

# Curating Identifiable Data for Sharing: The Databrary Project

Rick O. Gilmore  
The Pennsylvania State University  
University Park, PA  
rogilmore@psu.edu

Karen E. Adolph & David S. Millman  
New York University  
New York, NY  
karen.adolph@nyu.edu, dsm@nyu.edu

**Abstract**—Video captures the nuances and dynamics of human behavior more cheaply and completely than other recording methods. Although the detailed fidelity of video provides significant potential for reuse, this potential goes largely unrealized because videos are rarely shared. Moreover, the large size of video files, diversity of file formats, and incompatible software tools pose technical challenges, and recordings of faces and voices pose issues regarding participant identifiability. The Databrary (databrary.org) data library, based at NYU, has developed solutions for: securely sharing research video while respecting participants’ privacy, storing and streaming shared video, and managing video datasets and associated metadata. Video data are big data, and interest in recording, analyzing, and sharing video for research, education, and policy purposes continues to grow. Databrary makes video data sharing convenient and attractive for researchers, thereby increasing transparency and enhancing the potential for discovery.

**Keywords**—video; data sharing; open science; data science

## I. INTRODUCTION

Video captures the nuances and dynamics of human behavior unlike any other recording method. This makes video especially attractive for thousands of behavioral scientists across a wide range of fields. Because video is largely self-documenting, it presents significant potential for reuse without the requirement of extensive metadata. However, this potential goes largely unrealized because videos are rarely shared and existing metadata are not captured reliably and organized to enable discovery. Moreover, research video contains information about participants’ identities, making the materials challenging to share. The large size of video files, diversity of formats, and incompatible software tools pose additional technical challenges.

With funding from NSF (BCS-1238599), NIH (NICHD U01-HD-076595), and the Society for Research in Child Development, we have built the Databrary (databrary.org) digital data library that is specialized for storing, managing and sharing research video. Databrary has overcome critical barriers to sharing video, including solutions for respecting participants’ privacy; for storing, streaming, and sharing video; and for managing video datasets and associated metadata [1]. Databrary’s technology and policies lay the groundwork for securely sharing research videos. As of August, 2016, in only

two and half years of operation, Databrary has collected thousands of individual videos—6000+ hours of recordings—featuring more than 5,000 infant, child, and adult participants. More than 600 authorized researchers and affiliates from more than 250 institutions in North and South America, Europe, Asia, and Australia have gained access to the library, and the numbers grow daily. Databrary has also developed a free, open-source video-coding tool called Datavyu (<http://datavyu.org>) to enable researchers to add annotations that are time-locked to individual frames or video segments.

In this paper, we describe how video data are being currently used in a variety of fields, and we show how collecting and sharing of video in contexts where it is not currently used can strengthen scientific transparency and reproducibility. We discuss some of the challenges that have previously limited video data sharing and describe the solutions Databrary has devised to overcome them. We briefly describe the process of video annotation using the Datavyu tool, and conclude with some thoughts about the role of video in combination with other data streams.

Video data are big data, and interest in recording, analyzing, and sharing video for research, education, and policy purposes continues to grow [2,3]. Databrary makes video data sharing convenient and attractive for researchers, thereby enhancing the potential for discovery in the sciences of human behavior and learning.

## II. VIDEO AS DATA

Researchers who study human or animal behavior, have long recognized the power of visual media to capture the richness and complexity of behavior as it unfolds in real time [4-6]. Video has become the backbone of research for thousands of scientists who study learning and development, each of whom collects hundreds to thousands of hours of video each year [7].

The scale of video collection in some large collaborative research projects is even larger. For example, the Measures of Effective Teaching Project, 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 and streamed to registered viewers across the country. The NSF-funded

HomeBank project, affiliated with the TalkBank/CHILDES archive, 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 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 and animal behavior.

At the same time, video has unique virtues that may be attractive to scientists in other fields who do not currently collect or analyze video but who are concerned about reproducibility, transparency, and openness in scientific research [8-10]. 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 [4-6, 11, 12]. As such, video can capture essential details about empirical procedures that are overlooked or omitted in the most detailed and carefully written methods sections of scientific papers. Video can record how participants gave consent to participate in research, what tasks participants performed and in what order. Video recordings can capture behavior in laboratory or classroom settings or spontaneous behavior in more public settings. Video can capture the dynamics of computer-based tasks or displays used in laboratory research. Videos of empirical procedures can and should be viewed as the gold standard of documentation across the behavioral sciences. Indeed, were the use of video for this purpose more widespread, many disagreements about whether empirical replications truly reproduced the original experimental conditions would be moot [9,10].

Our surveys of the developmental science community [7] suggest that many researchers already use video to record experimental procedures to train lab staff and research collaborators. The next step is for these researchers and others to regularly share these videos with their colleagues. The *Journal of Visualized Experiments (JOVE)* has arisen in part to fulfill this need, but we argue that other, more widely available and inexpensive mechanisms for sharing videos depicting research methods are needed.

Whether recordings represent raw data or the documentation of experimental procedures, the use of video in scientific research will continue to grow and to pose new challenges.

### III. THE CHALLENGES OF VIDEO

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 by a different observer for a different purpose. This makes video an especially valuable raw material for discovery. Capitalizing on this value, however, requires overcoming a unique set of challenges.

*Personally identifiable information on video poses problems for the protection of participants' privacy.* For years, policies have existed for sharing de-identified text-based data [11]. But video cannot be readily de-identified in the same ways as text data. Most videos of people contain 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 record from 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 files into formats that can be used and understood by others [14]. Most researchers lack training and expertise in standard practices of data curation [1]. Video 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 federal data sharing policies [15]. It also makes curating datasets a challenge for repositories.

*Extracting behavioral patterns from video presents technical and practical barriers.* Although machine-assisted image and video-tagging has made significant advances in recent years, the rich and diverse information contained in video still requires time-consuming, laborious work by human observers to extract behaviors of interest and record them in the form of time-locked annotations or tags. The extracted data are represented in specialized ways, including paper and pencil, and are not easily exportable to other tools or statistical analysis software. In principle, researchers could build on the videos and tags generated by others. But in practice, most researchers do not share coding files with researchers outside their labs. Moreover, coding files often contain proprietary and incompatible data formats making them difficult to push along the analysis pipeline and to share with other researchers. As a result, the hard-won, expensive-to-acquire human insights about behaviors extracted from research video remain difficult to analyze and largely hidden to the greater scientific community.

#### IV. MEETING THE CHALLENGES

Mindful of these challenges but motivated by the scientific promise of video data sharing and with substantial financial support from NSF and NICHD, the PIs established Databrary, the first-of-its-kind web-based digital data library for storing, managing, and sharing research video and associated metadata. Databrary has targeted the developmental and learning sciences community that is the PIs' intellectual home. But, the team specifically designed Databrary to be adapted for and used by other researchers in the behavioral sciences.

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 metadata 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 simultaneously solving some of the thorniest problems associated with sharing data that contain personally identifiable information.

##### A. Databrary's policies enable sharing of identifiable data

Databrary's policy framework recognizes that to maximize the potential for reuse, the content of recordings must not be altered. Thus, Databrary does not attempt to de-identify videos. However, to enable sharing of unaltered research video containing identifiable information Databrary developed new policies to protect participants' privacy. First, Databrary restricts access to researchers who register and secure formal authorization from their institutions, and second, Databrary shares identifiable data only with the explicit permission of the participants. Databrary has created template language for seeking participants' permission to share data which researchers may adapt for their own use. An online user guide fully describes these policies.

Unique among data repositories, Databrary authorizes both data use and contribution. However, users agree to store on Databrary only materials for which they have IRB/ethics board 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. Hundreds of researchers and their institutions have agreed to Databrary's framework.

##### B. Databrary overcomes technical barriers to video data sharing

To address the problem of diverse video formats, Databrary uses NYU's high performance computing services to automatically transcode each recording 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. As of August 2016, Databrary's assets total more 20TB and are stored on NYU's central IT storage, which provides one off-site mirror and regular long-term tape backups.

##### C. Databrary's design overcomes practical barriers to sharing

Video requires little metadata to be useful, and the only metadata that are strictly required are participants' data sharing preferences. Databrary has developed a novel *active-curation* framework that reduces the burden of post hoc data sharing [1]. 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.

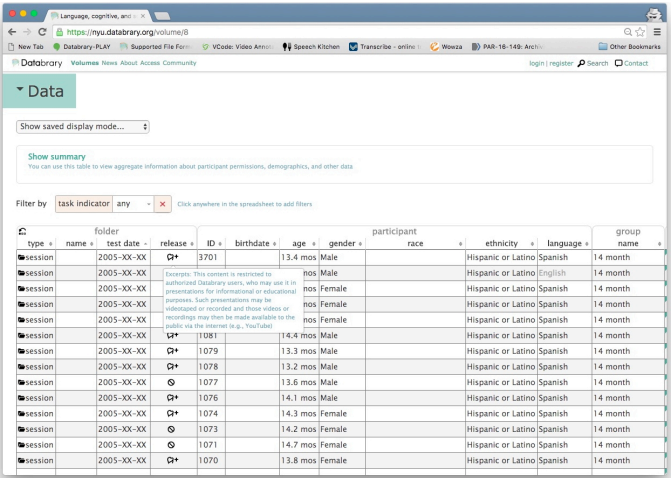
To encourage immediate uploading, Databrary provides a complete set of controls so that researchers can restrict access to their own labs or to other users of their choosing prior to sharing. Datasets can be shared with the broader research community at a later point when data collection and ancillary materials are complete, whenever the contributor is comfortable sharing, or when journals or funders require it. 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.

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.

Figure 1 shows Databrary's spreadsheet interface that enables researchers to store and share individual-level metadata about participants, including indicator icons showing the level of data sharing permitted. Figure 2 shows Databrary's timeline interface that data contributors can use to upload multiple video files or other data streams. Users can view these videos, create highlights from them, and tag them in the web browser. Shared materials must be made available in findable, accessible, interoperable, and reusable formats to be maximally useful to others [16]. To that end, Databrary allows researchers to search for videos that meet their particular specifications using the interface depicted in Figure 3. Each data set or study on Databrary has its own unique web page that when shared receives its own persistent identifier that may be used to cite the resource.

Active curation poses few new burdens on 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. Moreover, by entering participant- and study-level metadata into Databrary, researchers make it possible for others to search for participants or studies that meet specific criteria. Thus, in the very near future, researchers who wish to reuse

materials from Databrary for new studies will be able to find exactly the videos and related metadata they need to address their new question. To our knowledge, no other data repository promises this sort of capability.



The screenshot shows a web browser displaying a Databrary spreadsheet. The spreadsheet has columns for session, test date, release, ID, birthdate, age, gender, race, ethnicity, language, and group name. The data is organized into rows for different sessions and participants.

session	test date	release	ID	birthdate	age	gender	race	ethnicity	language	group name
session	2005-XX-XX	Q+	3701	13.4 mos	Male	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1079	13.3 mos	Male	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1078	13.2 mos	Male	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1077	13.6 mos	Male	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1076	14.1 mos	Male	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1074	14.3 mos	Female	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1073	14.2 mos	Female	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1071	14.7 mos	Female	nos	Hispanic or Latino	Spanish	14 month	
session	2005-XX-XX	Q+	1070	13.8 mos	Female	nos	Hispanic or Latino	Spanish	14 month	

Fig. 1. Databrary spreadsheet for recording participant metadata.

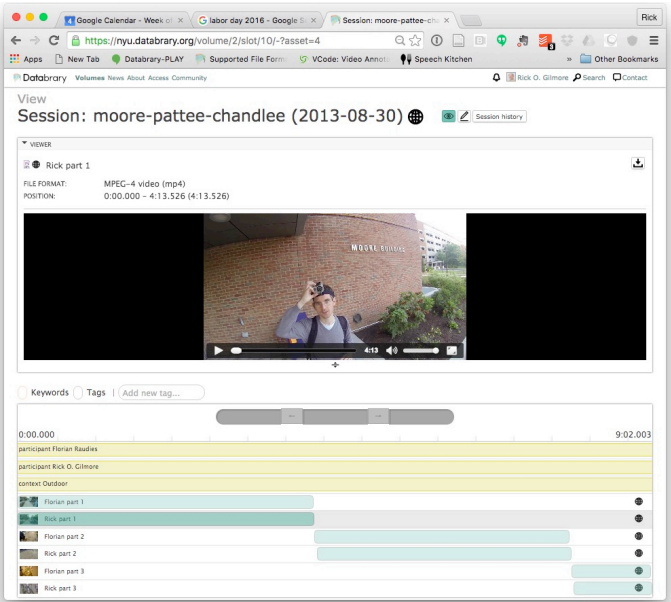


Fig. 2. Timeline for depicting and viewing temporal relations among videos.

D. The Datavyu (datavyu.org) coding tool enables discovery

Most researchers who collect video deploy trained human observers to view the recordings and annotate them with specific tags that label the onset and offset of particular behaviors or events, the category or type of behavior, transcriptions of speech, and qualitative judgments about mood or other psychological characteristics. In the developmental sciences, general purpose spreadsheets and paper-and-pencil are the most commonly used tools for annotating video. But an increasing number of researchers use academic or commercial tools specialized for video and audio annotation. The tools

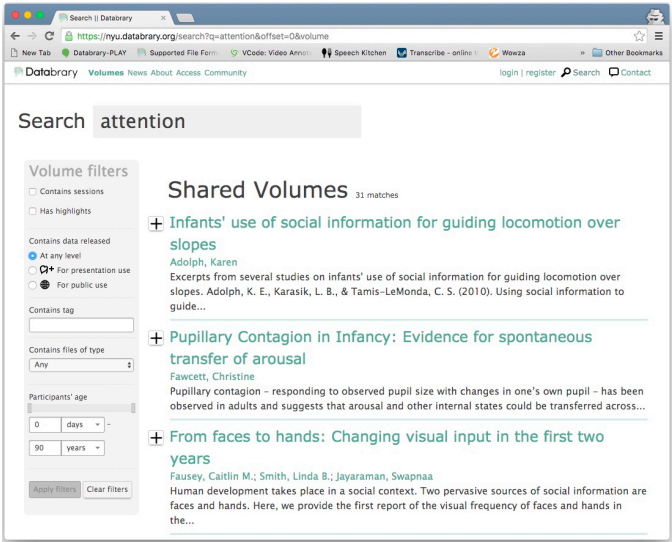


Fig. 3. Search interface for finding videos.

allow coders to play video forward and backward at varied speeds time-locked to the codes. The tools thereby enable researchers to unpack the multi-layered complex patterns of human behavior as it unfolds in real time. Following one or more coding passes, each of which focuses on a subset of behavioral dimensions, the tags or annotations are then exported for visualization and analysis using other tools.

One of us (Adolph) has pioneered the use of video coding tools for mining the data contained in video, and so the development of and support for video coding tools has been integral to the Databrary project from the beginning. Among our initiatives was the publication of a web-based best practices guide for coding video that is agnostic about the tool or tools a researcher chooses to use [17]. We have also continued to develop a free, multi-platform (Windows and Mac OS), open-source, scriptable video coding tool called Datavyu (datavyu.org). Datavyu is highly customizable, making it suitable for a variety of video coding use cases, and its Ruby scripting API makes it possible for users to customize the program and automate many routine tasks that would otherwise require significant time investment from researchers and often error-prone human intervention.

Datavyu files, called spreadsheets, may be uploaded to Databrary and shared along with the videos they are linked to. This allows users in geographically separate locations to share video coding files and to validate or build upon each other's codes. The rich set of time-locked annotations contained in Datavyu files cannot yet be visualized on the Databrary timeline or searched. But, empowering Databrary to import, display, make searchable, and export annotation from Datavyu and related video coding tools remains an important project goal. In the very near future, a researcher will be able to search across Databrary for specific time segments in which a particular type of behavior occurred. For example, a researcher could search for all instances of crying, or all instances of vocalizations by English-speaking children under 24 months of age, and so on.

## V. THE FUTURE OF VIDEO AND BIG DATA SCIENCE

As boyd and Crawford [18] observe "The era of Big Data has begun. Computer scientists, physicists, economists, mathematicians, political scientists, bio-informaticists, sociologists, and other scholars are clamoring for access to the massive quantities of information produced by and about people, things, and their interactions" (p. 662). We argue that video has a vital, often overlooked, role to play in the effort to understand human behavior. The Datavyu tool and Databrary library make it possible for today's researchers to code, store, manage, and share their own research videos and related metadata, and to find, reuse, and build upon video collected by others. And yet, relatively few researchers currently reuse others' video, and despite Databrary's rapid recent growth, only a fraction of developmental and learning scientists who collect video use the system actively.

Accordingly, the Databrary team continues to focus on expanding the range, quality, and quantity of shared video to make it more attractive for researchers to gain access and take advantage of Databrary's assets. The PIs and staff continue to refine the system's features to make active-curation the norm within the behavioral sciences community, and we urge colleagues to take advantage of the value of video in documenting experimental procedures to enhance transparency and reproducibility in behavioral research. Databrary provides a natural home for these sorts of materials.

More broadly, although human-applied image and video annotations will remain the gold standard for the foreseeable future, we are excited about the pace of development in machine learning, automated speech recognition, and computer vision. These emerging technologies hold significant promise for behavioral scientists who want to mine research videos to answer questions not previously possible to address [19]. For example, in the near future, it may be possible for speech, person/face, action, or object detection algorithms to operate in the background on Databrary's videos and when complete, to provide to researchers a set of tags upon which further analyses can build. In addition, Databrary's existing catalog of natural videos collected in home, laboratory, and classroom settings can provide a corpus of materials valuable to technologists who are developing robust machine-based image and video annotation systems that work outside of a carefully curated set of training data. Databrary will continue to explore potential collaborations with researchers in these fields as a way of making existing shared videos more useful to others and as a way of creating incentives for researchers to contribute new video.

Finally, because many behavioral scientists collect video along with other temporally dense data streams (e.g., heart rate, electrodermal responses, electroencephalography, kinematic measurements, eye position, accelerometer), Databrary will explore ways to integrate data from these streams into the library. Eventually, Databrary will store and share data sets reflecting the multiple spatial and temporal dimensions of human behavior all of which are anchored by video. We think this multi-level, multi-scale view of human behavior will lead to significant new discoveries and that by making video

recording the core measurement, those insights will be even more meaningful.

## VI. CONCLUSIONS

Databrary has demonstrated that is feasible to securely share identifiable research video and related metadata. The Databrary team has developed and continues to refine policies and technologies that overcome many of the barriers to widespread video sharing within the developmental sciences community. We believe that Databrary's policies, active curation framework, and technologies have broad potential in the behavioral and social sciences to transform our understanding of human and animal behavior and the reliability, reproducibility, and transparency of research in these fields. Databrary seek partnerships with others who share our vision of a transformed science of human behavior that is anchored in video.

## ACKNOWLEDGMENTS

We thank Alison Dewhurst, Nancy Daneau, Eric Rasumussen, Mark Righter, and David Ackerman for their support in developing Databrary's policy framework and data security plan.

## REFERENCES

- [1] Gordon, A., Millman, D.S., Steiger, L., Adolph, K.E., & Gilmore, R.O. (2015). Researcher-library collaborations: Data repositories as a service for researchers. *Journal of Librarianship and Scholarly Communication*, 3(2), <http://doi.org/10.7710/2162-3309.1238>.
- [2] Adolph, K.E., Gilmore, R.O., Freeman, C., Sanderson, P., & Millman, D. (2012). Toward open behavioral science. *Psychological Inquiry*, 23(3), 244–247. <http://doi.org/10.1080/1047840X.2012.705133>.
- [3] Gilmore, R.O. (2016). From big data to deep insight in developmental science. *Wiley Interdisciplinary Reviews Cognitive Science*, 7(2), 112–126, <http://doi.org/10.1002/wcs.1379>.
- [4] Adolph K.E. (2016) Video as data: From transient behavior to tangible recording. *APS Observer*, 29, 23–25.
- [5] Gesell, A. (1946). Cinematography and the Study of Child Development, *The American Naturalist*, 80(793), 470–475.
- [6] Gesell, A. (1991). Cinemanalysis: A Method of Behavior Study, *The Journal of Genetic Psychology: Research and Theory on Human Development*, 152(4), 549–562.
- [7] Gilmore R.O., Adolph K.E. (2016) *Video use survey of ICIS and CDS listserv subscribers and Datavyu and Databrary users*.
- [8] Nosek B.A., Bar-Anan Y. (2012) Scientific Utopia: I. Opening scientific communication. *Psychological Inquiry* 23(3):217–244.
- [9] Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716.
- [10] Gilbert, D.T., King, G., Pettigrew, S., & Wilson, T.D. (2016). Comment on "Estimating the reproducibility of psychological science." *Science*, 351(6277), 1037–1037. <http://doi.org/10.1126/science.aad7243>.
- [11] Goldman R, Pea R, Barron B, Derry SJ (2014) *Video Research in the Learning Sciences* (Routledge).
- [12] Curtis S. (2011) "Tangible as tissue": Arnold Gesell, infant behavior, and film analysis. *Science in context* 24(3):417–443.
- [13] 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>

- [14] Ascoli, G.A. (2006) The ups and downs of neuroscience shares. *Neuroinformatics* 4(3):213–215.
- [15] National Science Foundation Dissemination and Sharing of Research Results. Available at: <http://www.nsf.gov/bfa/dias/policy/dmp.jsp>
- [16] The FAIR Data Principles - FOR COMMENT (2014) FORCE11. Available at: <https://www.force11.org/group/fairgroup/fairprinciples>
- [17] *Best Practices for Coding Behavioral Data From Video*. Available at: <http://datavyu.org/user-guide/best-practices.html>.
- [18] boyd, d. & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, 15(5), 662–679. <http://doi.org/10.1080/1369118X.2012.678878>
- [19] Raudies, F., & Gilmore, R.O. (2014). Visual motion priors differ for infants and mothers. *Neural Computation*, 26(11), 2652–2668. [http://doi.org/10.1162/NECO\\_a\\_00645](http://doi.org/10.1162/NECO_a_00645)