to the intermediate distribution frame (IDF) cabinets and onto various server cabinets and IT equipment. A separate SAN distribution frame (SDF) serves the Storage Area Network (SAN) equipment to allow SAN network connections through IDF zone cabinets onto the SAN equipment. Redundant and diverse routes of structured cabling run throughout the space.

Command Center: The Command Center is staffed 24x365. The Building Management System allows the Command Center staff to monitor all power, cooling, network, and security for the facility.

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

NYU Digital Library Technology Services (DLTS)

New York University Libraries’ Digital Library Technology Services group (dlib.nyu.edu) collaborates with a wide range of scholars and curators to process, preserve, and enable access to digital materials in many forms. DLTS currently manages over 30 distinct digital collections, comprised of over 200,000 individual items, using over 150 terabytes of storage. DLTS has received past support from the Library of Congress (for web archiving and multimedia preservation), the Institute for Museum and Library Services (for collections and repository interoperation research), the National Endowment for the Humanities (for traditional collection digitization as well as papyrology collections and research tools, and the MediaCommons network of new-form scholarly publishing), and the Andrew W. Mellon Foundation (for media preservation research, media collections, archival collection management software).

NYU Office Space

PI Adolph has office and laboratory space provided by her home department at NYU. Co-I Simon has office space provided by IHDSC that is shared among the current Databrary staff. Millman has office space provided by the NYU library.

The Pennsylvania State University
Facilities, Equipment & Other Resources

Laboratory: Department provides lab space.

Clinical: N/A

Animal: N/A

Computer: Department provides computers for faculty.

Office: Department provides office space for faculty.

Other: N/A

Major Equipment: N/A

Other Resources:
Proposal Development and Grants Management: The Grants and Contracts Office offers an array of services to its faculty including, but not limited to, identifying funding sources, interpreting sponsor guidelines, proposal budget development, completion of forms, proposal editing, compiling the complete proposal, and obtaining the necessary university approvals.

When an award is received, the same grants and contracts staff assist faculty in the post-award administration of their projects by establishing grant accounts and overseeing the departmental budget clerks to process grant expenditures, provide updated reports to PIs, ensure that expenditures meet sponsor guidelines, and close out grant accounts.

DATA MANAGEMENT PLAN

1. Types of data produced

The project will collect data that consist of research video and audio recordings of behavior and text based tags of those recordings. These data will be in the form of video and audio files; information and metadata about the recordings in PDF, spreadsheet, word processing, image, and text files (TXT and CSV); and coding files containing annotations—in open formats such as .csv and proprietary formats—from commonly used desktop video coding software (CHAT, Datavyu, ELAN, Mangold Interact, Noldus Observer XT, and Transana).

2. Data and metadata standards

Allowing contributors as much flexibility as possible in what data formats they can provide is crucial to the success of the project. We will allow video, audio, text, and coding files to be contributed in a variety of formats, as provided by the users who are creating these data. We will transcode all deposited video and audio data into a standardized format (currently H.264 video codec, AAC audio codec in an MPEG-4 container for video). Access copies of these videos will be provided over the web via the native HTML5 video element. Data from desktop coding software and coding definition data will be exported both in their original file formats and converted to an open standard such as XML or CSV.

3. Policies for access and sharing

Data will be viewable and downloadable from Databrary only by authorized investigators who have been granted password-protected access. Researchers who wish to have access to the data must formally apply for access. Applicants agree to uphold Databrary's ethical principles and to follow accepted practices concerning the responsible use of sensitive data. Only researchers from institutions with Institutional (Human Subjects) Review Boards or similar review entities will be authorized for access. An official from an authorized investigator's institution must co-sign the Databrary Access Agreement (44). Full privileges are granted only to those applicants with independent researcher status at their institutions. Others may be granted privileges if they are affiliated with a researcher who agrees to sponsor their application and to supervise their use.

Ethics board or IRB approval is not required for non- or pre-research uses of Databrary. IRB approval is required to contribute data and for research uses. Once authorized, a user has full access to shared data on the site, and may browse, tag, download for later viewing, and conduct non- or pre-research activities. These policies are spelled out fully in an online user guide (46).

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

To support contributors in creating research data that may be easily shared, Databrary has extended the principle of informed consent to participate in research to include the act of sharing these data with other researchers. To formalize the process of acquiring these permissions, Databrary developed a Participant Release Template (45) with standard language

recommended for use with study participants. This language helps participants to understand what is involved in sharing video data, with whom the data will be shared, and the potential risks of releasing video and other identifiable data to other researchers.

Participants choose from four different levels of release: None, Shared, Excerpts, or Public. *None* implies that identifiable data may be uploaded to Databrary, but shared only with people selected by the data owner, usually members of a research protocol. Data that are missing a release level (e.g., participant wasn't asked, permission level was lost) are treated as *None*. *Shared* data may be shared with other authorized investigators and affiliates on Databrary. *Excerpts* means that photographic images or short audio or video clips may be shown by authorized investigators in public settings for educational or research purposes. *Public* data are available to anyone. Databrary automatically makes de-identified data about individual participants and metadata about datasets available to the public when a dataset is shared.

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

4. Policies for reuse and redistribution

Data will be made available for educational and research purposes. Access will be provided using the web-based Databrary application whose software is open source (23). Materials generated under the project will be disseminated in accordance with the policies of NSF and participating institutions. Publication of data shared on Databrary by users shall occur during the project, if appropriate, or at the end of the project, consistent with normal scientific practices. Users are provided the tools to cite data sources hosted on Databrary using an automatically-generated persistent identifier. No data may be redistributed outside the principles of the Databrary Access Agreement (44).

5. Plans for archiving and & preservation

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

SUSTAINABILITY PLAN

The first Databrary grant was funded by NSF in 2012, and current NIH funding runs through the spring of 2018. NYU libraries have committed to preserving the data stored on Databrary indefinitely beyond the end of grant-related funding for the project. We note that TalkBank (37), one of the most successful data repositories in the behavioral sciences, has been funded by competitive NIH and NSF grants for more than 30 years. Accordingly, we see that continued grant-seeking remains the most viable and promising means of sustaining and building the Databrary library over the short to medium term.

The current proposal, if funded, would support Databrary through the fall of 2019. We are planning new grant submissions to NICHD and to the Sloan and Gates Foundations, and we have other ideas about enhancements to the library and research projects based on the library's holdings that we will target to NSF, NIH, and other Federal agencies. The PIs have successful track records of seeking NSF and NIH funding, so we are optimistic about these prospects.

In the long term, we are working with NYU development staff to seek a private endowment fund for Databrary that would ensure resources for storage, maintenance, and development staff for an indefinite period. We estimate that Databrary could be made fully self-supporting with an annual budget in the range of $300,000-350,000.

We continue to work with entities like ICSPR, TalkBank, and others to advocate for long-term, stable funding sources for data repositories. Databrary is very early in its development, so we think that charging institutional or researcher-specific subscription fees is premature, but a possibility in the near term. We note that the ArXiv PDF preprint repository has had success with a model (53) mixing institutional contributions, foundation grants, and core support from the host institution. Another avenue we will explore will be contributions from professional and scientific organizations whose members benefit from Databrary's repository services. For example, a fee of $25/year per member could generate $125,000 from the estimated 5,000 members of the developmental science community (ICIS and SRCD memberships). Finally, we will explore ways for researchers with federally funded research grants to include Databrary storage fees into grant budgets based on the projected costs of storage and on staff/support costs needed to maintain the library. For example, a researcher who collects 5 hrs of HD video per week over a 3 year NSF grant could generate 7.8 TB of video data. Databrary's current internal cost for storage alone is $450/TB/year. So, a grant of this size could be charged $3,510 for the cost of storage plus an additional amount for the staff and support costs. NSF budgets are always tight, but a $4-5,000 fee for storage across a 3 year award period seems feasible.

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

TECHNICAL PLAN

Expertise

Technical aspects of the project will be carried out by Co-I Simon and the Front-end Developer.

Dylan Simon, Ph.D., Co-Investigator

Dr. Simon, Databrary's System Architect and Technical Lead, will spearhead the technical development of the project. Simon earned his Ph.D. in psychology from NYU and his Bachelor's degree in engineering and applied science from Caltech. Dr. Simon designed and implemented the Databrary digital library. He has expertise in a wide range of technologies critical to the project's success. Dr. Simon will devote 50% of his effort to building the proposed new features, and he will help hire, train, and supervise the Front-end Developer.

Front-end Developer (TBD)

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

Databrary Architecture

Databrary is an open-source web application (23) built in Haskell using wai/warp. The backend is a PostgreSQL relational database. Apache Solr supports search. The user interface is built primarily on the AngularJS JavaScript framework, and all data access is performed through a JSON API.

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

To enable upload, storage, and download of native format coding files, Databrary's current file upload and download capability will be upgraded. At present, the system only supports uploading, storing, and downloading of Datavyu (13) coding files. To enable import of coding files and the coding passes contained within them, we will modify and extend Databrary's video tagging feature. This feature currently allows users to select segments of video within Databrary's timeline user interface and to apply short text-based tags to the user-defined segments. Thus, the system's data model already provides the foundation for applying user-defined tags to video segments.

Databrary already enables users to upload, store, and download coding manuals in text, PDF, word processing, and spreadsheet formats. Rather than build tools that try to mine these idiosyncratically formatted files for relevant code definitions, we will build a data model that captures and organizes code definitions and a user interface that will be used to enter, display, and edit them. This type of backend will make it easier to import coding definition information from those tools that capture it in coding template files (Mangold Interact, Noldus Observer XT, Transana).

Communication with Target Community

Community engagement has been the focus of the Databrary project from the beginning. New Databrary features will be announced via the Databrary and Datavyu mailing lists, and with the cooperation of our TAC partners directly to their users. We will also gather feedback from users via electronic surveys to the same audiences. Project staff will provide email, phone, webinar, and in-person support to Databrary users, continuing current practices. Staff will demonstrate new features at training workshops, supported by NSF and SRCD, held at scientific conferences and in targeted regional locations chosen for their high concentrations of video-using researchers.

Schedule and timeline

Project 1.0: Coding file import. The sub-projects associated with importing coding files (1.1-1.6) have been allotted 3 months each and are planned to run sequentially beginning in September 2016. In practice, there may be aspects of the data import sub-projects that operate in parallel, some may take longer than others, and we may alter the implementation order.

Project 2.0: Timeline Interface. The Front End Developer will take a leading role in this work, scheduled to start in September 2016. Projects 2.1 - 2.3 are planned to take 3-4 months each.

Project 3.0: Coding manual import. Project 3.1 is scheduled for 2 months, and will begin at the start of Year 2, after the timeline interfaces have been designed. Project 3.3 also involves UI/UX work, and this work is scheduled to take 3-4 months. Project 3.2, the back-end modifications are scheduled to begin after the coding file import projects (1.1-1.6) are complete, in mid to late Year 2. The back-end modifications are slated to take 4-6 months.

Project 4.0: Search. Design work on search could begin early on, but will not be the primary focus of implementation and testing until Year 3, most of which is devoted to developing the search engine (4.1), scheduled for 6-9 months, and the search interface (4.2), scheduled for 4-6 months.

We will plan to hold TAC meetings in October of 2016, 2017, and 2018. Annual Databrary Advisory Board meetings occur in May.

DOCUMENTATION OF COLLABORATIVE ARRANGEMENTS

TECHNICAL ADVISORY COMMITTEE (TAC) ROSTER:

Brian MacWhinney, Ph.D.
Professor of Psychology and Modern Languages
Carnegie Mellon University
Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
http://psyling.psy.cmu.edu/
macw@cmu.edu

Vasant Honavar, Ph.D.
Professor and Edward Frymoyer Chair of Information Sciences and Technology Director,
Center for Big Data Analytics and Discovery Informatics
Associate Director, Institute for Cyberscience
Professor of Genomics and Bioinformatics and of Neuroscience
301A Information Sciences and Technology Building
The Pennsylvania State University
University Park, PA 16802
http://vhonavar.ist.psu.edu
http://ailab.ist.psu.edu
vhonavar@ist.psu.edu

Han Sloetjes
Lead Developer, ELAN The Language Archive (TLA)
Max Planck Institute for Psycholinguistics
P.O. Box 310
6500 AH Nijmegen
The Netherlands
https://tla.mpi.nl/forums/software/elan/
han.sloetjes@mpi.nl

Niek Wilmink
Product Portfolio Manager
Noldus Information Technology
Nieuwe Kanaal 5
P.O. Box 268
6700 AG Wageningen
The Netherlands
http://www.noldus.com
N.Wilmink@noldus.nl

David K. Woods, Ph.D.
Researcher, Lead Transana Developer
Wisconsin Center for Education Research
University of Wisconsin-Madison
1025 West Johnson Street, Room 345-A
Madison, Wisconsin 53706
http://www.transana.org
dwoods@wcer.wisc.edu