Data Coding, Analysis, Archiving, and Sharing for Open Collaboration: From OpenSHAPA to Open Data Sharing

Final Report
NSF Workshop held 15-16 September 2011
National Science Foundation Headquarters, Arlington, VA
Supported by NSF Award #1139702

Karen E. Adolph, New York University
Penelope M. Sanderson, The University of Queensland

Summary

On 15-16 September, 2011, Karen Adolph (New York University) and Penelope Sanderson (The University of Queensland) hosted a workshop at NSF headquarters titled, "Data Coding, Analysis, Archiving, and Sharing for Open Collaboration: From OpenSHAPA to Open Data Sharing." Participants included 35 researchers in developmental science, educational research, computer science, and cognitive science, along with program officers from NSF, the National Institute for Child Health and Human Development (NICHD), and the Institute of Educational Sciences (IES). Discussions focused on the promises and pitfalls of open video-based data sharing in developmental science. This report summarizes the results of those discussions and the subsequent activities it has stimulated.

Background

Behavioral development emerges from a confluence of factors interacting across multiple domains and timescales. Its study demands recording instruments that can capture this complexity and richness. Researchers studying behavioral development comprise thousands of psychologists, clinicians, movement scientists, and other developmental scientists. Most rely on video as the backbone of their research programs because, relative to other recording methods, video is cheap and easy to use, and in most cases only video faithfully renders the phenomena that researchers aim to describe and explain. Video coding—applying user-defined "tags" to events of interest—transforms the behaviors captured on video into quantitative data for statistical analysis.

Traditionally, developmental scientists have not engaged in open sharing of video data [1]. This culture of isolation arose for historical, legal, ethical, and pragmatic reasons, and past reliance on analog video limited widespread data dissemination. More recently, various data sharing efforts have set precedents for addressing issues such as participant anonymity and researcher attributions [2–6], and digital video overcomes the dissemination problem. However, sharing video presents its own unique set of challenges [2,3]—establishing an infrastructure for storing and accessing data; creating tools for tagging videos to enable exploration and analysis; and fostering an academic culture of transparency and collaboration that embraces the benefits offered by open video data-sharing.

Open sharing of digital information is both a scientific imperative and common practice in the biomedical [7], physical [8], and earth sciences [9], and it is an emerging mandate for federal funding in the behavioral sciences. NSF's National Science Board encourages "individual scientific communities to establish data-sharing and management practices that align with NSF data policies" [10]. Similarly, the NIH exhorts that "all data should be considered for sharing." [11]

With the encouragement of NSF program officials, we convened a workshop to address two primary questions:

(1) What are current barriers to open video data sharing in the developmental sciences?
(2) How might these barriers be overcome?

Participants and Schedule

Participants and their affiliations are listed in Appendix A. The original workshop schedule is listed in Appendix B. The mix of senior and junior researchers provided a wide variety of perspectives during the many opportunities for discussion built into the schedule.

Organizers were Adolph and Sanderson. Clinton Freeman (from UQ) and Jesse Lingeman (NYU) helped to envision an initial plan for open video data sharing in the developmental sciences based on the OpenSHAPA video coding tool. Freeman (the systems architect) and Lingeman (software engineer) collaborated with Adolph and Sanderson on the development and use of OpenSHAPA.

Senior participants were selected to cover key areas in developmental science (Alibali, Aslin, Bornstein, Galloway, Goldstein, Lockman, Messinger, and Smith), data sharing and archiving (Altman, Bertenthal, Davis-Kean, MacWhinney, and Stamper), and data visualization and human-computer interaction (Borner, Gray, Hoffman, Quek, Wong, and Yu). Three senior participants (Lockman, Quek, and Aslin) also served as discussants charged with providing an overview and perspective on issues discussed at the end of the workshop.

Additional participants included junior faculty members, postdoctoral researchers, and doctoral students. The junior researchers brought first-hand experience of data coding and data management needs to the workshop, as well as distinctly contemporary views on the technical possibilities for data analysis and sharing.

Dr. Mark Weiss, Director of the NSF's Behavioral and Cognitive Sciences Division, opened the workshop. Weiss emphasized the timeliness of the workshop for NSF given its growing emphasis on data management practices.

Day 1. The workshop began with an introduction by the OpenSHAPA development team (Sanderson, Adolph, Freeman, and Lingeman). We outlined the workshop goals, discussed how video coding and exploration tools can help researchers to get more out of their data, demonstrated OpenSHAPA, and presented a vision for how open video data-sharing might be supported with the appropriate cyber infrastructure and data management tools.

Smith followed with a data driven description of how micro video analyses and data mining can reveal mechanisms to explain macro-level constructs. Stamper provided an illustration of the mature data analysis and data sharing environment developed in the University of Pittsburgh/Carnegie Mellon University Science of Learning Center. In the afternoon, Messinger and Yu presented examples of data mining, analysis, and visualization, using plugin architectures, and Borner described how data exploration and visualization tools could transform scientific thinking. Bornstein discussed the different issues and needs involved in traditional collaboration versus open data sharing. Bertenthal described the SIDGrid data-sharing project and emphasized the importance of community buy-in and participation. Altman discussed issues of acknowledgment, attribution, citation, and data archiving and long-term preservation. Finally, Davis-Keen outlined professional and technical issues around data sharing and secondary data analysis. Issues of permissions, access, and acknowledgment were emphasized.

Day 2. The day opened with MacWhinney's description of the TalkBank data-sharing project, including a history of CHILDES, a well-known data sharing effort in the developmental

behavioral sciences with a 30-year history of success. Alibali and Goldstein presented examples of alternate academic video coding tools—Transana and ELAN—for exploring and analyzing multiple data streams. Galloway described the integration of disparate data sources—video, EMG, force plate data, parental reports, and so on, using the OpenSHAPA tool. The need for interoperability among analysis tools was highlighted. Then Hoffman, Gray, and Wong provided a view from cognitive science and cognitive engineering and presented examples of the difficulties of working with others' analyses, the achievement of end-to-end data analysis and modeling, and tools for extracting themes from data.

Before the lunch break on Day 2, workshop attendees identified a set of critical issues to resolve for an open video data-sharing system based on the OpenSHAPA tool, for later plenary discussion by attendees. These issues included the kinds of data to be shared; obtaining participant permissions for sharing; constraining access by users; acknowledgments and citation of the original investigators; the timing of when researchers could contribute data; video coding, exploration, visualization, and data management tools for maximizing extraction of information from videos; interoperability among tools; pros and cons of standardizing metadata; training users; and building a community of contributors and users.

The workshop closed with summaries by the senior discussants. Aslin reviewed the motivations and obstacles to open video data sharing. Lockman described a variety of use cases for shared data and emphasized the promise and importance of getting more from data already collected and coded. Quek summarized technical obstacles to sharing and possibilities for overcoming these obstacles.

Presentations from workshop participants are posted on the OpenSHAPA website [12].

Workshop Highlights

In this section we highlight the key issues discussed in the workshop. The strong consensus of the participants was that the barriers to open sharing of video-based data sharing were surmountable, potential benefits were many, and that the timing is right to proceed with a data sharing initiative in developmental science.

Nature of Developmental Data

Developmental scientists ask fundamental questions about the origins and causes of complex behavior—who we are and how we got this way. The field is broad and varied, spanning topics in cognitive, social, linguistic, perceptual, emotional, and motor development. Research designs range from naturalistic home studies to experimental lab paradigms, and from basic research with typically developing children to applied studies and clinical interventions. Ages cover the entire lifespan. Sampling may be once or multiple times. Yet, embedded in this breadth and diversity are commonalities. Developmental scientists focus on understanding change, and to do so, they record and interpret behavior over time, typically using video.

What kinds of data should be shared? Rather than limiting the library contents to a particular domain of development, the consensus was that the shared data should represent the diversity of work in developmental science. This diversity will help to support increased transparency in research and to speed progress in developmental science. The innovation of this effort is the emphasis on open sharing of raw video data, but to make the data maximally useful, shared studies should also include relevant metadata including code books, coded spreadsheets, and resulting manuscripts.

Challenges of Video

Video, the most common medium for recording behavior, has unique challenges and virtues [2,3]. Video often accompanies other data streams (physiological recordings, brain imaging, EMG, motion tracking, eye tracking, verbal transcripts). Multiple data streams—including two or more camera views—require tools for synchronization and integration and benefit from multivariate time series analyses.

Many labs lack the resources to adopt tools for detailed video coding and complex analyses. Therefore, tools should be free and open source. The bar for entry should be low enough to encourage participation from labs without extensive technical staff or training, while the tools should be sophisticated and flexible enough to suit a variety of research needs. Furthermore, the widespread use of video creates a data explosion: Our investigations after the workshop revealed that the typical developmental lab collects 8-12 hours of video per week in widely varied formats [13]. Thus, sharing digital video requires substantial storage capacity, powerful search and streaming tools, and significant computational resources for transcoding videos into common, preservable formats.

Why Share Video Data?

As stated by the NIH, "Everyone benefits [from sharing], including investigators, funding agencies, the scientific community, and, most importantly, the public. Data sharing provides more effective use of resources by avoiding unnecessary duplication of data collection. It also conserves research funds to support more investigators. The initial investigator benefits, because as the data are used and published more broadly, the initial investigator's reputation grows." [11]

Researchers will be motivated to ***contribute*** data because a history of data sharing and a plan to commit data to an open repository will enhance the likelihood of federal funding. Their work will receive more attention and citations by users [14], and their data and tools will survive in useable form beyond their lifetimes. Researchers will be inspired to ***use*** data in the repository because effective data sharing can transform discovery in developmental science. More rapid progress could be made if developmental researchers could point readers and reviewers to raw video data that illustrate procedures and findings; if users could browse for exemplars to stimulate new work; if researchers could gather preliminary data, expand samples, run replications, examine cohort effects, and assess effects of geographic location or population by using data in a shared archive; and if instructors could search for suitable examples to illustrate methods and findings to their students. We could lay a firmer foundation for the science if there were true transparency such that researchers could view each other's methods [15]. Open sharing would allow researchers to extend prior work by building on each other's codes and to score and analyze video files in ways unimagined by the original researcher. Researchers could use tools contributed to the archive to enhance understanding of their own data and use data to test their analytic tools. They could collaborate with like-minded researchers in a sub-area of development to create corpora with shared coding schemes and analysis tools as exemplified by repositories such as TalkBank.

Nevertheless, despite these virtues, the discussion at the workshop also highlighted barriers to data sharing, concerns about misinterpretation of shared data, issues with data attribution and citation, and what appropriate incentives to data sharing might be. Some of the issues were integrated into concept maps by participant Hoffman (see Figures 1, 2, and 3).

An important point made by agency personnel attending the workshop was that communities should determine their own conception of data sharing and initiate their own practices before a model is imposed by agencies. Further discussion revolved around whether open data sharing should evolve organically as individual researchers release data to a sharing environment, or

whether a core group of influential senior researchers should commit to open data sharing and so create an initial “critical mass” of open data-sharing to attract others. The latter has been the model on which MacWhinney’s CHILDES project [5] works and has been a key to its success for the child language community.

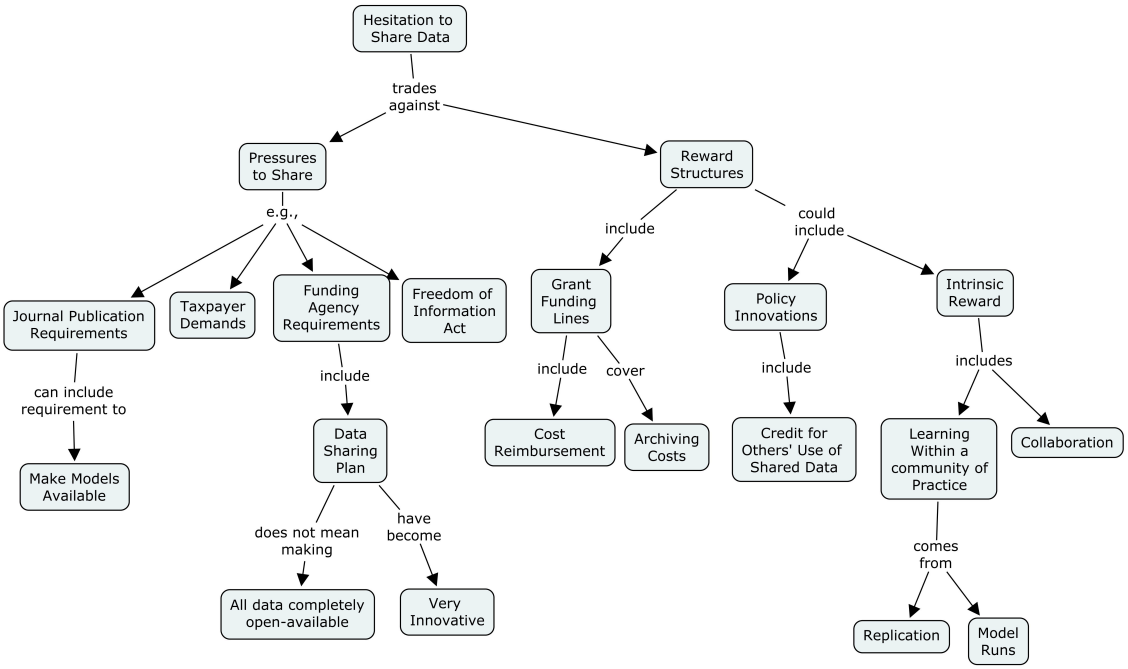

Figure 1: Concept map based on workshop discussion around researchers’ motivation to share data. Courtesy of Robert Hoffman.

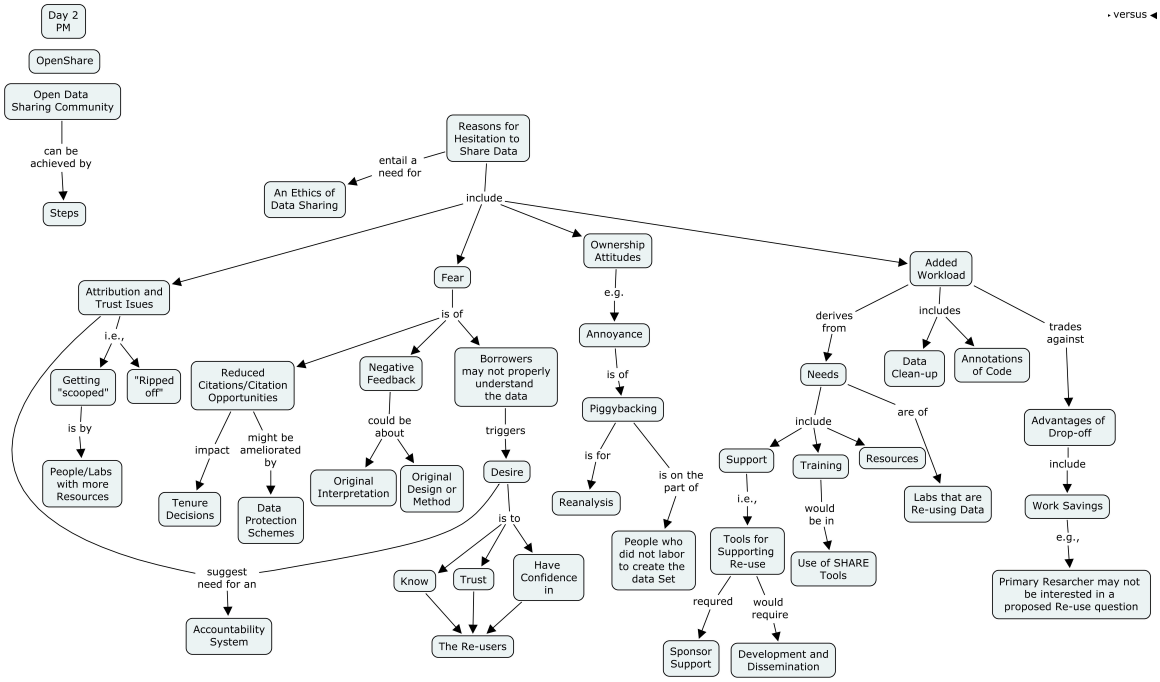

Figure 2: Concept map based on workshop discussion around reasons for researchers' hesitation to share data. Courtesy of Robert Hoffman

Precedents for Data-sharing

Data sharing of text and image files, and to some extent of video files, is not new. Examples in neuroscience include the Biomedical Informatics Research Network (BIRN) [16] and the International Electrophysiology Epilepsy Data Portal (IEEG) [17]. In the behavioral/social sciences, four representative projects are the following:

- The Inter-university Consortium for Political and Social Research (ICPSR) [18]
- TalkBank and CHILDES [19]
- The Social Informatics Data Grid (SIDGrid) [20]
- The DIVER project [21]

The first three were introduced and discussed at the workshop, whereas the fourth came to our attention afterwards. For 50 years, ICPSR has housed and curated quantitative, demographic, and sociological data related to education, aging, criminal justice, substance abuse, terrorism and related fields. It is affiliated with the Institute for Social Research at the University of Michigan and offers membership to institutions. For 30 years, the TalkBank/CHILDES repository at Carnegie Mellon has housed and curated a set of databases with transcript, audio, and video data relevant to the study of language and developed transcription and analysis tools. SIDGrid is a computing infrastructure that provides integrated computational resources through the TeraGrid for processing multimodal data on social communication. It combines data collection, storage, coding, analysis, and sharing, and represents communication in real-time combining different data types. DIVER, developed at Stanford, is a tool for viewing and annotating video and sharing it with others over the web. It is important to build on the successes of these efforts.

Challenges of Sharing Video Data and Associated Metadata

We discussed the challenges of sharing and accessing video data at length. Many of the discussions continued after the workshop, as we worked with colleagues to develop research proposals offering resolutions to these challenges. Some of the fundamental challenges covered at the workshop are outlined below.

Permission, access, and security. Sharing video poses special ethical problems to contributors and users. Video data cannot be made anonymous without reducing information content. Concealing participants' identity (e.g., by pixilation or blocking faces) would limit the utility of the data. Given these restrictions, researchers risk violating participants' privacy if digital images are viewed or released to the public without authorization. Before data sharing can happen, methods need to be developed to provide access to video by authorized users while protecting participants' confidentiality.

Licensing. There also needs to be guidelines for how shared data can be used by researchers. We discussed data licensing arrangements such as the creative commons licenses [22] that permit users to "remix, tweak, and build upon" the work for non-commercial uses so long as the original data depositors are credited and the new work is licensed under the same terms.

Attribution/citation. Most researchers will only contribute the raw data from a study after completing data collection and a manuscript is submitted or in press. It is important to develop practices for acknowledging and citing the researchers and organizations that originally invested resources in the contributed data that is later used in further analyses.

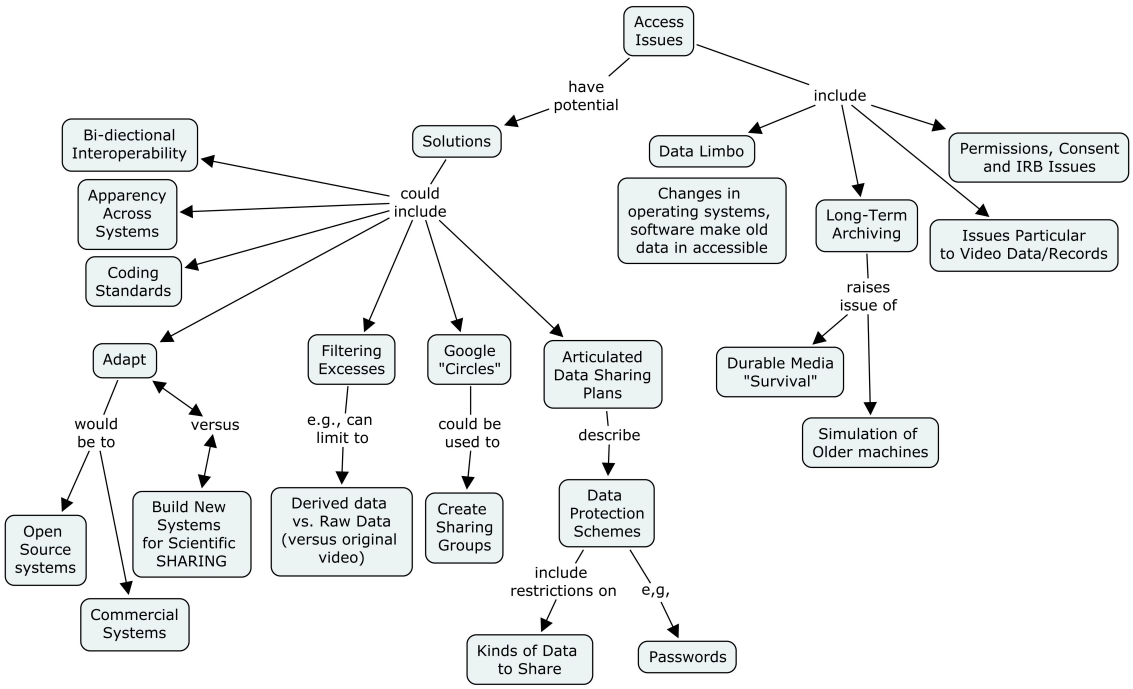

Figure 3: Concept map based on workshop discussion around issues relating to how researchers would manage appropriate access to shared data. Courtesy of Robert Hoffman.

Data coding tools. Commercially distributed video coding tools include Mangold Interact, Noldus Observer® XT, and StudioCode. Their strengths include technical support, richly featured mature software bases, and extensions for specialized analyses. But it was recognized that they have important limitations. They are expensive; users cannot alter fundamental functionality; software upgrades are commercially driven; and proprietary formats impede sharing files across coding tools.

Several non-commercial coding tools are available under open-source licenses, and have been used by workshop attendees. ELAN [23] is free and Transana [24] charges a nominal fee. Like many academic tools, ELAN and Transana are specialized in certain research domains (language and education research, respectively) and do not easily convert into other formats.

OpenSHAPA [12] is a free, platform-independent, open source tool developed by the PIs, inspired by MacSHAPA, which was developed during 1990-1994 with intermittent upgrades until the early 2000s. OpenSHAPA is a general-purpose tool that offers powerful features and high-level flexibility through a scripting interface, and is a logical foundation for an open-data repository for developmental science. We discussed the capabilities of OpenSHAPA, and workshop attendees suggested directions that its development should take (see Figure 4). The discussion focused on data management, visualization and plug-ins, interoperability between OpenSHAPA and related tools, and questions relating to creating an OpenSHAPA community of users. The potential to specify and document data analysis workflow was seen as a key data management strategy to support data sharing. Much discussion focused around describing OpenSHAPA's current plugin architecture and OpenSHAPA's use of the Ruby scripting language to help researchers to build their own data management practices. The interoperability of databases was seen as essential: OpenSHAPA would need to be able to read in databases from other

data coding tools if it is to become a unifying tool for data sharing. Finally, we discussed barriers to entry, the need for OpenSHAPA training workshops, and online community building.

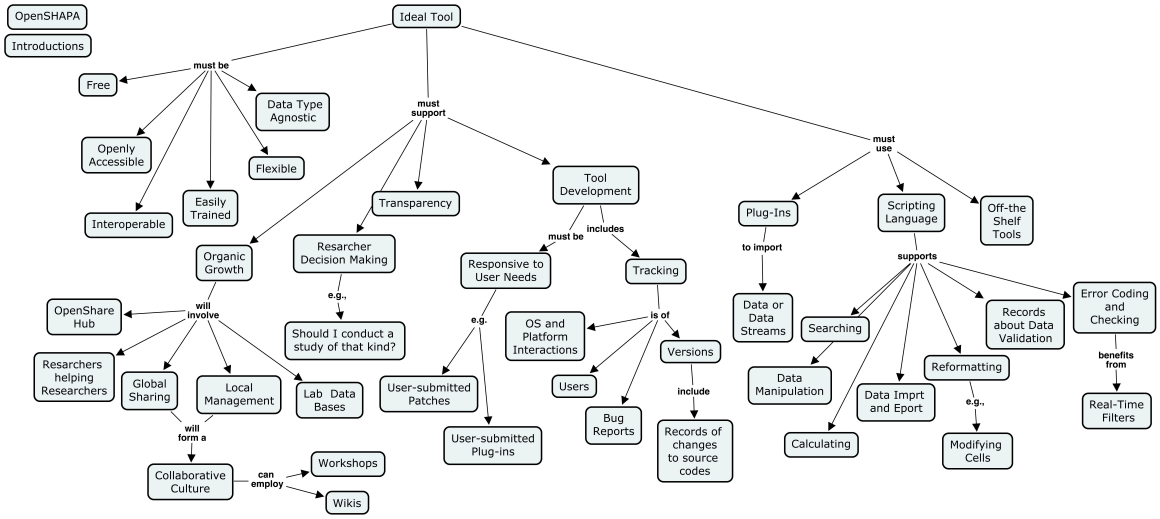

Figure 4: Concept map highlighting main points of the discussion around an ideal video data coding and analysis tool. Courtesy of Robert Hoffman.

Types of metadata. One important issue is how video or other data can become “discoverable” when in a large shared library. We discussed different approaches to data curation. In developmental science, behaviors of interest vary across studies and few tags are universally used, except for children’s age and sex. Thus we must trade against the need to maintain flexibility given the diversity of work in developmental science and the benefits of standardization.

Nonetheless by exploiting the metadata that typically accompany developmental studies (see Figure 5), we may be able to keep manual curation to a minimum, as shown in other fields [25]. Typical developmental studies include global metadata (manuscripts, protocols, codebooks, etc.) that are already sufficient to guide search for appropriate datasets. The raw data typically comprise multiple video files and associated metadata. OpenSHAPA, for example, creates spreadsheet files linked to each video file that include time-linked tags to events of interest, study design information, and participant demographics. These tags can also be used to search for appropriate data sets and to browse their videos. Thus, the challenge of searching a shared repository of developmental data can potentially be mitigated by combining the existing, investigator-defined metadata with powerful search, data mining, and meta-analysis tools such as those deployed in the neuroimaging community [26,27]. The challenge of keeping contribution time and cost to a minimum can be met with the appropriate data management tools.

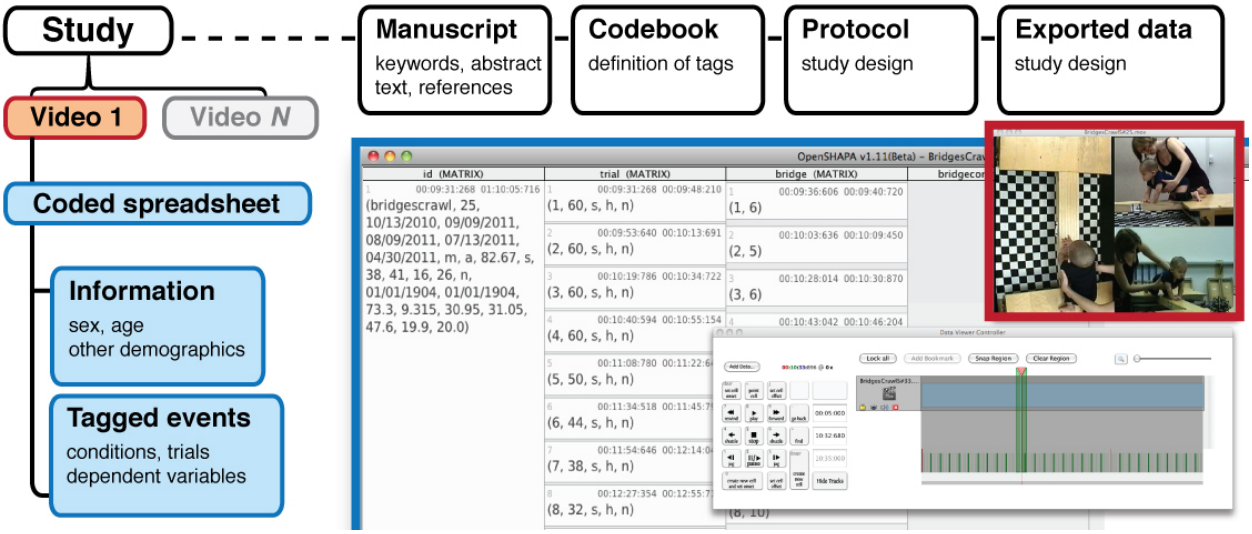

Figure 5. Depiction of relationships among raw video data and metadata. Each study typically produces global forms of metadata (manuscripts, codebooks, protocols, etc.). Within each study, each video file is associated with metadata that describe its contents. Screen shot shows an OpenSHAPA spreadsheet (with participant demographics, user-defined behavioral codes and visualization of nested data structure), video controller, and one video frame.

Identified gaps and opportunities

It was clear from the workshop and from subsequent discussions that no large-scale scientific repository exists for sharing digital video data that meets these criteria. The developmental science community is faced with creating a system that makes contributing video and metadata convenient and reliable, permits powerful and transparent searching through deposited datasets, and enables discovery. The system should be open source so that it can be “owned” and developed further by the community, and wherever possible the system should incorporate existing tools and infrastructure.

Discussions at the workshop made it clear that many previous data-sharing efforts have failed not because of technological limitations, but because the research community did not fully embrace the effort. Thus, a central priority is to build an engaged and diverse user/contributor community of developmental scientists who are committed to open data sharing and to the creation of standards for data access and quality.

Workshop outcomes

Since the workshop, we have worked steadily towards the submission of proposals to NSF, NICHD, and other agencies. We have engaged in intense consultation with colleagues and with leaders of existing data sharing initiatives to find an approach most likely to succeed for the developmental science community. Some of our steps and their outcomes are outlined below.

Work towards major funding proposals

After the workshop, Adolph, Sanderson, and Freeman took steps towards preparing major funding proposals to support data coding, management, and sharing in developmental science. We added two colleagues to the leadership team. They bring great strengths that complement the interests and expertise of the original workshop organizers.

Rick Gilmore, Ph.D., Associate Professor of Psychology and Director of Human Imaging, Social, Life, and Engineering Sciences Imaging Center (SLEIC), The Pennsylvania State

University. Gilmore studies the neuroscience of perception, memory, and action in infancy and early childhood, using behavioral, neuroimaging (EEG and MRI), and computational methods. Alongside his role as Director of SLEIC, he also coordinates the Human Developmental Neuroscience Initiative (HDNI) under the auspices of the Child Study Center (CSC) with support from the Social Sciences Research Institute (SSRI). Gilmore brings further developmental science expertise as well as knowledge of data management and visualization.

David Millman, Ph.D., Director of Digital Library Technology Services, New York University. Millman came to NYU in 2008 from Columbia University, where he was Senior Director of Research, Teaching, and Learning Technologies in the Columbia IT unit. He has developed and managed internet-based services since the late 1980s, including public information systems, reference book databases, art museum collections and electronic scholarly publications. He has also been a PI or co-PI on several NSF grants. Millman brings expertise with digital media data archiving and data sharing.

Sought out commitments to contribute data in the future

At the core of any data sharing initiative is the commitment of researchers to make contributions to the shared data library and to use the data in the library. In the process of developing the NSF and NIH proposals, we secured commitments from more than 90 researchers [28] with PI status at their institutions as "committed contributors" to the projected developmental science data-sharing environment. These researchers represent the diversity of work in developmental science (including basic and applied work, typically developing and clinical populations, human and animal work across cognitive, linguistic, social, perceptual, emotional, and motor development) and in institutions that support developmental research (large research I universities and small teaching colleges). Many of the contributors are working at NYU, PSU, CMU, Indiana University and University of Iowa, around members of the leadership team and advisory board. These user clusters will be instrumental in developing the OpenSHAPA coding tool, data management tools, and cyber infrastructure.

Derived draft permission language with community input

For data sharing to be possible, participants, contributors, and users must make various formal agreements about sharing. (1) Participants and all others depicted on the video must give permission for the video file to be shared at one of several levels of access. (2) Contributors must ensure that video and metadata contributed to the shared environment are released with the level of access originally granted by participants. (3) Users of shared data must respect the level of access that was originally granted by the participants.

With input from members of the developmental science community, including many on the advisory board of the project (see section below) we developed draft wording for a document that participants (their parents or guardians in the case of minors) would sign, granting permission to one of several levels of future shared access by researchers, or not granting permission on any level of shared access [28]. Contributors to the project have committed to work with their respective IRBs to finalize agreements that would permit open data sharing in their current and future studies. The draft permission form is included in Appendix C.

Consultations with NYU and others on technical matters

A data-sharing initiative requires a technical home where servers and disk arrays are housed, websites are hosted, and backup and maintenance are performed. We have consulted with senior personnel at NYU Digital Library Technology Services (DLTS) as well as the Research Computing and Cyberinfrastructure (RCC) unit at PSU. We have established that NYU will host the project and PSU will be a mirror site. These facilities at both NYU and PSU have significant

prior experience with hosting major data sharing projects, including many NSF-funded cyber infrastructure and archiving projects.

Formed advisory board including leaders of existing data sharing initiatives

Given our goal to create a culture of data sharing within developmental science, we are eager to draw on the experiences of those who have led related projects. Accordingly, we formed an advisory board to guide the project (See Appendix D). The directors of ICPSR [18] (Alter), TalkBank [19] (MacWhinney), SIDGrid [20] (Bertenthal), BIRN [16] (Kesselman), and DIVER [21] (Pea) have agreed to serve. They have already provided counsel on critical issues that include the following:

- Need for effective mechanisms for controlling access to a repository (ICPSR)
- Importance of fostering a scientific community that advocates for the resource (TalkBank; SIDGrid; ICPSR)
- Challenges of creating software that scientists will actually use (SIDGrid; BIRN)
- Desirability of making data contribution easy (BIRN, ICPSR)
- Tensions between instituting standard search and coding ontologies versus allowing these to emerge organically from the community (TalkBank, BIRN, ICPSR)
- Utility of adapting and adopting free, open source, existing tools whenever possible (BIRN)
- Importance of a broad base of institutional support (BIRN, ICSPR, SIDGrid, TalkBank) and its link to sustainability (ICPSR)
- Need to ensure that the infrastructure provides sufficient incentives for investigators to contribute to and draw from the data repository (TalkBank, SIDGrid, ICPSR, BIRN).

Developed 5-year plan to create a video-based data library (Databrary) and upgrade OpenSHAPA

Based on the above activities and input, we have developed a five-year plan to create a shared developmental science data library and to upgrade OpenSHAPA to support the library. We named the shared data library “Databrary” to avoid possible conflicts with previous uses of the name “OpenSHARE.” The technical work plan is organized around three specific aims, listed below with the general nature of our activity indicated under each aim.

Aim 1: Create the infrastructure and data management tools to enable open sharing of video data among developmental scientists.

To fulfill this aim we will build a team of developers, collect seed datasets, develop a hardware and storage environment for data, create user authentication and access protocols, develop software that helps users contribute their data, develop searching and browsing tools for users to explore data stored in Databrary.

Aim 2: Expand and develop the OpenSHAPA coding tool to facilitate annotation, exploration, visualization, analysis, and sharing of video data.

To fulfill this aim we will upgrade OpenSHAPA’s video and audio replay capabilities, construct further plugins for other kinds of data such as physiological or motion capture data, and provide data visualization and statistical analysis capabilities.

Aim 3: Build the capacity for data sharing and coding based on the expertise and needs of the developmental science community

To fulfill this aim we will engage in community-building activities. We will train Databrary and OpenSHAPA users at workshops in our laboratories and at conferences, and we will provide onsite and web-based training. Finally, we will also seek further sources of support so that the Databrary and OpenSHAPA projects are sustainable into the future.

Summary of aims. We will develop a web-based open video data-sharing library, Databrary.org, and seed it so that users can engage the system early on. We will develop data management tools to enable sharing and mechanisms that permit interoperability between coding tools. We will develop template IRB language for a data-sharing permission form so that investigators can begin to collect and contribute datasets to the library, and we will develop appropriate contributor and user agreements to ensure appropriate levels of access. We will expand the OpenSHAPA coding tool to meet diverse research needs, such as incorporating multi-media data with video. We will pursue our aims with the endorsement and participation of the developmental science community. And finally, we will develop and implement plans for long term sustainability.

Submitted proposal to NSF: In discussion with NICHD and NINDS

In March 2012, we submitted a major research proposal to NSF based around the above outcomes. We are currently in discussion with NICHD and NINDS on the procedure for submitting companion proposals to NIH. The objective is for two or more agencies to share support for the Databrary and OpenSHAPA project.

Conclusion

The conclusion of the workshop is that the challenges for open video data sharing in developmental science are surmountable, the benefits for developmental researchers are immense, and the time is right to pursue this project.

Intellectual Merit. By creating tools for open video data sharing, we expect to deepen insights and accelerate the pace of discovery in developmental science. The contribution of a particular dataset will no longer depend on the private analytic and interpretive activities of researchers from one laboratory, but instead benefit from the critique of many researchers with different viewpoints. This strengthens the credibility of the research enterprise as a whole. Researchers will be able to view one another's datasets and reanalyze them to test competing hypotheses or address new questions beyond the scope of the original study. The intellectual merit of this workshop is that it has enabled the development of a proposal aimed at achieving these ends.

Broader Impacts. The infrastructure the workshop PIs and their colleagues on the Databrary project will develop will have impacts beyond the developmental science community. The entire behavioral science community can benefit from the OpenSHAPA and open data-sharing tools, and we expect that insights that emerge from this project can facilitate open data sharing and better data management practices in other research communities. Moreover, through our community building, we will train a new generation of developmental scientists who will be empowered with an unprecedented set of tools for discovery. Some of the researchers who benefit will come from institutions without substantial resources to support scientific research. Finally, since some data sets will be available for public viewing, we will raise the profile of developmental science and bolster interest in and support for scientific research among the public at large.

References

1. *To Share or not to share: Research data outputs*. (2008). UK Research Information Council..<http://www.rin.ac.uk/our-work/data-management-and-curation/share-or-not-share-research-data-outputs>
2. Derry, S. J. (2007). *Guidelines for video research in education: Recommendations from an expert panel*. Data Research and Development Center; University of Chicago. <http://drdc.uchicago.edu/what/video-research-guidelines.pdf>
3. Derry, S. J. *et al.* (2010). Conducting video research in the learning sciences: Guidance on selection, analysis, technology, and ethics. *Journal of the Learning Sciences* **19**, 3-53.
4. MacWhinney, B. (2001). From CHILDES to TalkBank. *Research in Child Language Acquisition*.
5. MacWhinney, B. (2007). The TalkBank Project. *Department of Psychology*. <http://repository.cmu.edu/psychology/174>
6. National Academies workshop on data attribution and citation. (2011) <http://sites.nationalacademies.org/PGA/brdi/PGA_064019>
7. Kaye, J., Heeney, C., Hawkins, N., Vries, J. de & Boddington, P. (2009). Data sharing in genomics -- re-shaping scientific practice. *Nature Reviews Genetics* **10**, 331-335.
8. Young, J. R. (2010). Crowd science reaches new heights.*The Chronicle of Higher Education*. <http://chronicle.com/article/The-Rise-of-Crowd-Science/65707/>
9. Kleiner, K. (2011). Data on demand. *Nature Climate Change* **1**, 10-12.
10. *Digital Research Data Sharing and Management*. (2011). National Science Board. <http://www.nsf.gov/nsb/publications/2011/nsb1124.pdf>
11. *Data Sharing Workbook*. (2004). National Institutes of Health. <http://grants.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf>
12. OpenSHAPA site <https://openshapa.org>
13. Gilmore, R. O. & Adolph, K. E. (2012). *Video use survey of ICIS and CDS listserv subscribers*.
14. Piwowar, H. A., Day, R. S. & Fridsma, D. B. (2007). Sharing detailed research data is associated with increased citation rate. *PLoS ONE* **2**, e308.
15. Gelman, S. Technology could help. (2012). *Rigor without rigor mortis: The APS board discusses research integrity*. <http://www.psychologicalscience.org/index.php/publications/observer/scientific-rigor.html#gelman>
16. Biomedical Informatics Research Network <http://www.birncommunity.org/resources/tools/>
17. International Electrophysiology Epilepsy Data Portal <http://ieeg.org>

18. Inter-university Consortium for Political and Social Research (ICPSR) <http://www.icpsr.umich.edu/icpsrweb/ICPSR/>

19. TalkBank site <http://talkbank.org>

20. Social Informatics Data Grid (SIDGrid) <http://sidgrid.ci.uchicago.edu/home>

21. DIVER: Digital Interactive Video Exploration and Reflection. <http://diver.stanford.edu/>

22. Creative Commons licenses <http://creativecommons.org/licenses/>

23. ELAN: Language archiving technology <http://www.lat-mpi.eu/tools/elan/>

24. Transana:Qualitative analysis software for video and audio data <http://www.transana.org/>

25. Besser, H. & Van Malssen, K. (2007). Pushing metadata capture upstream into the content production process: Preliminary studies of public television. <http://ils.unc.edu/digccurr2007/papers/besserVanmalssen_paper_4-1.pdf>

26. Neurosynth.org <http://neurosynth.org>

27. Brainmap.org <http://brainmap.org>

28. Databrary project site <http://databrary.org>

Appendix A: Participants

Organizers and OpenSHAPA team (4)

Karen Adolph, Professor of Psychology and Neuroscience, New York University.

Penelope Sanderson, Professor of Cognitive Engineering and Human Factors, The University of Queensland.

Clinton Freeman, Senior Research Engineer, National Information and Communication Technology (NICTA), Australia

Jesse Lingeman, Research Scientist, Department of Psychology, New York University

Senior participants (19)

Martha Alibali, Professor of Psychology and Educational Psychology, University of Wisconsin-Madison

Micah Altman, Senior Research Scientist, Institute for Quantitative Social Science, Harvard University, and Archival Director of the Henry A. Murray Research Archive

Richard Aslin, William R. Kenan Professor of Brain and Cognitive Sciences, University of Rochester

Bennett Bertenthal, Rudy Professor of Psychological and Brain Sciences, Indiana University

Katy Borner, Victor H. Yngve Professor of Information Science, Indiana University

Marc Bornstein, Senior Investigator and Head of Child and Family Research, National Institute of Child Health and Human Development

Pamela Davis-Kean, Research Associate Professor, Research Centre for Group Dynamics and Center for Human Growth and Development, Institute for Social Research, University of Michigan

James Cole Galloway, Professor of Physical Therapy, University of Delaware

Michael Goldstein, Associate Professor of Psychology, Cornell University

Wayne Gray, Professor of Cognitive Science, Renssalaer Polytechnic Institute

Robert Hoffman, Senior Research Scientist, Institute for Human and Machine Cognition

Jeffrey Lockman, Professor of Psychology, Tulane University

Brian MacWhinney, Professor of Psychology, Carnegie-Mellon University

Daniel Messinger, Professor of Psychology, of Pediatrics, and of Electrical and Computer Engineering, University of Miami

Francis Quek, Professor of Computer Science, Virginia Tech

Linda Smith, Distinguished Professor and Chancellor's Professor of Psychological and Brain Sciences, Indiana University

John Stamper, Technical Director, Pittsburgh Science of Learning Center DataShop, and member of research faculty, Human-Computer Interaction Institute, Carnegie-Mellon University

William Wong, Professor of Human-Computer Interaction and Head of the Interaction Design Centre, Middlesex University

Chen Yu, Associate Professor, Departments of Psychological and Brain Sciences and of Cognitive Science, Indiana University

Junior Participants (12)

Paulo Carvalho, Doctoral Student, Department of Psychological and Brain Sciences, Indiana University

Kaitlin Fausey, Postdoctoral Fellow, Department of Psychological and Brain Sciences, Indiana University

John Franchak, Postdoctoral Fellow, Departments of Psychology and Neuroscience, New York University

Damian Fricker, Doctoral Student, Department of Psychological and Brain Sciences, and of Cognitive Science, Indiana University

Amy Joh, Assistant Professor, Department of Psychology and Neuroscience, Duke University

Lana Karasik, Assistant Professor, Department of Psychology, The City University of New York, Staten Island

Celeste Kidd, Doctoral Student, Department of Brain and Cognitive Sciences, University of Rochester

Michele Lobo, Research Scientist, Department of Physical Therapy, The University of Delaware

Evan Patton, Doctoral Student, School of Cognitive Science, Rensselaer Polytechnic Institute

Kasey Soska, Postdoctoral Fellow, Department of Psychology, University of Virginia

Catarina Vales, Doctoral Student, Department of Psychological and Brain Sciences, Indiana University

Tian Xu, Doctoral Student, Departments of Computer Science and of Cognitive Science, Indiana University

Appendix B: Workshop Schedule

Day 1	**(Thursday, September 15)**
8:00-8:30	*Breakfast*
8:30-10:00	Introduction
10:00-10:30	Morning break
10:30-12:00	**Data Coding, Management & Sharing** Linda Smith, John Stamper, Brian MacWhinney
12:00-1:30	*Lunch*
1:30-3:00	**Data Mining, Visualization, Analysis, Plug-Ins** Daniel Messinger, Chen Yu, Katy Borner
3:00-3:30	Afternoon break
3:30-5:30	**Data Sharing: Professional & Technical Issues** Marc Bornstein, Bennett Bertenthal, Micah Altman, Pamela Davis-Kean
7:00	Workshop dinner at Willow
Day 2	**(Friday, September 16)**
8:00-8:30	*Breakfast*
8:30-10:00	**Managing Multiple Data Streams** Martha Alibali, Mike Goldstein, Cole Galloway
10:00-10:30	Morning Break
10:30-12:00	**Data Annotation, Exploration, Visualization** Robert Hoffman, Wayne Gray, William Wong
12:00-12:15	Identification of 3-5 key issues for discussion after lunch
12:15-1:30	*Lunch*
1:30-3:00	Problem-oriented group discussions
3:00-3:30	Afternoon break
3:30-4:00	Short summary presentations from group discussions
4:00-5:00	**Summary** Richard Aslin, Francis Quek, Jeff Lockman
7:00	Workshop dinner at Rock Bottom

Appendix C: Data Sharing Permission Draft (as of March 1, 2012)

Data Sharing Permission for Child [Adult]

This form requests your permission to include [your data and] your child's data in a data library on the Internet (Databrary.org). The library allows researchers interested in development and behavior to share findings. Data sharing will help researchers to learn more from the data they collect and will lead to faster progress in our understanding of human development.

Giving permission to share data is entirely separate from giving consent to participate in the research study. You do not have to give permission to share [your or] your child's images or other data in the library. If you choose not to share [your or] your child's data with the library it will not affect your receipt of payment if offered or [your or] your child's participation in this or future studies.

What data will be shared?

With your permission, we will include video recordings of [you or] your child's behavior in the data library. We will also include basic information about [you or] your child such as age, sex, race, ethnicity, and geographic region. We may also include information from interviews and questionnaires and/or information collected with other recording methods such as motion tracking, eye tracking, and brain imaging. If researchers code and analyze the video files and/or other sources of information, this information will also be shared.

If [you or] your child [have] has a medical diagnosis, we may wish to include that information in the library. If you provide health-related information about [you or] your child's condition, prior treatments, medications, or family history of illness during this study, it may be shared in the library. But, because this information may be more private, we will request your permission to include these data in the library on a separate data release form.

Will the data be confidential or anonymous?

All of [your or] your child's data will be filed by an identification code, not by name. No information will be included in the data library about how to contact you or your child (child's last name, parents/guardians' last names, address, phone number, email, etc.). Authorized users of the library must agree not to try to contact you or your child.

Your child's image and/or voice will be visible on the video. Your child's name and/or your name may be spoken out loud on the video. Your image and/or voice and those of visitors or other members of your family may also be seen on the video; if the study takes place in your home, aspects of your home may be seen on the video. Thus, it is possible that you, your child, or other people could be identified from the video by accident. However, authorized users of the library must agree not to mention the name of any person on the video in published reports of data from the library.

Who can Access the Data in the Library?

Videos and other data describing the behavior of individuals can only be viewed and downloaded from the library by authorized users. Researchers who wish to have access to the data must formally apply for access. Only researchers whose research is supervised by Institutional (Human Subjects) Review Boards or similar organizations that supervise research

will be authorized for access. Researchers must renew their authorization for access to the library every year.

Authorized users must sign a data use agreement. It requires them to maintain the confidentiality of the data, treat human subjects ethically, and **not** use the data for commercial purposes. Authorized users must treat data in the library with the same high standards of care that they would treat data collected in their own laboratories.

How long will the data be in the library?

Data in the data-sharing library will be preserved indefinitely in a secure facility.

Permissions:

Please check the appropriate boxes below:

☐ I do not give permission for any of [my data or] my child's data to be included in the data library.

or

☐ I give permission for [my data or] my child's data to be shared in the library.

I wish to limit who can view data, images, and video of [me or] my child, as follows:

☐ [My data or] my child's data can be viewed *only by authorized users and researchers under their supervision* in the conduct of scientific research.

☐ [My data or] my child's data, including photographic images and video excerpts can be viewed by *public audiences* for scientific purposes (e.g., professional conferences, talks) and/or educational purposes (e.g., classroom lectures, workshops).

I wish to limit how digital files of images or video excerpts of [me or] my child can be shared outside the library, as follows:

☐ I do not give permission to release any digital files outside the library.

☐ I give permission to release digital files to individuals who are not authorized users of the data library. I trust that authorized users of the data library will exercise professional judgment and uphold ethical principles in determining which files to release, to whom, and how.

Compensation

There will be no compensation to you or to your child for the use of data in the library.

Print child's name _____

Print parent/guardian's name _____

Parent/guardian's signature _____ **Date** _____

Researcher obtaining consent _____ **Date** _____

We will give you a copy of this form for your records. If you have any questions about the data-sharing library, please email to **[Databrary EMAIL CONTACT]**. For questions about your rights as a research participant, you may contact **[APPROPRIATE CONTACT FOR LOCAL INSTITUTION]**.

Appendix D: Databrary Advisory Board

David Ackerman, Executive Director of Digital Library Technology Services, *New York University*

Martha Alibali, Professor of Psychology and Educational Psychology, *University of Wisconsin at Madison*

George Alter, Professor of History, Director of ICPSR, *University of Michigan*

Richard Aslin, William R. Kenan Professor of Brain & Cognitive Sciences, Director of Rochester Center for Brain Imaging, *University of Rochester*

Roger Bakeman, Professor of Psychology Emeritus, *Georgia State University*

Bennett Bertenthal, James H. Rudy Professor of Psychological and Brain Sciences, Director of SidGRID, *Indiana University*

Howard Besser, Professor of Cinema Studies, Director, Moving Image Archiving & Preservation Program, *New York University*

James Cole Galloway, Professor of Physical Therapy, Psychology, Human Development and Family Studies, and Biomechanics and Movement Sciences, *University of Delaware*

Paul Horn, Senior Vice Provost for Research, Distinguished Scientist in Residence, *New York University*

Carl Kesselman, Professor of Industrial and Systems Engineering, Fellow of the USC Information Sciences Institute, Director of BIRN, *University of Southern California*

Yann LeCun, Silver Professor of the Courant Institute of Mathematical and Computer Sciences, *New York University*

Brian MacWhinney, Professor of Psychology, Director of TalkBank and CHILDES, *Carnegie Mellon University*

Carol Mandel, Dean of the Division of Libraries, *New York University*

Susan McHale, Professor of Human Development, Director of the Social Science Research Institute, *The Pennsylvania State University*

Peter Molenaar, Distinguished Professor of Human Development and Psychology, *The Pennsylvania State University*

Roy Pea*,* David Jacks Professor of Education and Learning Sciences, Director of DIVER, Director of the H-STAR Institute*, Stanford University*

Jan Plass, Paulette Goddard Professor of Digital Learning and Media Sciences, *New York University*

Padma Raghavan, Professor of Computer Science and Engineering, Director of the Institute for CyberScience, *The Pennsylvania State University*

Nilam Ram, Assistant Professor of Human Development and Psychology, *The Pennsylvania State University*

Linda Smith, Distinguished Professor and Chancellor's Professor of Psychological and Brain Sciences, *Indiana University*

Luke Zhang, Associate Professor of Information Sciences and Technology, *The Pennsylvania State University*