Sinai Health | Lunenfeld-Tanenbaum Research Institute

Data Management Strategy

Background

In 2017, a RDM Advisory Committee was formed by the Toronto Area Health Sciences Network (TAHSN) hospitals affiliated with the University of Toronto and have encouraged hospitals to develop their own strategies while coordinating with the University of Toronto libraries.

In 2018 the Canadian Tri-Agencies released a draft of the *Tri-Agency Research Data Management Policy for Consultation,* which will require institutions to create an institutional RDM strategy[3]. This strategy complies with that requirement.

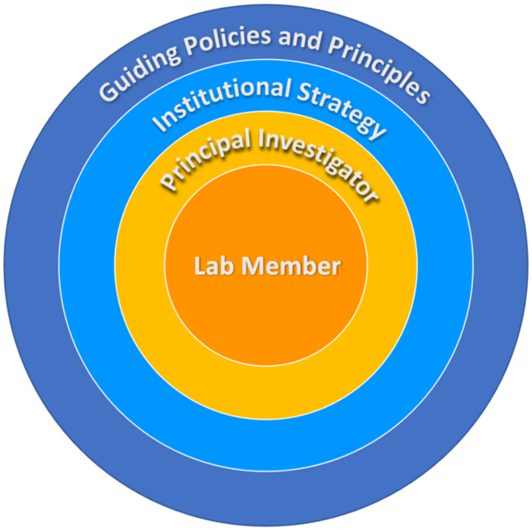
The aim of this strategy is to develop and foster a culture that supports researchers in adopting responsible RDM practices. In addition, as the research landscape evolves, this strategy will evolve to meet the changing needs and requirements that the researcher faces.

The Strategy

The overall strategy is wholly encompassed by the "Guiding Policies, Principles and Legal Requirements" that our employees and researchers abide to. It will also strive to be in alignment with similar Data Management Strategies at the University of Toronto and the TAHSNr Member Workgroup.

The aforementioned will mediate the "Institutional Strategy", where we set out guidelines, procedures, infrastructure, and allocation of resources for the scientists.   At the Principal Investigator level, the senior scientist will set out the Project Data goals, data flows, data chain of custody, and data lifecycles. At the core, the Lab Member will execute the data plan for the project as defined by the Principal Investigator.

Each section will be discussed in detail below.



Guiding Policies, Principles and Legal Requirements

This strategy will be consistent with the following:

- Sinai Health System policies, statements, and agreements
- Tri-Agency Research Data Management Policy
- Tri-Agency Framework Responsible Conduct of Research (2021)
- Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans - 2nd edition
- Tri-Agency Statement of Principles on Digital Data Management
- Sinai Health System IT Policy - I-h-15
- Personal Information Protection Electronic Documents Act (Federal PIPEDA – 2004)
- Personal Health Information Protection Act, (Ontario PHIPA 2004)
- Access to Information Act (Federal, 1985)

This strategy will aim to follow Information Technology Standards such as:

- ISO 27001:2022 Information security, cybersecurity and privacy protection
- FAIR Principles – the set of guidelines published on March 2016 set out with specific emphasis on enhancing the ability of machines to automatically find and use the data, in addition to supporting its reuse by individuals
    - https://www.nature.com/articles/sdata201618

This strategy will honour and respect objectives as outlined in:

- C.A.R.E. [indigineous]

Institutional Strategy

At the Institutional level, the primary objectives of the strategy are the following:

- At the highest level - provide resources, infrastructure and training for the scientific staff to adhere to the paths set out in **Guiding Policies, Principles and Legal Requirements.**

Currently, this has been accomplished with heavy investments into Institutional standardized hardware, network, security, and software infrastructures. This ensures a common secure platform that researchers will build upon their individual project data management plans on. Examples of this is the provision of the unified networking hardware ecosystem, up-to-date firewall systems, unified security user management authentication systems, and centralized data storage and backup systems. The infrastructure will be the base for the scientist to build their lab and project's "**Data Management Plan**". This will define whether institutional resources are sufficient, or procure further funding to extend the base infrastructure to suit their project needs.

How have we been achieving these platform goals?

The Hardware and Base Infrastructure

Networking

The Institute has been providing a stable, steadily updated networking infrastructure. Networking architecture design is focused upon stability and redundancies that are resilient against common network failures. For example, critical network traffic routes are designed with dual redundant fibre optic paths that allow communications to continue through major physical failures. Primary core trunks are at least 40 Gbps, and leaf nodes are beginning to be upgraded to a base of 10 Gbps speeds to the end user. This is in anticipation for the large data requirements in upcoming research projects. As the end-of-life period for this equipment approaches, we will replace them with modern trunk speeds of at least 100 Gbps switches.

Data Centre and Virtualized Host Infrastructure

In order to provide a stable, safe, and secure platform, all core systems are virtualized with redundant failsafe hardware configurations. The primary core Data centre that houses the majority of server and storage equipment is located in the Main Computer Room (MCR) of the Sinai Health System building – 600 University Ave. The Secondary (Mirror) data centre is located across the street at the 25 Orde St. building. Regular hardware refresh schedules are executed as these core equipment nears end-of-life status.

Storage

The Institute has recently made large investments in Enterprise-grade best-in-class petabyte size storage systems with integrated remote site backup. This ensures a stable trusted storage system which the Institute scientists can reliably access and store their data on. Previous versions of data can also be retrieved by the user, within reasonable size and time limits.

Security

On the outermost portion of the Institute network (ingress/egress) - fast, Enterprise-grade dual redundant firewalls are situated at the edge of the network to protect the internal systems against unauthorized access.

The Institute maintains core user authentication systems based on Microsoft Active Directory and Unix/Linux based LDAP Systems. As such, all laboratory and administrative systems are locked down via these systems to ensure that the data is accessible only by the correct personnel. HR processes are linked with IT user database maintenance so that users are properly removed from access when necessary.

Access to data stored on Institute servers must be pre-authenticated via one of the above mentioned authentication systems. Users are provisioned secured storage that is accessible only by that user, and collaboration group work data is situated in server locations that are specifically set to allow multiple sets of users to access.

External access is provisioned via VPN servers that are slated to be replaced with next generation models in 2023.

The Institute has also recently made investments in advanced threat protection appliances to provide network-wide visibility and intelligence to detect and respond to targeted attacks and advanced threats. This will greatly enhance the Institute's ability to protect its data from external threats. The Institute will also be making investments by this quarter of the year into security product subscriptions that will more deeply analyze incoming data streams and increase security on mobile devices.

The above security systems are in place as accordance to privacy and intellectual property requirements.

<u>The Software</u>

A base level of core supported operating systems and software, with funding for sufficient upgrade licences, ensures that the systems remain secure and the data remains viable and uncorrupted for the course of a scientific project. Standardization on the most commonly used software ensures data compatibility and consistent metadata formats in the present and ensures a higher success rate in being able to access the data in the far future.

The Data

As the definition of what is "data" is broad, this must ultimately include in the planning, the physical forms of data, for example, but not limited to: lab books, paper notes, tissue samples, measurement readouts, photographic pictures/negatives, analytical blots, cell line libraries, genomic DNA libraries, peptide libraries, and compound libraries. This part of the strategy mostly focuses on the electronic forms of data medium, and will note that it is currently deficient in including strategy discussions on physical data.

For most research applications, the Institute has provisioned enough electronic data storage space for typical scientific activities. However, we are seeing that more research is becoming more data-centric, and the available storage space is forecasted to be insufficient to the projected data rate of growth.

Short term – Data creation and collection and classification

Historically, the Institute has made available a set of standardized enterprise-grade storage systems for scientific use and the labs at the LTRI are well versed in depositing their work onto these systems for safe keeping. The majority of equipment where data is created is already securely locked down with the available authentication systems and the researchers are aware to follow data deposit protocols to ensure that the data is securely saved onto Institute servers.

Data classification, labelling, and enforcement exercises are currently primarily left to the Senior Scientist and subordinates to determine. *Immediate future efforts* will need to survey the local requirements and the common standards used at other institutions. A committee will need to be formed to digest the survey results and provide a data classification standard for the Institute members to follow.

Medium Term – Analysis, Collaboration, and Metadata labelling

As data is collected, analysed or completed, the security access, privacy and intellectual property requirements will be periodically reviewed by the lab member or the Senior Scientist throughout the project life for any changes or deficiencies that need to be addressed.

Long Term – Documentation, Metadata labelling, Archival

As the project draws to an end, the work should shift towards cleaning, final documentation, classification, labelling, and storing away the data. This may involve existing internal institutional resources, or utilizing external archival warehouses. Proper documentation is essential for future use of this data.

Long Term Data Archival – Currently under review

Formal procedures are in the process of being defined for handling research electronic and physical data when a Senior Scientist retires or is no longer employed at the Institute. Currently handled ad-hoc, mainly dependent on remaining lab staff to identify and direct handling of the data.

Data Curation Committee needed – Define policy and strategy. Provide oversight on what data and physical records should be kept, how to be archived, how/what should be destroyed, make recommendations on equipment and resources required.

- Standard operating procedures when closing down a lab
- Historically, data has been archived by leaving the data "in-place" on production systems. This has worked somewhat due to the relatively slow data growth in the past. However, the rate of data growth has expanded exponentially over the past decade and has pushed current production storage systems to the limit.
- Will need to investigate whether our researchers can deposit data into the **University of Toronto Dataverse** servers, or do we need to create a LTRI Dataverse server footprint, or is a "Dataverse" even an appropriate solution for long term data archival.
- The Data Curation Plan will need to set out:
    - Data classification guidelines
    - Procedures on expiring data
    - Labelling of Intellectual Property
    - Labelling of Confidential/Privacy data
    - Labelling of Public/Private use data
    - Labelling of data with MTA/contractual obligations
    - Data/metadata labelling standards
    - Software compatibility standards
    - Define levels of data descriptions
        - Gold standard – XML files with clearly defined fields and descriptions, following international data labelling standards
            - Alternatively use the Dublin Core Metadata Element Set (ISO Standard 15836)
        - Better than nothing – A brief unstructured text defining the data, the source, and anything else that would help future data mining efforts
        - Below Satisfactory – No descriptions. Data is deposited with only basic filenames, file types, and data creation metadata attached
    - Data retention guidelines
    - Maximum data corruption limit rates
        - Number of allowable error bits per terabyte per year
    - Minimum Data retrieval speed/cost guidelines (i.e. will it take 1 year to retrieve needed data, or $100,000)
    - Obligations to granting agency data
    - Obligations of Institution to scientist on data storage (linked to original employment contracts)
    - With those points in mind – the question remains on what type of hardware, budget, and manpower will be needed to sustain the data archives.
    - In the near term, a potential solution would be to utilize obsolete storage servers for archival, and to migrate the data off these systems as more reliable hardware become available. This is risky, as the obsolete systems have an unknown projected lifetime and there is no manufacturer support to maintain them.

Data Classification Standards

Properly archiving data for the long term will require formalized procedures to classify data and populate the metadata with enough meaningful descriptors such that the data remains useful for future investigators. Procedures for this at the Institute does not currently exist. Will investigate current data classification standards in ISO, government/military entities, pharmaceutical and intellectual property examples.

Part of the need for Data Classification Standards also stems on the need to classify data so that it is treated properly throughout the data lifecycle. Based on the classification,

- Does the data need to be sterilized before releasing into the wild/ can the data be released to the public?
- Are there PHIPA considerations
- Is there intellectual property that must be reviewed beforehand
- Undetermined how data classification procedures will be executed. Will there be an ongoing reviewer/auditor who will ensure that the labs follow these standards? Will there be training/consultation services available?
- A working group committee will need to be created
- May require someone in the project that can review and redact the data before release to the public?
- What happens with data created by people who are no longer employed at the Institute

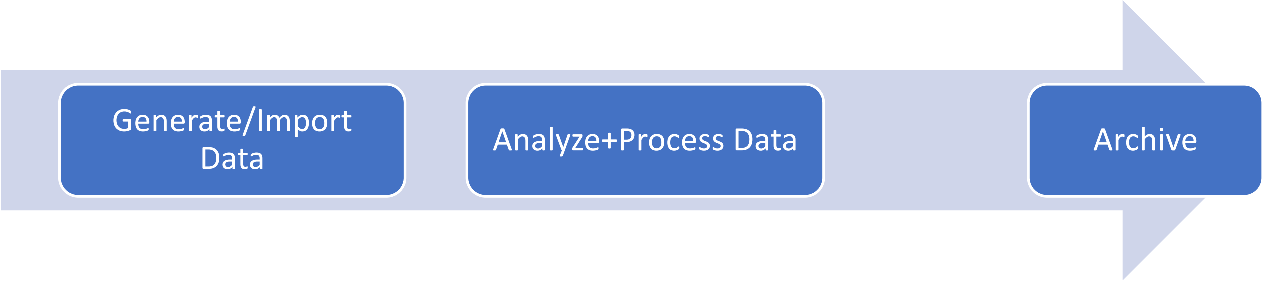
The Funding

The Institute has been fortunate to have been able to secure limited funding that allowed it to start on the path on building a data management infrastructure. However, there are major issues on funding stability that puts significant risks on data management sustainability. Long term funding for data management will greatly stabilize this risk uncertainty.

Principal Investigator

The Principal Investigator will be defining the high level data life cycle and data flows specific for their lab and projects.

Data life cycle



Generally consists of:

Generate/Import Data

The **beginning** of the data life cycle entails these key points:

- identifying the sources and nature of the data
- obtain MTA and define contractual obligations
- delegating and communicating the responsibilities
  - define the **chain of custody of the data**
  - important to include ownership and responsibilities of the data at the end of the project
- transmitting and storing the data
  - identify levels/granularity of backup needed
  - identify levels of compression/noise, allowable compression losses
- define local physical storage/remote storage needs
- authorizing/deauthorizing security controls
- allocating appropriate lab/institutional resources in relation to the data
- initial classification of the data
  - if necessary – enforce encryption of data based on classification level
- estimate the data growth rate and how much data storage space will be needed
- define the end date (if any) for the data
- define data integrity specifications, which may involve one or more of the following points:
  - auditing requirements
  - backup requirements
  - data resolution requirements
  - maximum allowable error/noise
  - cost estimates for retrieving and storing this data

<u>Analyze and Process Data</u>

<u>Mid-point</u> – During the data analysis process, attention will be given to:

- ensure curation of the data and analysis results
    - proper documentation of the data and it's relation to the project
- confirm that metadata is updated as necessary
- when necessary, revisit with lab members on data chain of custody responsibilities/security controls/data classification
- document any changes in the data chain of custody
- confirm that the data growth rate is in-line with early estimates
- confirm that appropriate lock-down of data is undertaken
- confirm that appropriate deletion of data is undertaken
- confirm that backups are being performed as necessary depending on the data classification
- confirm that temporary data has been cleaned off equipment and production servers
- confirm that confidential/patient data has been handled properly
- confirm that MTA and contractual obligations are being adhered to

<u>End Point</u>

- confirm to Expire/Delete any data that was defined at the beginning of the data lifecycle that is appropriate for deletion
- Ensure lab member applies prescribed treatment of the data based on data classification assignment
    - Update/remove security access
- Ensure lab member assigns data with the proper descriptions and metadata - properly attached and visible
- Ensure that lab member begins archival or preparations for public access
    - Storing the data
    - Length of time required
        - If extreme long term – identify what resources, technologies, skill sets and funding are continually needed to migrate the data off obsolete systems
            - Tape technologies are a potential candidate, however, as an example, LTO tape drives are only read compatible to two prior generations, which could potentially mean that in 6 years, no tape drives will be available for sale that can read the archive tapes, unless we purchase it from used markets.
        - Does storage equipment need to be purchased in duplicate? How many backups of the archives are we required to maintain?
    - Cleaning the data
        - Ensure any private, confidential data has been redacted appropriately
    - Ensure that intellectual property has been secured before being publicly broadcasted
    - Ensure that released data is not currently bound by existing Material Transfer Agreements or other contractual obligations

- How accessible should these archives be made available?
    - Personnel
    - Space
    - Security and Authorization procedures
    - Cost estimates

- Ensure appropriate audit logs are retained as necessary to satisfy Tri-Agency Financial requirements
- Ensure that data will be following SSHRC **Research Data Archiving Policy (1990)** requirements
- Lab book, paper files, and all physical medias must be prepared according to accepted academic standard archival procedures. This may need to be adjusted accordingly, depending on the type of media and storage requirements

Where appropriate to the type of data, the Principal Investigator will be requiring those on the data chain of custody to be knowledgeable and cognizant of the relevant Guiding Policies, Principles and Legal Requirements.

A template Data Management Plan will be provided to the Principal Investigator to use as a guideline to include in forthcoming grant applications.

Lab Member

Operating under the enclosing Data Management Strategy framework, and the project direction and requirements from their Principal Investigator, the lab member will apply the appropriate treatment to the data that they are responsible for.

Throughout the period that the Lab Member is assigned to the data during their chain-of-custody, they will be required to continually satisfy these data management requirements.

- Execute requirements as mandated by the framework and Principal Investigator.
- Communicate changes and deficiencies to the Principal Investigator in a timely manner.
- Document data and relevant contexts.
- Store the data onto appropriate systems, and remove data from systems as appropriate in a timely manner
- Classify the data appropriately.
- Update metadata as appropriate to the project.
- Ensure that data in their chain of custody is being handled with the appropriate security levels, conforming to material transfer agreements, and contractual obligations.
- At the end of the project, prepare the data for archival and inform Principal Investigator of any issues that may have arisen.

Summary Points

The fundamental infrastructure and skill sets for data management are in place.
Items that require more ongoing attention at the Institutional and possibly agency levels are:

- Dissemination and training of DMP to all the relevant stakeholders
- Clear long term archival resources and procedures
- Data Classification training and resources will be needed in order to implement a viable Data Archival Strategy
- Consistent and predictable funding

Acknowledgements

- Large sections are sourced from the "Tri-Agency Research Data Management Policy" website, and sample TAHSNr member sample drafts
- Many thanks in particular to Dr. Sue McGlashan, the University of Toronto, and the Sunnybrook Research Institute, for their initial drafts