CIRM
CALIFORNIA INSTITUTE FOR REGENERATIVE MEDICINE

CIRM Data Sharing and Management Plan (DSMP)

Guidelines for Discovery Awards

PURPOSE OF DSMP

To leverage CIRM-funded data and enable reuse of data by other researchers, CIRM DISC awardees are expected to share their data consistent with FAIR (Findable, Accessible, Interoperable, and Reusable) and CARE (Collective Benefit, Authority to Control, Responsibility, and Ethics) data principles and reflective of practices within specific research communities. Development and execution of the CIRM **Data Sharing and Management Plan (DSMP)** is intended to facilitate:

- **F**indability of data through a public dashboard, the CIRM Data Explorer,
- **A**ccessibility of data by deposition in data repositories accessible to other researchers, and
- **I**nteroperability and
- **R**eusability of data by associating deposited data with necessary and sufficient metadata.

The DSMP intends to capture information (metadata) about the biological samples used for data generation, the methods and data analysis pipelines used during the funded studies and the Data Use Limitations that apply to the data. The information assembled in the DSMP should provide 1) sufficient detail for another researcher to repeat the data processing stages (replicate results), 2) sufficient context to use the data in new ways with confidence in their interpretation of the data and its provenance (reuse data), and 3) sufficient information on data use limitations, enabling adherence to CARE principles.

The expectation is that information captured in the DSMP will be included when data is deposited in a repository, with the goal of making the data FAIR. Data submission rules of data repositories must be followed.

CIRM appreciates your careful attention to this matter and your support of these aims.

SCOPE OF DSMP

CIRM requires DISC awardees to manage and preserve raw data, processed data and metadata, and share applicable data[1] and metadata, i.e. make applicable data and metadata available to the broader scientific community through data repositories accessible to other researchers. CIRM expects all applicable data generated under a CIRM award to be shared no later than the time of publication or by the end of the award period, whichever comes first. Even data not used to support a publication, including null or negative findings, are considered data.

For some programs and data types, CIRM has developed specific data sharing expectations (e.g., data types to share, relevant standards, repository selection, timelines) that should be reflected in a DSMP. When no specific CIRM data sharing expectations apply, researchers should propose their own approaches to data sharing and management.

CIRM requires that anyone deriving data from living humans must be prepared to ensure privacy and confidentiality protections (i.e., de-identification, Certificates of Confidentiality, and other protective measures), in accordance with applicable federal, Tribal, state, and local laws and regulations.

DATA TERMINOLOGY

- **Data:** The Intellectual Property Policy for CIRM Awards defines "Data" as: Scientific, clinical, or technical recorded information derived during the Project Period of an Award, regardless of form or the media on which it may be recorded, but not any of the following: financial, administrative, management data, other information incidental to contract administration, preliminary analyses, drafts of scientific papers, plans for future research, peer reviews, or communications with colleagues. "Data" excludes physical objects (e.g., laboratory samples).

- **Data generation:** generation of raw data

- **Data processing:** all data processing steps (dry lab) following generation of raw data

- **Data production:** overarching term, referring to both data generation and data processing

- **Data products:** the result of each data generation step and each data processing step (Each data product should be listed in the DSMP Metadata Catalog)
    - **Raw data:** data produced by an instrument (e.g., raw sequence data) or by other methods, such as measurements and surveys, or obtained from a data repository
    - **Processed data:** data produced from raw data and from subsequent processing steps (e.g., quantification files, alignment files, etc.)
    - **Final processed data:** data produced from last processing step (e.g., aggregated quantification, etc.), on which conclusions are based

- **Metadata:** data that provide additional information needed to make shared raw and processed data findable, interpretable and reusable. Metadata information is requested in the DSMP.

[1] Definition of *applicable data* in 'Data terminology'

Metadata categories in CIRM Data Explorer

- **Data Product Details**: methods used for data generation (machine, instrument), data processing (software toolkits, pipelines) and data sharing (data repositories).
- **Biological Material Details**: information about the source and modifications of the biospecimens and the final cell product used for data generation
- **Goal of Experiment**: information about diseases studied and/or biological questions addressed
- **Sample Preparation:** information about experimental approaches used to prepare the sample for data generation
- **Protocols & Publications**

Additional metadata

- **Map of unique identifiers:** a document that details the persistent unique identifiers or other standard indexing tools, assigned by data repositories and used to track projects and samples, enabling other researchers to find related data deposited in different repositories.
- **Data Dictionary:** a document that defines field names, such as male/female is represented by 0/1 or 1/2 or m/f etc. (only needed if not using an existing Data Standard, such as this LOINC code for sex at birth)

- **Data standards:** guidelines or formal rules for producing, structuring, naming, and describing data.
  - CIRM expects that an awardee will apply data standards that are common to their field of study in the production of data and to metadata that are deposited in a Data Repository. Examples of data standards can be found at CDISC or LOINC.
  -
- **Data sharing:** making data available to the broader scientific community by depositing in a data repository accessible to other researchers

- **Applicable data:**
  - All data that are needed for another researcher to replicate results and to reuse data. Minimally this includes raw data, final processed data and metadata.
  - CIRM does not anticipate that researchers will preserve and share all data produced in a study. Researchers should decide which data to preserve and share based on ethical, legal, and technical factors that may affect the extent to which data are preserved and shared. The rationale for these decisions must be provided in the DSMP Questionnaire.
  - Data not used to support a publication, including null or negative findings, can be considered applicable data.

- **Replicate results:** another researcher uses shared data and same code/software as original researcher to obtain the same results

- **Reuse data:** another researcher uses shared data and different tools / software to obtain new results, or uses shared data in combination with their own data

CIRM
CALIFORNIA INSTITUTE FOR REGENERATIVE MEDICINE

INSTRUCTIONS FOR SUBMITTING AND UPDATING DSMP

The CIRM DSMP has 3 components:

- Metadata Catalog
- Data Use Limitations (DUL) Institutional Certification
- Questionnaire

Together, these components outline how the data, anticipated, and then generated, for the funded project, will be shared with the scientific community.

For all **data** you propose to generate, please prepare a Data Sharing and Management Plan (DSMP):

1. Join/log into **CIRM Data Explorer**
2. Complete the **Metadata Catalog** for expected data
3. Complete, sign, and submit the **Data Use Limitations (DUL) Institutional Certification form**
4. Complete the **DSMP Questionnaire**

The Metadata Catalog will be a living record:

- **Initial Metadata Catalog:** Prior to CIRM issuing the Notice of Award (NoA), the initial Metadata Catalog is submitted to CIRM. It contains minimal information about the anticipated data types and experimental design of the project.
- **In progress Metadata Catalog:** Throughout the project, the Metadata Catalog is continually updated as data is produced and metadata is collected. The goal is to avoid last minute scrambles to assemble all information needed for data deposit at the end of the project. The most up to date version of the Metadata Catalog is submitted to CIRM as part of each scientific progress report and is subject to CIRM review and approval.
- **Final Metadata Catalog:** At the end of the award, the Metadata Catalog, as well as the DUL form, Map of Unique Identifiers and Data Dictionary, are finalized and serve as a record of metadata that is shared with raw and processed data.

Once data has been deposited by the awardee, the metadata provided in the Metadata Catalog and the DUL information, will be made public and displayed in the CIRM Data Explorer, a dashboard that scientists can use to discover CIRM-funded data and determine where they are deposited.