**Data Management and Sharing Plan**

*Data Management and Sharing Plan template for studies using* ***Human Tissue with no human contact****. Remove grey and blue text prior to submission. Instructions are in grey italics. Example text is in* blue*.*

*Add an introductory statement of (1) the entity/ies providing data collection or management services for the study, such as the organization serving as the Data Coordinating Center, central and core labs, and (2) the main information systems used to process data such as REDCap, IDEAS, or a commercial Electronic Data Capture (EDC) system vendor*.

***Example text:*** Study data will be managed by the Clinical Research Informatics (CRI) Division, Department of Population Health Sciences, Joe R. and Teresa Lozano Long School of Medicine at the University of Texas Health Science Center at San Antonio (UTHSA) serving as the study Data Coordinating Center. The REDCap information system (described in the Equipment section of the application) will be used to manage study data.

*Add the following Core Lab statement for each core lab supporting the study.*

***Example UTHSA Institutional Core Lab text:*** This project will generate data through the [*insert the appropriate core lab/s Flow Cytometry, Structural Biology, Mass Spectrometry, Bioanalytics, and Single-Cell, Optical Imaging, and Biobanking]* UTHSA Institutional Core Laboratories.

**Element 1: Data Type --** Data Expected to be Collected and Used in the Study

*The data expected to be collected and used in the study section should cover all sources of data. Individual data elements or variables don’t need to be stated. We suggest using a table of data sources that cover all Scientific Data, e.g., data sources from which data are generated. The NIH DMS Policy requires statement of data sources. At a minimum, these should cover pre-specified primary and secondary outcomes intended for or that would otherwise support a* publication, are expected to be shared unless sharing is not possible*. Data Type: the general category/ies of data that are generated or acquired from the associated data source. For example, the Data Type “Carbamazepine concentration” would be generated from a data source, “Clinical Lab instrumentation” or more precisely, “Beckman Coulter AU Clinical Chemistry System; Emit ® 2000 Carbamazepine Assay”. Data source is the method through which data are generated or from which data are acquired; the more precise, the better. Types of data from the same source should be lumped into one row, such as a data type of “Sample Clinical Annotation” collected through a “UTHSA Clinical Data Warehouse query” data source. Data Volume should indicate the total order-of-magnitude estimate of data size from that source and for how many experimental or observational units, e.g., samples. Proposed Repository/ies should be chosen from among the discipline-specific repositories listed in* [*NOT-OD-21-016*](https://grants.nih.gov/grants/guide/notice-files/NOT-OD-21-016.html)*. If a discipline-specific repository does not exist, a repository should be selected from the general, i.e.,* [*not-discipline-specific repositories*](https://www.nlm.nih.gov/NIHbmic/generalist_repositories.html) *listed on the NIH page or one that otherwise meets the requirements for repositories such as persistent identifiers, web-searchable metadata, and adherence to the* [*FAIR principles*](https://www.go-fair.org/fair-principles/)*. Use of data standards for each data type and source should be indicated in the right-most columns of the table (EX: exchange standard, DE: standard data elements, CT: Controlled Terminology, DM: Common Data Model). They are named and enumerated later in the plan. The researcher and core labs will be the most likely to know if discipline-specific repositories or data standards exist. If none exist state N/A rather than deleting the table; if the NIH Program Official reviewing the plan knows of any that should be used, this will prompt them to say so during the Just In Time (JIT) period when the budget can more easily be adjusted to account for any additional cost incurred by implementing the standards or by a change in repository. The actual standards used are stated later in the DMS Plan.*

***Example text:*** in blue

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Table 1**: Data Types (Element 1A) and Sources (Element 1B) | | | | Data Standards | | | |
| Data Type | Data source | Data Volume | Proposed repository (If sharing) | EX | DE | CT | DM |
| Sample Clinical Annotation | Beckman Coulter AU Clinical Chemistry System; Emit ® 2000 Carbamazepine Assay | < 1 MB, 20 samples | Cell Depot | N/A | X | X | X |
| Single cell RNA sequencing (scRNA-seq) | NextSeq 6000 | < 1 GB, 20 samples | Cell Depot | Repository standard | X | X | X |
| Morphology from Flow Cytometry Images | Attune CytPix | < 1 GB, 20 samples | Cell Depot | Repository standard | X | -- | X |
| Flow Cytometry | Attune Flow Cytometer | < 1 GB, 20 samples | Cell Depot | Repository standard | X | -- | X |

Standards (EX: exchange std., DE: Data Element std. or common data element set, CT: Controlled Terminology, DM: Common Data Model)

Data collection and processing methods:

*In the data collection and processing methods section, for each data source in the table, state three things:*

1. *whether the data are generated by humans or electronically,* 
   1. *for electronically generated and processed data how the instrumentation is calibrated, such as use of standards or manufacturer-specified calibration procedures and how the testing processes are quality controlled like running against standards controls*
   2. *for data that are generated or processed by humans, how the processes involving humans are quality controlled, such as two raters with measured Inter-rater reliability.*
2. *any data processing that occurs after the data are generated or acquired by the researcher, and*
3. *the procedures that will be followed for human-based operations such as Standard Operating Procedures (SOPs), instrumentation Manuals, a laboratory procedure or protocol. SOPs, calibration of equipment and use of a second independent rater or independent review of a representative sample are common ways that this definition and control occurs* *for processes involving humans.*

*Systematically addressing these three things for each data source communicates that the researcher is in control of the data generation and processing and instills confidence that processes are in place to prevent errors, to detect and correct errors when they occur, and ultimately that the data will be capable of supporting conclusions.*

***Example text:*** Data are managed according to the lab protocols (flow cytometry, single cell RNA sequencing and imagery data). (1) Data entry of CRF data to specific folders by lab members on the day the experiment is conducted or per lab protocol. (2) Data cleaning, i.e., to identify missing and aberrant data and events will run to prevent, catch, or correct errors as early as possible, to assure that expected data and samples are received. The data cleaning will be carried out by the PI or research scientist. (3) FCS flow cytometry standard format, sc-RNA sequencing per Genome Sequencing Facility (GSF) SOP. (4) Transformation of the data to the analysis datasets and to the sharing format required by the repositories will be performed by the Biostatistician.

***Example UTHSA Institutional Core Lab text:*** The *[Core Lab name]* Institutional Core Laboratory will [*delete as appropriate: generate, store, curate, re-format]* and distribute the [*insert the type of data from Table 1 such as* Flow Cytometry, Structural Biology, Metabolomics, Lipidomics, Single-Cell, Optical Imaging, Biological sample inventory, *etc. …]* data. The \_\_\_\_\_\_\_\_\_\_ data are generated from the \_\_\_\_\_\_\_\_\_\_\_\_ analyzer/equipment. The analyzer/equipment is calibrated by \_\_\_\_\_\_\_\_ [*insert statement such as running standards prior to analyzing samples for the study, or according to the manufacturer’s specifications.]* Raw data are processed by \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. We anticipate \_\_\_\_\_ Mb/Gb/Tb of data will be produced (Table 1). *[Insert a statement about process control such as laboratory personnel in the core lab work according to a set of standard protocols/SOPs or Study-specific lab procedures will be created to assure consistency in data generation and processing.]* The core laboratory will store this data for a period of \_\_\_\_\_ weeks during active data collection. *If needed,* The data processing performed by staff of the core laboratory will consist of \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_. After this, all data will be transferred to storage devices under the responsibility of the PI.

Scientific Data to be Shared:

*In the Scientific Data to be Shared section, provide any needed additional detail for the data types and sources in Table 1*.

***Example text:*** Scientific data to be shared include final analytic datasets on which study publications are based and include pre-specified primary and secondary outcomes from Table 1 data sources.

Metadata and associated documentation to be shared:

*In the Metadata to be shared section, state any documentation that will be shared along with shared data. In general, documentation needed to interpret and use the shared data to replicate the analysis or to correctly reuse the data should be shared with the data. The list provided in the example is extensive and most often fewer associated documents need be shared.*

***Example text:*** Study metadata for shared data sufficient to meet repository and data citation requirements will be provided with shared data. Metadata describing shared data include final versions of the experimental protocol, data dictionary for shared data files, relevant procedures from the Lab Manuals of Operations (MOP), and Statistical Analysis Plan (SAP) unless limited by the repository.

**Element 3: Standards**

*In the Standards section, list the standards indicated in Table 1 that will be applied to the scientific data and associated metadata. If community or discipline-specific standards do not exist, that should be stated.*

***Example text:*** Flow Cytometry FCS files will be shared. Sc-RNAseq data will be shared in h5ad file where the expression matrix is stored in CSC as specified by the CellDepot Repository.

**Element 2: Related Tools, Software or Code**

*In the related tools, software or code section, list specialized tools, software, or code needed to access or manipulate shared scientific data. If no specialized software, code, or tools are needed, state this.*

***Example text:*** No specialized tools, software, or code are needed to access or manipulate shared scientific data.

**Element 4: Data Preservation, Access, and Associated Timelines**

*The data preservation, access, and associated timelines section should state when and how long the scientific data will be made available and through which repository.* *NIH has provided additional information to assist in selecting suitable repositories for scientific data resulting from funded research (NOT-OD-21-016). Local considerations for data sharing may apply such as:*

* *The UTHSA Library Records Retention Officer is available for consultation regarding the records retention policy and relevant resources at libraryrecordsmgtsvc@uthscsa.edu, or 210-450-8260.*
* *For very large data collections for which the investigator needs additional space local data storage prior to sharing or for retention locally, UTHSA IMS offers* [*local data storage options*](https://ims.uthscsa.edu/computer_networking/servers_storage.aspx)*.*
* *The Office of Technology Commercialization (OTC) is available to help with identification of and decisions about protection of Intellectual Property. If there is any chance that data should not be disclosed, contact* [*OTC*](https://otc.uthscsa.edu/)*.*

*In general, this information should be stated for each data source for which data will be shared. A table may be the most efficient and systematic way to present this information. Columns in the table include the following. Level of de-identification (if any human data are involved): Removal of the HIPAA identifiers and any information that could indirectly identify an individual. In general, de-identified information would be shared unless specific authorization has been obtained from research participants. It is best practice to state even de-identified data sharing in the Informed consent form for a study. If only animal data are used, the cell should be populated with “N/A animal data”, or “de-identified samples” as appropriate. Repository persistent identifier is the type of permanent identifier that the repository assigns to shared data. The repository metadata standard is the standard that is used by the repository to describe the study that generated the data and the shared dataset; the PI is often the best person to complete this information and will need to provide this study-level descriptive information to the repository along with deposition of data to be shared. Data sharing start is the time at which you anticipate data sharing occurring. Note that sharing upon publication is expected unless there is justification for not doing so; in which case, the justification should be stated below the table. Data sharing duration is the time period for which data will be shared. The records retention period is optional and was added to indicate the local retention duration.*

**Table 2**: Data Preservation, Access, and Associated Timelines  ***Example text:*** in blue

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Data Type or Source | Repository persistent unique ID | Repository Metadata Standard | Data Sharing Start | Data Sharing Duration | Records Retention Period |
| Sample Clinical Annotation | Project DOI | h5ad | Publication\* | 10 years | AV |
| Single cell RNA sequencing (scRNA-seq) | Project DOI | h5ad | Publication\* | 10 years | AV |
| Morphology from Flow Cytometry Images | Project DOI | FCS | Publication\* | 10 years | AV |
| Flow Cytometry | Project DOI | FCS | Publication\* | 10 years | AV |

\* Data will be shared at the time of an associated publication or end of the performance period, whichever comes first. AV: retention period as long as Administratively Valuable according to UTHSA record retention schedule.

***Example text:*** The scientific data generated from this study will be shared in the Cell Depot Repository. We anticipate 2 publications from this work. Analysis datasets for each publication will be shared through the NCI/MPD repository and indexed by a project-level repository generated DOI. Row-level shared data are linked by a unique study ID. Data for this study will be retained in the PI’s lab following the award period. Data for this study will be retained in the UTHSA institutional archive for 10 years after the award period as specified by the Records Retention Schedule described in UTHSA HOP 2.2.1

**Element 5: Access, Distribution, or Reuse Considerations**

The access, distribution or reuse considerations section should s*pecify any of the following factors that impact sharing.*

1. *Factors affecting subsequent access, distribution, or reuse of scientific data*
2. *Whether access to scientific data will be controlled*
3. *Protections for privacy, rights, and confidentiality of human research participants*

***Example text:*** Access to shared data will be controlled according to the repository policy which requires a request and merit review and will be available according to repository processes. The de-identified shared data will shared under a repository standard Data Use Agreement (DUA). No additional protective measures will be taken.

**Element 6: Oversight of Data Management and Sharing**

The oversight of data management and sharing section states how compliance with this DMS Plan will be monitored and managed, frequency of oversight, and by whom at out institution as required by the NIH DMS Policy. The institutional oversight text below has been provided by the UTHSA VPR’s Office and OSP. Aside from inserting the PI’s name and statement of any Quality Management System in place in the lab/research group/research setting, we strongly recommend no changes. Description of the QMS under which the data will be collected and managed instills confidence in readers that the data will be capable of supporting research conclusions.

***Text Provided by UTHSA VPR’s Office and Office of Sponsored Programs:***

The PI for the project *[Investigator Name]* at the University of Texas Health San Antonio (UTHSA) will be responsible for the day-to-day oversight of data management activities and data sharing. The PI will conduct regular meetings with key study personnel to ensure quality of data entry and the timeliness of data collection, analysis, storage, and sharing. Program services are available to assist investigators in the creating and internal monitoring of NIH compliant data management plans. The Vice President for Research Office and Office of Sponsored Programs at UTHSA have created a data management and sharing plan compliance program as part of the process for submitting the annual NIH progress report. In addition, the Offices will conduct an audit at the time of publication to identify award specific data DOIs and include that information in the annual report. If needed, the Offices will work with the PI to evaluate and update the DMS as the project progresses.

*The PAGE LIMIT was initially stated by NIH as 2 pages. We understand that the limit has been extended, but a recommendation of 2 pages is still stated on the NIH site.*

[*https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan*](https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan)