

# Databrary

## Standard Operating Procedures

### 1. What is Databrary?

[Databrary](#) is a data library intended to foster sharing of information among researchers. It is an ecosystem for sharing behavioral and neural science data and the tools to analyze those data among a community of researchers. Databrary aims to facilitate data sharing and greater transparency within labs, among collaborators, with other researchers and members of the larger developmental and behavioral science community, and with the general public.

Databrary contains De-Identified Data, Non-Identifiable Data, and Metadata. Databrary also contains Identifiable Data – video/audio recordings and photographic images Databrary calls Recordings – and other sensitive materials, such as Self-Reported Health-Related Information. The tools stored in Databrary consist of software that aids the extraction, visualization, and analysis of time-series-based research data. These tools may be open source, freeware, or commercially licensed, and may be supplemented by sample data sets, web links to data sets shared elsewhere within Databrary, or documentation. Every data set and tool will have its own unique, permanent, uniform resource locator (URL) – web address – so that researchers may both take credit for their own Databrary contributions and properly and consistently cite the Databrary resources they use from others.

The Databrary ecosystem also includes a database of researchers, called Investigators or Affiliate Investigators, who have authorization to use and contribute data. The database keeps track of the identities of each Investigator or Affiliate who contributes to, browses data within, or withdraws data or tools from Databrary.

This document sets out Databrary’s Standard Operating Procedures (SOPs). Section 2 describes the challenges that must be met in creating a data sharing system for sensitive, identifiable data. Section 3 describes Databrary operating procedures. Sections 4 and 5 describe other documents involved in Databrary operations and [definitions](#) of key terms which are italicized throughout the text, respectively. For the latest versions of Databrary policies, see the [website](#) or [GitHub repository](#).

### 2. The Challenges of Databrary

Implementing the Databrary ecosystem poses several challenges. The most crucial challenge concerns research ethics. Databrary must keep the promise of confidentiality given to research participants who provide identifiable data while upholding the principle of informed consent.

Any identification of a research participant or unauthorized disclosure of confidential information violates the promise of confidentiality given to the research participant. This requirement poses a challenge. Many of the sources targeted for sharing in Databrary contain personally identifying information about research participants and other individuals involved in a study. Images and/or voices may be visible or heard. Names may be spoken out loud. Images and voices of visitors or other people who are not research participants may be recorded. If a study takes place in a home or other setting, aspects of that home or setting may be recorded. Thus, individuals or locations could be identified from recordings. Those depicted individuals include research participants and other private persons who are not the targets of a research study.

It is technically possible to try to turn recordings into de-identified data through blurring, distortion and other means of modification, but doing so diminishes scientific value. Therefore, Databrary must do everything possible to ensure that data sets are used solely for research purposes and not for the investigation of specific individuals. Databrary and its users must do everything practicable to minimize the chances that individuals can be identified from information shared within Databrary. Databrary staff and Investigators must understand, certify competence in, and abide by the highest standards of ethics concerning research with human participants. This means that no data files can be deposited, stored, or shared within the Databrary ecosystem that contain personally identifying information beyond that incidentally contained in the recordings — faces, voices, names, and the physical location of the recording.

At the same time, the principle of informed consent requires that depicted individuals must give permission for their recordings to be shared with others in Databrary. Similarly, individuals who provide sensitive information, such as self-reported health-related information should be given the opportunity to give or refuse permission for this information to be shared. Depicted individuals and research participants must decide what level of sharing permission they wish to give: no permission at all because they do not wish to share recordings or self-reported health-related information, or sharing permission with the assurance that shared data will be treated with care by other researchers. Only recordings for which permission to share has been given by all depicted individuals may be shared openly in the Databrary. Similarly, only self-reported health-related information for which permission to share has been given may be shared in Databrary. In turn, Databrary must ensure the level of access that depicted individuals have granted. This means that information about sharing permissions must be permanently linked with individual data files in Databrary. It also means that researchers who access files in Databrary must agree to follow the wishes of individuals who grant permission for their recordings or other sensitive data to be shared.

### 3. How Databrary Will Work

This section describes in detail how Databrary will work, from accessing data, metadata, and tools stored in the library to the contribution of new information.

**3.1 Accessing data sets or tools from Databrary** Some information may be easily shared with minimal restriction so long as personally identifying information has been removed and there are no restrictions on the how long the information may be stored. Sharing recordings or other data with personally identifying information or sensitive information such as self-reported health-related information must be shared only among researchers who understand, embrace, and formally agree to abide by principles of research ethics that include maintaining participant confidentiality. Thus, Databrary must monitor and control who has access to data sets and what level of access they are granted.

**Levels of access permitted to data or software** There will be two levels of access: (1) public/unrestricted, and (2) restricted, but open to authorized Databrary Investigators and Affiliate Investigators. We considered a restricted form of sharing among researchers identified by investigators, but rejected it as incompatible with the data sharing requirements of federal funding agencies.

The Databrary system will have a set of pages that allow the public unrestricted access to archival materials (e.g., Gesell movies), materials that do not contain sensitive information (abstracts, researcher information, protocols, visual displays, coding manuals, etc.), and links to data analysis/software tools. If depicted individuals have given permission to share excerpts from recordings, the public preview pages will include excerpts. Identifiable or sensitive data — recordings, coded spreadsheets linked to recordings, self-reported health-related information — and metadata will be assigned to the restricted, but open access level in almost all cases.

**Means of access, user accounts and privileges** There are three means of access to Databrary's assets: browsing, downloading, and contributing. Individuals who wish only to browse Databrary's public/unrestricted assets may do so without authorization from Databrary, however for security reasons the system will log access by internet protocol (IP) addresses (see Access Logging/Usage Metrics). Unauthorized/unregistered individuals may not comment in Databrary's comment sections, contribute data, or carry out other functions reserved for Databrary Investigators or Affiliate Investigators. Those who wish to browse Databrary's restricted, but open assets must apply for Databrary Investigator or Affiliate Investigator status. Who May Apply for Databrary Investigator or Affiliate Investigator Status Databrary Investigators must be eligible for Principal Investigator (PI) status at their home institutions. Graduate and undergraduate students, postdoctoral researchers, research scientists, and research staff may be granted Databrary Affiliate Investigator status if they are carrying out research under the supervision of a Databrary Investigator. Databrary Investigators who supervise Databrary Affiliate Investigators must (1) grant permission to researchers under their supervision

to view data on their behalf, and (2) agree to take responsibility for the data use of the sponsored members under their supervision, and (3) take responsibility for the security of the data (where data are stored, etc.) by following best practices in data security (best-practices.pdf).

**Timing of Databrary Investigator or Affiliate Investigator Application for Researchers Who Want to Contribute Data** A researcher may apply for Databrary Investigator status before formal permission has been granted by an IRB for collecting or contributing a given data set, but no contributions can be shared with Databrary until Databrary Investigators demonstrate that they (1) have permission to share a given data set from a research supervision entity (e.g. IRB), or (2) have demonstrated that no official permission is needed.

**Application Process** Application for Databrary Investigator status may be made via a web form. The following information will be sought:

*Required for all Databrary Investigators and Affiliate Investigators:*

- Name, title, institution, surface address, email, phone.
- Unique Databrary username and password. The user name may be the applicant's official institutional access ID (e.g. [rog1@psu.edu](mailto:rog1@psu.edu)).
- Certification of PI status eligibility from appropriate research supervision entity.
- Research ethics with human participants training certification (e.g., Collaborative Institutional Training Initiative, CITI; <https://www.citiprogram.org>), and expiration date of current authorization.
- Information about research supervision entity (IRB or equivalent) that governs the applicant's scientific activity, if applicable, including name(s) of representative(s) of the institution, surface address, telephone number, email.
- Required for Databrary Affiliate Investigator applicants: Information about the research supervisor (Name, Databrary user ID, and email)
- An electronically signed copy of the [Databrary Investigator Agreement](#).

The agreement obligates applicants to follow principles of ethical research with human participants, to seek permission to share recordings or self-reported health-related information from participants, to convey those permissions to Databrary, to keep their research supervision offices informed about their activities, to follow [best practices in data security](#), to follow the Databrary [Bill of Rights](#), and other principles.

*Optional, but recommended*

- Applicant web site, Research Gate (<http://www.researchgate.net>) profile, social media accounts, ERA Commons ID.

**Application review** Once a Databrary Investigator or Databrary Affiliate Investigator application has been completed and submitted, applicants and the representative(s) of the institution will receive an automated email confirmation that an application has been received and is under review. Databrary staff will review applications to determine the following:

- Applicant identity information is valid and correct.
- Applicant has human participants ethics clearance from CITI, a governing research supervision entity or its equivalent.
- Applicant has PI eligibility or its equivalent at their institution.
- Applicant has not violated Databrary use standards in the past.

For individuals seeking Databrary Affiliate Investigator status, the supervising/sponsoring Databrary Investigator will be notified via email that an application has been filed. The supervising/sponsoring Databrary Investigator will log on to Databrary, approve (or not) the Databrary Affiliate Investigator status application, and provide information about the level of access to be provided to this individual: browsing only, or browsing + downloading, or browsing + downloading + uploading.

Project staff under the supervision of the Project Director (Adolph) and Associate Director (Gilmore) will conduct the review and approval. The Databrary Ethics committee will audit the procedures annually. The Databrary Ethics Committee will consist of staff, Advisory Board members, and the research community at-large.

If an application requires revisions, project staff will communicate the nature of those revisions via email. All email correspondence will be logged. Once an application is approved, researchers and the official(s) from the research supervision entity will be notified by email. The welcome/notification email will include reminders about Databrary policies regarding data use and citation.

**Access logging/usage metrics** Databrary will log access to the system by Databrary Investigator ID and by IP address. Most access logging information will be used by Databrary for internal security, user metrics, and management functions. Some summary information about files browsed, downloaded, shared, or modified may be announced via Databrary Twitter, Facebook, RSS, and other web/social media feeds. For example, Databrary may identify and publicize which data sets are most often browsed, downloaded, and cited. Databrary Investigators will have access to information about Databrary activity of Affiliate Investigator members of the Investigator's research team through an administrative control panel.

**Browsing and downloading** Any member of the public may browse data or download tools with unrestricted/public level of access. Individuals with Databrary Investigator or Affiliate Investigator status may also browse or download assets with restricted, but open access. When Investigators or Affiliates request data sets to download, the Databrary system will remind them to cite data sets and associated papers, to obtain appropriate approvals for conducting research, to follow [best practices for data security](#), and to remove and destroy data that are no longer needed. Databrary will also send periodic email reminders to researchers who download data reminding them to follow these practices.

Many Databrary uses will involve browsing or downloading data sets without conducting research on those data (e.g., finding excerpts for teaching, searching for exemplars to facilitate study design, pre-research activities to determine whether formal analyses are feasible or promising). If Investigators or Affiliates intend to conduct research with data sets, they may be required to obtain permission from a research supervision entity at their institution. Decisions about whether Investigators or Affiliate Investigators may conduct research on data sets or the type of research proposed are beyond Databrary's control. Thus, whenever permitted by the research supervision entity that supervises their research, Investigators or Affiliate Investigators may conduct research on data sets derived from Databrary. It is the explicit responsibility of the Investigator or Affiliate Investigator to obtain the necessary permissions to conduct research by their respective research supervision entity. This is clearly stated in the [Databrary Investigator Agreement](#). When such approval is sought and secured, the Investigator or Affiliate Investigator must provide Databrary with the approved protocol name, IRB or equivalent protocol identification number, and expiration date.

**Account renewal** Investigator and Affiliate Investigator accounts must be renewed annually. Databrary will send email reminders that accounts are due to expire 1 month, 2 weeks, and 1 week prior to the expiration date and ask Investigators and Affiliates to update their account information. To renew their accounts, Investigators and Affiliates will also electronically sign the Databrary Investigator Agreement. Information about account renewals will be sent to research ethics staff at each member's home institution. At the annual renewal, Databrary will also send Investigators information about how their data sets have been accessed. If a Databrary Investigator or Affiliate Investigator fails to renew his or her account by the renewal deadline, the account will be suspended. The account will be reactivated on completion of the renewal procedures.

Accounts that are not renewed within 6 months of the suspension date will be archived. Data contributed by Investigators or Affiliate Investigators who do not renew their accounts will remain on Databrary.

**3.2. Deposits** Researchers who wish to contribute to Databrary must do some planning in advance. They must:

- Determine the source(s) of the data to be deposited;
- For recordings or self-reported health-related information, determine whether depicted individuals or research participants must give permission to share; in most cases, the answer is yes.
- Seek and be granted, as appropriate, permission to collect and share data from the research supervision entity that governs research at the Investigator or clinician's home institution;
- Secure permission to share from depicted individuals or research participants using the Databrary permission to share templates or equivalent language that has been approved by the Investigator's research supervision entity.
- Apply for and be granted Databrary Investigator or Affiliate Investigator status with Databrary;
- Prepare recordings for sharing with Databrary, including tagging files with the level of permission granted by depicted individuals. Remove personally identifying information from other data and metadata.

**Sources** There are two main sources of recordings: (1) recordings collected by researchers or clinicians in laboratory or clinical contexts, and (2) recordings collected from incidental sources not intended for research or clinical purposes. The latter may include home movies or videos, recordings from security cameras or microphones, cameras or other recordings made in public places, and recordings shared with the public via web portals such as [YouTube](#) or [Vimeo](#). Laws governing whether recordings made in public places may be used for research purposes vary. Researchers who wish to deposit recordings from these or other non-laboratory/non-clinical setting sources should consult with their local research supervision entities for specific permission to share recordings with Databrary.

There are two primary types of studies involving recordings and other data streams: (1) new recordings with associated metadata (coding spreadsheets, analyses, manuscripts, etc.); and (2) reused recordings (complete recordings, subsets, or files culled from different original sources) with new associated metadata (e.g., new codes). The same Investigator or research team may reuse original data deposited in Databrary. Databrary will implement a version control/date system to record modifications to, reanalyses of, and extensions to data. Versioning does not change the status of study types. Versions and reused data sets will retain links with the original and subsequent studies.

**Seeking permission To share data** Researchers who wish to share data with Databrary should seek permission to do so from their governing research supervision entity, such as an IRB. Individuals whose scientific activities are not governed by a research supervision entity may choose to have their protocol reviewed by another entity with similar authority. In preparing applications to the research supervision entity, researchers are strongly encouraged to refer to and use the document(s) provided by Databrary (see Section 4).

Links to all of the relevant Databrary documents concerning permissions and access may be found on the Databrary [website](#).

**Applying for approval to share data from a specific study** In addition to securing Databrary Investigator status, researchers who wish to deposit data must file a separate web-based application to share data from a specific study or source. That application seeks information commonly included in an application to an IRB for approval to conduct a study. In particular, the following information will be required:

- Databrary ID(s) of all principal or co-investigators on a project.

- Agreement from all data owners to share data.
- Verification of Databrary Investigator status for all study PIs/data owners.
- Study name, target population, and thematic tags.
- IRB protocol identifier or number, title, and expiration date.

Databrary Investigators or Affiliate Investigators who wish to conduct research by reusing or reanalyzing an existing data set should provide comparable information, typically at the time approval is sought from a research supervision entity. In most cases, research on data in Databrary requires an application to conduct secondary data analysis.

**Study-level application review** Once study-level applications have been completed and submitted to Databrary, applicants and their research supervision entity official(s) will receive an automated email confirmation that an application has been received and is under review. Databrary staff will review applications to determine that the proposed data to be deposited does not violate Databrary’s Bill of Rights (Appendix E) or best practices in data security (best-practices.pdf).

Project staff will carry out the study-level review under the supervision of the Project Director (Adolph) and Associate Director (Gilmore). The Databrary Ethics Committee will audit the approval and review process annually.

If the proposal requires revisions, the project staff will communicate the nature of those revisions via email. Individuals whose applications are rejected by Databrary project staff may appeal the decision to the Databrary Ethics Committee. The decision of the Databrary Ethics Committee is final.

**Data set contribution approval** Once Databrary staff have confirmed that an application to deposit data meets Databrary standards, the Investigator(s) and the research supervision entity official(s) will be notified by email. That email will include information about how and where the data may be uploaded to Databrary. Each data set or tool contribution will be assigned a unique and permanent uniform resource locator (URL), such as

<http://databrary.org/study/12345>

or

<http://databrary.org/tool/4321>

At this point, the data owner(s) may begin data collection, or if the data have already been collected or the tool readied for sharing, may proceed to upload the data or tool.

**3.3 Securing permission from participants to share data** Under most circumstances, the collection of recordings requires that individual participants, and for minors their parents or legal guardians, give written informed consent to participate in some specific research activity. Databrary will require that Investigators demonstrate that they have secured explicit written permission (or for minors assent and parent/guardian permission) to share recordings with Databrary from each individual depicted in a recording — not just research participants — or that the research supervision entity has waived the requirement for explicit written permission. Similarly, Investigators who wish to share self-reported health-related information must also gain permission or assent to share these data with Databrary.

**How will permission to share data be secured?** Databrary will publish permission to share templates. Examples can be found on [GitHub](#). These templates may be submitted with research supervision entity applications for research, and should be used by Investigators when they seek permission to share from research participants or depicted individuals.

**What kinds of permissions to share may be granted** Depicted individuals (or their parent or legal guardians) or research participants (or their parent or legal guardians) may grant different levels of permission to share. Permission/assent to share recordings should be sought separately from permission to share self-reported health-related information. The sharing levels for each are as follows:

*No sharing (Level 0)*

No recordings or self-reported health-related information may be shared beyond the research team on the IRB-approved protocol. De-identified data (coding spreadsheets, survey data, physiological data) and aggregate results and statistical analyses may be shared with Investigators and Affiliates as long as the research supervision entity has placed no time limitation on the duration of data storage.

*Sharing (Level 1)* Recordings or self-reported health-related information may be shared with all Databrary Investigators and Affiliates.

Permission to share recordings is separate from permission to permit excerpts Databrary requires that separate permission be sought to share excerpts with the public from the permission to share recordings with other researchers. An excerpt is a selection from a larger recording that illustrates a particular point or demonstrates a finding. It is typically no more than a few seconds or minutes in length and may be either a video or audio clip or a still image.

Many researchers now ask research participants for permission to share excerpts in various public settings — classrooms, scientific talks, in textbooks, with the media, and on research web sites. Databrary staff believe that from the point of view of a depicted individual, these distinctions between public venues are virtually meaningless and impossible to monitor or enforce. Researchers don't consistently control who is in their classrooms or presentations, who records a lecture or talk, or how materials are distributed. Thus, under current practice, well-meaning investigators may be inadvertently implying that they have more control over an excerpt than is actually the case.

Accordingly, Databrary permission to share templates requires that depicted individuals give separate permission to share excerpts with the public. The template permission document asks whether depicted individuals give permission for Investigators to select excerpts and to share them with the public for scientific and/or educational purposes. The template further states that in giving permission, depicted individuals trust that Databrary Investigators will use professional judgment and uphold ethical principles in determining which excerpts to share. Generally speaking, Investigators should refrain from creating excerpts for all individuals in a data set.

**When should permission to share data be sought?** Permission to share recordings or excerpts with Databrary should be sought after a testing session is complete. This creates a clear separation between the consent an individual (or parent or legal guardian) gives to participate in research and the permission to share. It also ensures that individuals can give informed consent about what sorts of behaviors were recorded during a testing session, and thus, what will be shared. Separating the consent to participate from the permission to share may also reduce the likelihood that participant in-take procedures are unduly prolonged by the additional procedures involved in sharing data. No Investigator should coerce depicted individuals or research participants to share.

**How will the levels of permission granted by participants be linked to files?** Investigators who collect identifiable data will be responsible for ensuring that every depicted individual on a recording gives permission to share with Databrary. Permission to share may be secured in writing on paper forms or by electronic signature, but depicted individuals must be given paper records if electronic signatures are given. Paper copies of the signed permission to share form(s) must be kept in secured files comparable to those used for storing informed consent forms. If no permission to share is secured or written documentation of a research participant or depicted individual's preference is lost or damaged, then the data in question may not be shared.

There are two ways the level of sharing permission may be associated with an individual's data files.

Manual linking of permission to share levels with individual data files

Investigators may keep electronic records, in tab or comma-delimited text files, spreadsheets, or comparable formats that provide information about the participant and the level of permission granted. For example, such a file might have the following format:

participant-id	test-date	months-at-test	sex	share	excerpt
study01_7821	20121230	4.25	m	0	0
study01_7822	20121231	4.75	f	1	1
study01_7823	20130102	5.15	f	1	1

In some cases, multiple individuals may be depicted on a recording, some who are research participants, and some who are not. In this case, the permission file may look as follows:

participant-id	test-date	months-at-test	sex	share	excerpt
study01_7827	20121230	5.15	m	0	0
study01_7827_01	20121230	NA	NA	1	1
study01_7827_02	20121230	NA	NA	1	0
study01_7825	20130102	240	M	1	1
study01_7825_01	20130102	NA	NA	0	0

In this case, the recording featuring participant 7827 includes two other individuals (7827\_01, and 7827\_02) both of whom who gave level 1 (sharing) permission. But, the participant's parent or guardian did not give permission (Level 0), so that level will be binding on the entire recording. Similarly, for the recording featuring participant 7825, level 1 (sharing) permission was given, but a non-participant (7825\_01) refused to grant sharing permission. Thus, no recording depicting this individual may be shared. If the recording can be edited to remove this person from the recording, the modified file may be shared.

When uploading a data set to Databrary, contributors will enter the permissions information into a web form along with any data that they upload, so that Databrary can associate permissions levels and any demographic information available with each data file. Investigators are strongly encouraged to provide participant's age-at-testing and sex, but are not required to provide any information. A missing permissions level will be treated as level 0 both for sharing and for excerpts. Automatic tagging of individual data files with permission to share level Some Investigators may choose to use Databrary's Labnanny data management environment once it is developed. Under this scenario, Labnanny would be used by the Investigator to collect metadata about research participants and their sharing preferences at the time of each lab visit. That metadata would be entered into the Labnanny system during or shortly after a visit, and the Labnanny system would contain a centralized, electronic version of the manual file described previously.

Then, when an Investigator decides it is time to share a given data set, the sharing permission levels and related individual-level metadata identified at the time of data collection will be transferred to Databrary via the Labnanny software.

**How will levels of permission to share be maintained within the Databrary system once files have been uploaded?** Databrary will maintain a database for each study, and each study will include individual-level records for each data set for which sharing permission has been granted. Any data with permission level 0 will never be shared beyond the PI's research team, and in some cases these files may



not be uploaded at all. However, Databrary will still collect demographic and de-identified data associated with these participants. This is necessary in order to allow meta-analyses and subsequent data set users to determine what proportion of the shared or published data was actually shared with Databrary. For files with permission to excerpt, Investigators may create excerpts for use in public. Investigators may share excerpts with the public, only with other Investigators or Affiliates, or not at all.

Files denoting the individual-level sharing levels will accompany each data set when it is downloaded. Before downloading data, Investigators will be reminded of their promise to upholding sharing permissions granted by the original participants and to their other obligations that are part of the Databrary Investigator Agreement.

A future version of the Datavyu coding tool will use the permission to share file to provide reminder alerts/warnings to investigators about the level of access permitted for a given data set. But, for the time being, investigators will have to monitor the use manually.

**3.4 Sharing with Databrary** Before a data set may be shared with Databrary, investigators must evaluate recordings and associated individual-level metadata for volatile or sensitive content. Even if a participant (or parent or legal guardian) has given permission to share, the ultimate responsibility for determining whether data should be shared lies with the investigator. If an investigator determines that an individual's data set should not be shared, the investigator will mark it as such in Databrary.

**Investigators who contribute data provide study-level metadata** Databrary will include investigator-level, study-level, participant-level, and measure-level metadata. Researchers provide investigator-level metadata at the time they apply for Investigator status and preliminary study-level metadata at the time they apply to deposit a specific data set. Prior to or during the process of uploading participant-level data to Databrary, data contributors may provide enhanced study-level metadata. This may include manuscripts, posters or talks, figures, statistical analysis scripts, code books, or grant proposals. The default will be for study-level metadata to be open to the general public (open/unrestricted).

**Contributors upload/deposit data to Databrary** Investigators or Affiliate Investigators may either upload data to Databrary manually, using the Databrary site's upload tool or use the automated file transfer system embedded within Labnanny. Investigators or Affiliate Investigators must log-on to the Databrary system in either case. With manual deposit, Investigators or Affiliate Investigators must select the study for which data are being uploaded, the participant permission file, and the individual files that will be uploaded.

Labnanny users or users of other lab management software compatible with the (to-be-developed) Databrary application program interface (API) will simply indicate that a data set is ready to share. The Labnanny or other software will manage the upload/transfer process.

**Databrary translates video and metadata into common format(s)** Once data sets have been uploaded to Databrary, they will be automatically converted into a standard set of formats. The specific formats will be determined based on an assessment of how best to preserve the data for the long term using open (non-proprietary) formats, but UTF-8 and H264 video formats are leading candidates for text and video files, respectively. Databrary will publish information about the data formats accepted for deposits and those used and supported within the data sharing system.

**Databrary data quality checking** Once the file conversion process has completed, a set of automated and manual data quality checks will commence. The automated checks will assess whether sharing permission levels have been provided for every individual participant in the uploaded data set and whether personally identifying information have been provided in the uploaded data. Manual data quality checks will assess transcoding quality—whether video/audio is readable. If the quality of the transcoded recordings is poor, they will be checked against the uploaded original files. If the original files are in good condition, they will be transcoded again. If the original files are in poor condition, the contributor will be asked to upload the files again.

Once Databrary has certified that an uploaded data set is ready to share, the data contributor will receive an email indicating that fact. Contributors may then review the data set. When contributors are ready to share the data set, they will log on to Databrary and share the data set.

**Databrary publishes data with permanent URL** When a new data set is released for sharing, Databrary will announce the event on its website with the data set’s description, author(s), and URL. Databrary will also announce new data sets on Twitter, Google+, Facebook, and other social media, as appropriate. Investigators may choose to subscribe to a data set or another Investigator’s news feed to receive updates about new contributions or revisions to old ones.

Investigators may choose to share partial data sets and update or modify them as new data becomes available or new analyses are conducted. Databrary will maintain a versioning system similar to that used by [Github](#) to maintain records about data provenance. Significant modifications or additions to existing data sets will also be published via the web, email, and social media.

### 3.5 Ensuring Ethical Practices, Handling Violations

**Databrary Ethics Standards** Databrary intends to uphold the highest standards of integrity. These standards include the following:

- every Databrary staff member, Investigator, and Affiliate Investigator will complete training in research ethics involving human participants from an official research supervision entity and keep current their human participants training certification with such an entity.
- every Databrary staff member, Investigator, and Affiliate Investigator is expected to embrace and enact the principle of informed consent. This means that individuals who are recorded on video, audio or other means have the right to determine whether their data may be shared, and if shared, with whom.
- Investigators must accurately record the sharing permission preferences expressed by users. Investigators must transfer information about individuals’ sharing preferences accurately to Databrary.
- Databrary must label files accurately. Investigators who browse or download data must respect the sharing preferences expressed by participants. Databrary staff and Investigators are expected to embrace and enact principles of participant confidentiality. This means that all individuals who are recorded on video, audio or other means have the right to have their identities kept confidential.
- Investigators must not upload or share files that contain personally identifying information beyond that contained in recordings. Investigators must not create new or modify existing data files in ways that increase the risk that a participant’s confidentiality may be violated.
- If an individual participant’s identity is determined inadvertently, that information must not be shared with anyone, including members of the Investigator’s research team. Information about the individual’s identity may not be recorded in any form.
- Databrary staff, Investigators, and Affiliate Investigators are expected to monitor the use of Databrary for violations of the principle of informed consent by others and to report any concerns about possible violations to the Databrary Ethics Committee at ([ethics@databrary.org](mailto:ethics@databrary.org)) or via the Databrary website (<http://databrary.org/contact/>)
- In selecting excerpts of data for public viewing, Investigators must exercise professional judgment in order to minimize the chance that an individual participant could be harmed.
- In selecting data sets to share with Databrary, Investigators must exercise professional judgment to avoid sharing data of an offensive or prurient nature. Databrary reserves the right to remove data or data sets that are deemed to be of an inappropriate nature.

**Databrary Ethics Committee** Databrary will form an Ethics Committee to monitor the project, review and modify ethics policies, and to adjudicate ethics complaints. The Ethics Committee will consist of five members of Databrary’s Advisory Board, staff representatives from the NYU and PSU compliance offices,

three members of the community of Investigators, and the Databrary Staff. The committee will meet via conference call or webinar quarterly.

#### 4. Related Policies and Procedures

- [Databrary Investigator/Affiliate Investigator Agreement](#)
- [Databrary Permission to Share Recordings](#)
- [Databrary Permission to Share Self-Reported Health-Related Information](#)
- [Databrary Best Practices for Data Security](#)
- [Databrary Bill of Rights](#)

#### 5. Definitions of Terms

**People** Investigator: An individual eligible for principal investigator status at an Institution who has received authorization from Databrary to view, download, and contribute data.

Data Contributor: An Investigator who contributes Data to Databrary.

Affiliate Investigator: An individual who has been approved for access to Databrary under the sponsorship and supervision of an Investigator.

Research Staff: Persons authorized by the Investigator's Institution, excluding the Investigator or any Affiliate Investigator under the Investigator's supervision, who will have access to Data. Research Staff include project staff or students conducting dissertation or thesis research.

Depicted Individual: A person depicted in Recordings.

Research Participant: A person observed for purposes of research.

Third-Party User: An individual who wishes to obtain access to Databrary, but has not yet been granted an Investigator or Affiliate Investigator account.

Private Person: Any individual and any private (i.e., non-government) partnership, corporation, association, organization, or entity, including family, household, school, neighborhood, health service, or institution.

Representative of the Institution: A person authorized to enter into contractual agreements on behalf of an Investigator's Institution.

Minor: A person under the age of 18

**Entities** Research Supervision Entity: An Institutional Review Board or equivalent authority charged with supervising research at an Institution.

Institution: The university or research organization at which the Investigator will conduct scientific activities, including research using Data.

**Documents** Databrary Authorized/Affiliate Investigator Agreement: Written or electronic agreement that sets out the rights and responsibilities of Authorized and Affiliate Investigators in their use of Databrary resources.

Permission/Assent to Share Recordings: Written or electronic documentation that an adult or minor Depicted Individual has given permission to allow Recordings to be shared.

Permission/Assent to Share: Written or electronic documentation that an adult Depicted Individual has given permission to allow Identifiable Data to be shared.

Permission/Assent to Excerpt: Written or electronic permission or assent given by Depicted Individuals to allow Investigators to show or distribute Excerpts from Recordings.

Permission to Share Self-Reported Health-Related Information: Written or electronic permission or assent to allow De-identified Self-Reported Health-Related Information to be shared with Databrary.

**Data-types** Identifiable Data: Collections of Personally Identifying Information that might be used singly or in combination to determine the identity of Research Participants or other individuals depicted in Recordings.

Non-Identifiable Data: Collections of information that do not contain Identifiable Data.

De-identified Data: Collections of information from which Personally Identifying Information has been removed or altered in such a way as to prevent the identification of individual persons.

Recordings: Video or audio recordings or still images.

Personally Identifying Information:

- Full name
- National identification number (e.g., Social Security number)
- Internet Protocol address
- Vehicle registration plate number
- Driver's license number
- Health certificate or insurance number
- Fingerprints, or handwriting
- Credit card or financial account or access numbers
- Digital identity (e.g., Facebook, Twitter, LinkedIn, email account names)
- Date of birth
- Birthplace
- Address
- Telephone number
- Faces and voices. See also: HIPAA identifiers ([http://privacyruleandresearch.nih.gov/pr\\_08.asp](http://privacyruleandresearch.nih.gov/pr_08.asp))

Shareable Identifiable Data: Identifiable data for which a Depicted Individual has granted Permission or Assent to Share Recordings or Data.

Metadata: Information that describes how, where, why, and by whom Identifiable Data, De-Identified Data or Non-Identifiable Data were collected.

Self-Reported Health-Related Information: Information about an individual or minor child's health status that is reported by the individual or an individual's parent or legal guardian, not by a health-care provider.

Data: Any Identifiable Data, De-Identified Data, Non-Identifiable Data or Metadata.

Excerpts: An excerpt is a selection from a larger video or audio recording that illustrates a particular point or demonstrates a finding. It is typically no more than a few seconds or minutes in length or may be a still image. Investigators or Affiliate Investigators may select segments of a recording as an excerpt for display in public settings if the Depicted Individuals have granted sharing permission that includes Public Showing.

Digital Files: Video or audio recordings or photographs stored in an electronic form that may be readily shared with other people who are not Research Staff over electronic networks like the Internet.

**Actions** Sharing Permission: Specific rights for sharing and public release granted by Depicted Individuals in a Permission to Share Data document.

Investigator Promise of Confidentiality: A promise to a respondent or Research Participant that the information the respondent provides will not be disseminated without the permission of the respondent; that the fact that the respondent participated in the study will not be disclosed; and that disseminated information will include no linkages to the identity of the respondent. Such a promise encompasses traditional notions of

both confidentiality and anonymity. Names and other identifying information regarding respondents, proxies, or other persons on whom the respondent or proxy provides information, are presumed to be confidential.

Deductive Disclosure: The discerning of an individual's identity or confidential information through the use of known characteristics of that individual. Disclosure risk is present if an unacceptably narrow estimation of an individual's confidential information is possible or if determining the exact attributes of the individual is possible with a high level of confidence.

Public Showing: Playing an audio or video segment or showing a photograph to an audience that is not limited to Research Staff.

Research: A systematic investigation, including research development, testing and evaluation, designed to develop or contribute to generalizable knowledge. 45 CFR 46.102 (<http://www.hhs.gov/ohrp/humansubjects/guidance/45cfr46.html#46.102>)