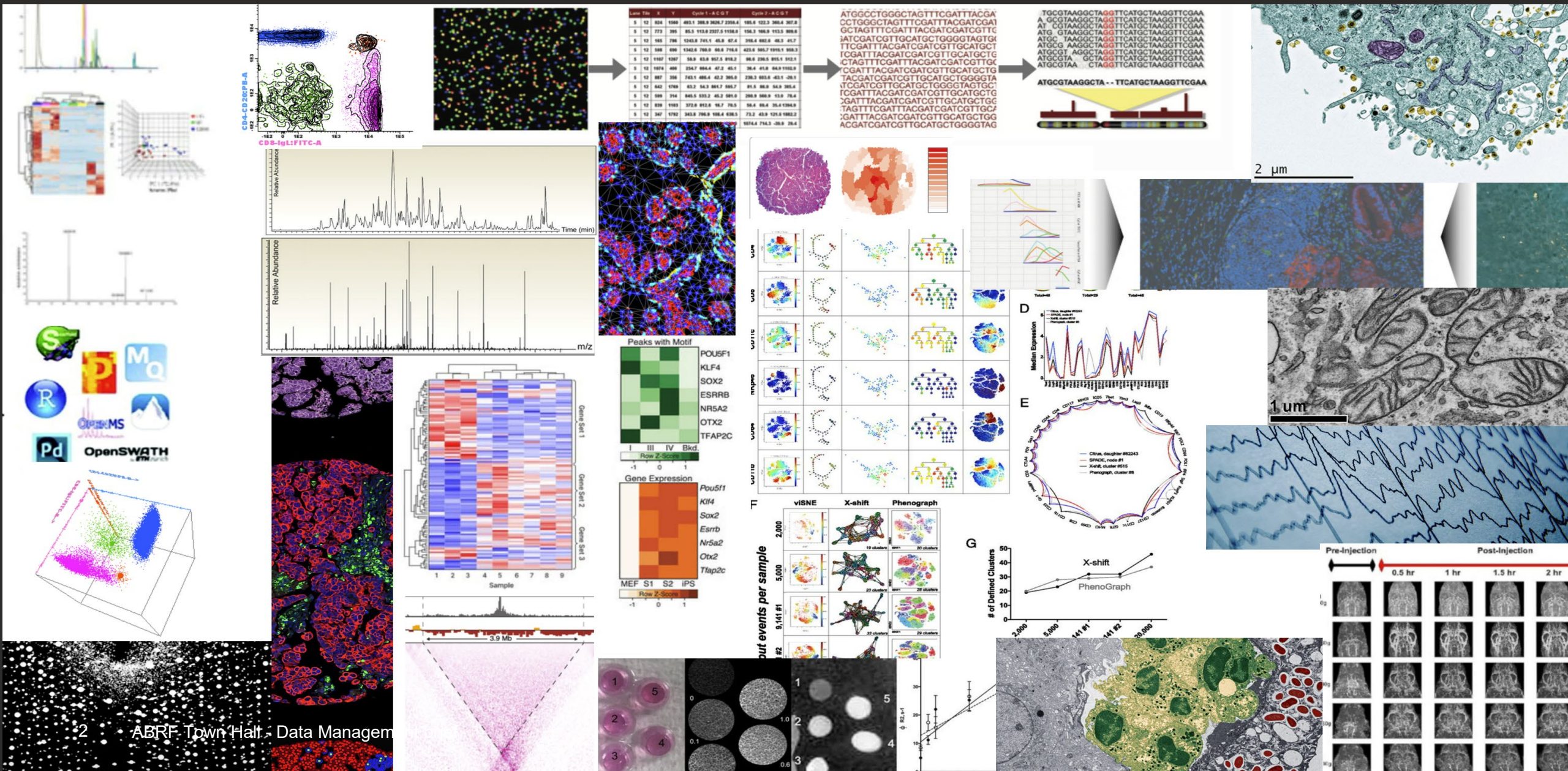
NIH Data Share Plan Requirements: Impact on Cores

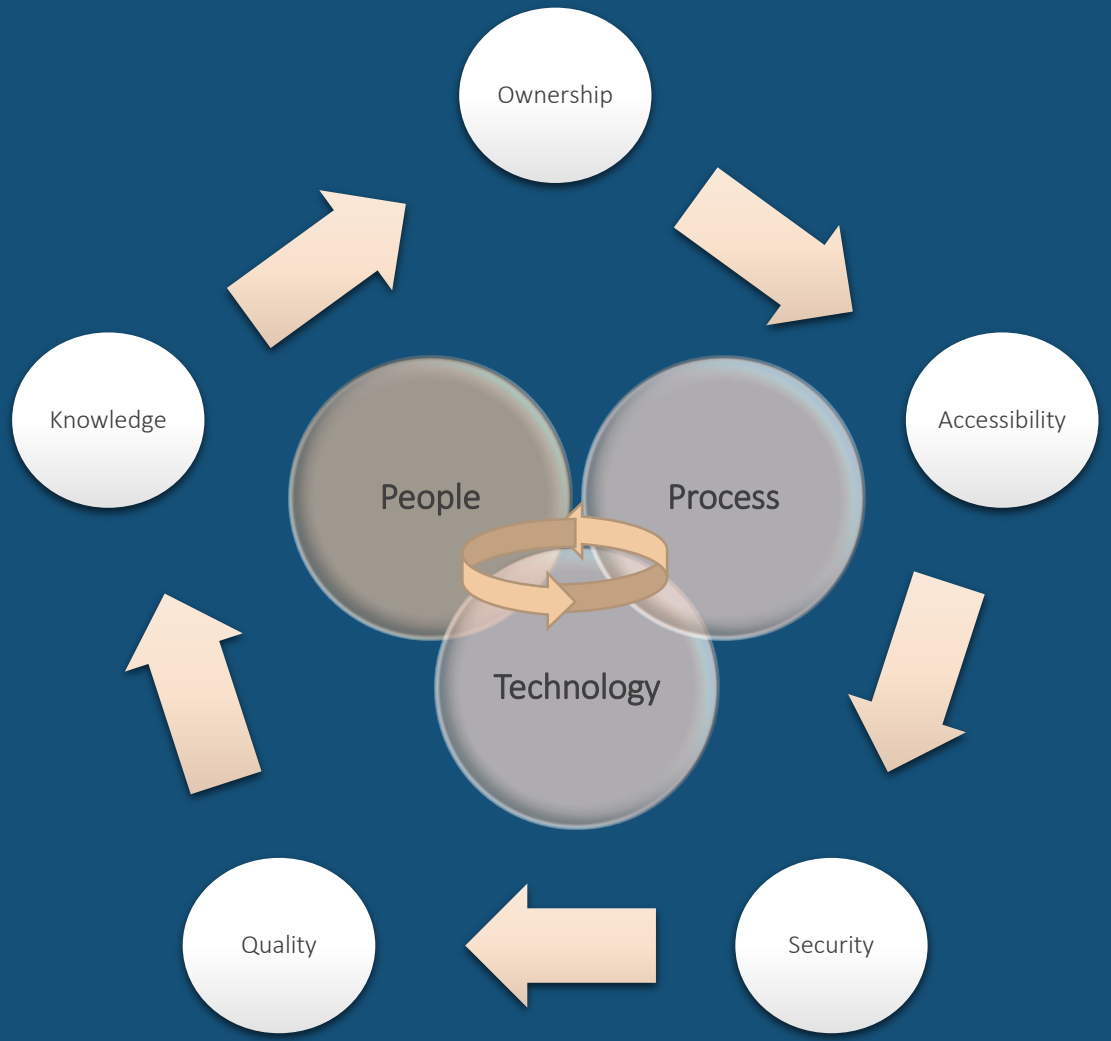
Nicole White & Sheenah Mische
ABRF DMWG 2022

Data is our key deliverable



Data Management?

A practice used to organize and maintain data processes to meet ongoing needs within the information lifecycle process.



Data Management vs. Data Governance

Data Management

- Method of how data is managed and arranged
- Implementing structure and how data is moved
- Focuses on technology and moving data to appropriate locations for access
- Manages the data as experts in the field, instrument, QC values, files types, etc.

Data Governance

- Provides a set of rules and policies on how data is managed
  - Retention policies
  - Archive polices
  - Storage costs
  - Storage rights and access
- Involves various members of the organization to develop
- First building block toward data management
- Focus is on business strategy

Some definitions:

- **Research Data:** any recorded factual material generated in research and commonly accepted in the scientific community as necessary to validate research findings, including all information useful for the reconstruction and evaluation of reported results of the Research and the events and processes leading to those results, regardless of the form or the media on which they may be recorded.
- **Data Provenance**: The origins, custody, and ownership of Research Data. Because datasets are used and reformulated or reworked to create new data, provenance is important to trace newly designed or repurposed data back to their original datasets. This includes the acquisition and management of information, including program configurations and the entire computational environment.
- **Ownership:** Research Data belong to the institution unless the institution expressly waives ownership rights under an applicable Sponsored Research Agreement, in which event the provisions of the Sponsored Research Agreement shall control.
- **Responsible Party**: the PI of the funded grant

Research results developed with NIH funding should be broadly available to the research community for furthering research

- Research papers and data products are key outcomes of the science enterprise
- Research sponsors are increasingly recognizing the value of research data. As a result, most funders now require that sufficiently detailed data management plans be submitted as part of a research proposal

- Final NIH Policy for Data Management and Sharing (NOT-OD-21-013) (released on Oct. 29, 2020, effective Jan 2023)
- Supplemental materials:
  - Elements of an NIH Data Management and Sharing Plan (NOT-OD-21-014)
  - Allowable Costs for Data Management and Sharing (NOT-OD-21-015)
  - Selecting a Repository for Data Resulting from NIH-Supported Research (NOT-OD-

NIH Data Management and Sharing Policy

- DMPs state how research data will be managed and shared
- DMPs need to comply with requirements from other relevant NIH institutes, Centers or Office(*e.g., DMPs for genomic data need to comply with the NIH GDS policy*)
- DMPS are reviewed but not scored
- Allowable costs include data curation and documentation, local data management and the preservation and sharing of data
- DMPs should include data type, related tools, SW and/or code, data standards, data preservation, access and associated timelines, access distribution and reuse considerations, oversights of data management and sharing

NYU Langone Health

ABRF members support FAIR Data Principles and ensure data provenance

- Instrument Standardization
- SOPs
- QA/QC
- Unbiased data acquisition
- Unbiased data analysis
- Documentation
- Source data | metadata | shared data
- Transparent reporting

Jianwu Wang et al., Proc IEEE Int Conf Big Data. 2015 ; 2015: 2509–2516. doi:10.1109/BigData.2015.7364047

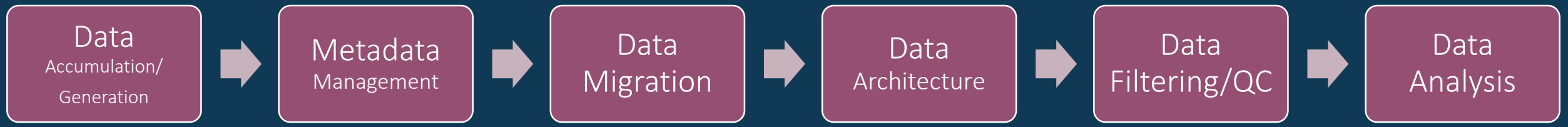
FAIR Principles

In 2016, a global consortium of researchers published "The FAIR Guiding Principles for scientific data management and stewardship" in *Scientific Data:*

- **F**indable
  - Data is described with rich metadata; data is assigned globally unique and persistent identifiers
- **A**ccessible
  - Data and metadata are retrievable
- **I**nteroperable
  - Data can be integrated with other data; data can be used in multiple applications or workflows
- **R**eusable
  - Data and metadata are well-described so they can be replicated

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) Sci Data 3, 160018. https://doi.org/10.1038/sdata.2016.18

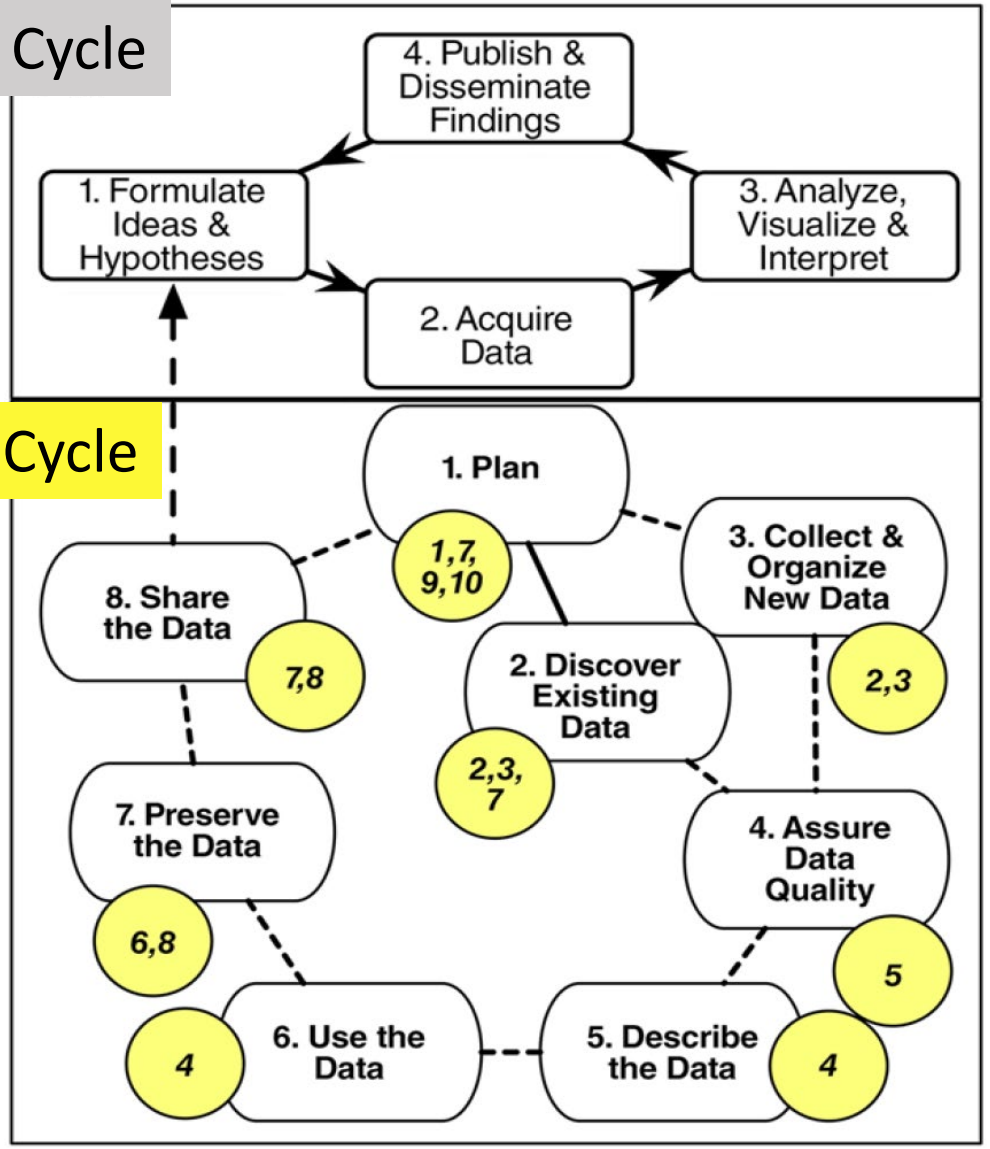
Key Parts to the Data Management Process



What is a Data Management Plan?

Research Cycle

Data Life Cycle



1. Determine Sponsor Requirements
2. Identify the Data to Be Collected Types | Sources | Volume | Data and file formats
3. Define How Data Will Be Organized
4. Explain How Data Will Be Documented
5. Describe How Data Quality Will Be Assured
6. Data Storage and Preservation Strategy
7. Define Project's Data Policies
8. Describe Data Sharing
9. Assign Roles and Responsibilities
10. Prepare a Realistic Budget

Michener PLOSComp Bio (2015) | DOI:10.1371/journal.pcbi.1004525

How can Cores facilitate Data Sharing Plan for NIH grant proposals? Employ Best Practices for All Data Management

***Persistent Unique Identifiers**:* Assigns datasets a citable, persistent unique identifier (PUID), such as a digital object identifier (DOI) or accession number, to support data discovery, reporting (*e.g.,* of research progress), and research assessment (*e.g.,* identifying the outputs of Federally funded research). The PUID points to a persistent landing page that remains accessible even if the dataset is deaccessioned or no longer available.

***Reuse**:* Enables tracking of data reuse (*e.g.,* through assignment of adequate metadata and PUID).

***Provenance**:* Maintains a detailed logfile of changes to datasets and metadata, including date and user, beginning with creation/upload of the dataset, to ensure data integrity.

***Metadata**:* Ensures datasets are accompanied by metadata sufficient to enable discovery, reuse, and citation of datasets, using a schema that is standard to the community the repository serves.

***Curation & Quality Assurance**:* Provides expert curation and quality assurance to improve the accuracy and integrity of datasets and metadata.

***Common Format**:* Allows datasets and metadata to be downloaded, accessed, or exported in a standards-compliant, and preferably non-proprietary, format.

***Long-term sustainability**:* Has a long-term plan for managing data, including guaranteeing long-term integrity, authenticity, and availability of datasets; building on a stable technical infrastructure and funding plans; has contingency plans to ensure data are available and maintained during and after unforeseen events.

Retention: Data is maintained in accordance with this Policy for the longer of (i) three (3) years after the final project close-out or (ii) six (6) years after any reporting, publication, presentation, or use in any grant application by the researcher of such Research Data.

Research Data relating to a student project must be retained at least until the degree is granted or it is clear that the student has abandoned the work.

***Secure**:* Provides documentation of meeting accepted criteria for security to prevent unauthorized access or release of data, (i.e., ISO 27001 (*https://www.iso.org/isoiec-27001-informationsecurity. html*) or NIST 800–53 controls (*https://nvd.nist.gov/800-53)*

***Privacy**:* Provides documentation that administrative, technical, and physical safeguards are employed in compliance with applicable privacy, risk management, and continuous monitoring requirements.

Example of a Core DMP created using *DMPTool*

Data Collection

- **What data will you collect or create?**
  - Image data
- **How will the data be collected or created?**
  - Image data will be generated by imaging platforms: .... in any of the following formats : … file formats are readable with the open source tool Bioformats.
  - Image data is stored and accessible through *OMERO Plus* and will be made publicly available through NYUGSM public *OMERO* interface and cataloged at NYULMC data catalog.

Documentation and Metadata

- **What documentation and metadata will accompany the data?**
- **Microscope Acquisition metadata**: available through *OMEROPlus* interface and stored within each image file.
- **Images generated post Immunostaining (any form), RNAscope, Spatial Transcriptomics and /or image analysis**: descriptions of detailed experimental conditions, reagents and/or code are provided via .txt, excel, .pdf or r code.

Ethics and Legal Compliance

- **How will you manage any ethical issues?** images contain no PHI or personal identifiers.
- **How will you manage copyright and Intellectual Property Rights (IP/IPR) issues?** Institutional management

Storage and Backup

- **How will the data be stored and backed up during the research?**
  - Storage of digital images is provided by NYUGSM through a research Isilon that is backed up weekly and mirrored at 3 different sites
  - All Exp Path images are stored within the research Isilon at a directory linked to *OMEROPlus* image data management server. Researchers can access data that was generated for their lab by directly downloading raw files or through OMERO webclient (a browser based interface).
- **How will you manage access and security?**
  - Access and security managed via LDAP (MCIT)
  - NYULH ID and password are required to access *OMERO* webclient, onsite access to the Research Isilon image directory (*OMERO* storage) is granted through LDAP credentialing. Advanced vpn permissions required for offsite access to *OMERO* storage
  - Collaborators must request access from MCIT with PI approval

cont'd Example of a Core DMP created using *DMPTool*

Selection and Preservation

- **Which data are of long-term value and should be retained, shared, and/or preserved?**
  - We preserve every image acquired by our systems and do not delete data once it is distributed to the research lab through *OMERO* webclient
  - Public data sharing is the responsibility of PI
- **What is the long-term preservation plan for the dataset?**
  - Under MCIT procedures data not accessed for >3 years → LT storage (details)

Data Sharing

- **How will you share the data?**
  - Image data can be shared publicly through *OMERO* webclient public user when requested by the principal investigator. accompanying metadata/descriptions will be made available by the PI through NYU Health Sciences Data Catalog who will provide links to *OMERO* webclient records and DOIs
- **Are any restrictions on data sharing required?**
  - No

Responsibilities and Resources

- **Who will be responsible for data management?**
  - ExpPath responsible for data preservation and provenance for all images produced
  - Individual PI responsible for downstream data annotations and public sharing
- **What resources will you require to deliver your plan?**
  - *OMERO* server access, *OMERO* storage access
  - Data repository resources
  - MCIT dedicated personnel
  - SW maintenance *Glencoe Software*

With many thanks to

- Valeria Mezzano NYUGSOM Exp Path
- Nicole Contaxis NYULH Health Sciences Library
  - Data Services Team, and Lead of the NYU Data Catalog
- FASEB DataWorks!
- ABRF DMWG members
- ABRF