DIRISA Storage Usage Policy

Version Number	1.0
Date	05 April 2022

Note: NICIS reserves the right to amend and/or extend this document with appropriate notice by way of the DIRISA storage mailing list. NICIS reserves the right to make final determination on the amendment. Any user who does not accept the amendments should voluntarily withdraw as a user. Any user who wishes to withdraw as a user should send an email to dirisa@csir.co.za.

Glossary

CHPC	Centre for High Performance Computing
DIRISA	Data Intensive Research Initiative of South Africa
GUI	Graphical User Interface
iRODS	Integrate Rule-Oriented Data System
Large Project User	A research group consisting of multiple users, requiring between 10 TB and 100 TB storage, typically a CHPC research programme, accessing the storage via the CHPC cluster.
NICIS	National Integrated Cyber-Infrastructure System
POPIA	Protection of Personal Information Act
PI	Principal Investigator leading a research group
RAID	Redundant Array of Independent Disks
SANReN	South African National Research Network
Standard user	Regular user of the storage, typically accessing it via a web based service with access to default 100 GB storage volume
Medium user	A standard user allocated more than the default 100 GB storage for Standard Users but less than 10 TB for Large Storage users, by special application
TB	Terabyte = 10^{12} bytes

1. Introduction

1.1. Purpose

The purpose of this policy is to specify the acceptable usage of the Data Intensive Research Initiative of South Africa (DIRISA) data storage system. The policy outlines the characteristics and requirements for users to apply for access to storage within the DIRISA data storage system. The policy also outlines the responsibilities that the user assumes by using the storage system.

The DIRISA data storage uses the Integrated Rule-Orientated Data System (iRODS) to provide a medium to long term data storage solution for research data. Research programmes such as those hosted by the Centre for High Performance Computing (CHPC) with requirements for storing large data sets. This may be for safe-keeping, re-use or publishing purposes.

Users who have such requirements to store large data sets and do not have the means to do so can apply for this storage with DIRISA. Access to the DIRISA Storage will be limited to research programs that are able to demonstrate not only compelling reasons for requiring storage, but also the ability and intent to manage to storage appropriately.

DIRISA subscribes to the FAIR (Findable, Accessible, Interoperable, Re-usable) and Open data principles. Individual researchers may hence also apply to use DIRISA Storage resources if their data can be shared and has value for further research.

1.2. Scope

- This policy covers all digital data assets that are transferred and stored within DIRISA's data storage system.
- This policy applies to all NICIS users who use the DIRISA data storage system to store digital data assets.

1.3. Objectives

- To establish a firm set of principles for managing the use of the storage system.
- To describe how NICIS and other users can use the DIRISA data storage.
- To ensure that users are aware of their responsibilities when storing their data at the DIRISA data store.

1.4. Summary

Three categories of users are distinguished, namely, “Standard Users”, “Medium Users” and “Large Project Users”

1.4.1. Standard Users

These are users that would be allocated 100 GB of data storage. These users would subscribe through the DIRISA Subscription Tool (SST). The user would have access to the iRODS system using a web interface. The provision of this service and storage is subject to the Acceptable Use conditions presented further herein. For CHPC users, more storage up to a maximum of 10 TB can be available on special application. In this case, the user would be regarded as a special application or medium user.

1.4.2. Medium Users

Medium users are anticipated to be similar to standard users, with the exception of requiring more than the default 100 GB, either through data growth over time or as a result of an unusually large initial storage requirement. Medium users will need to make a special application to DIRISA in order to obtain the additional storage.

1.4.3. Large Project Users

These are CHPC users that would be requiring more than 10 TB of data storage, up to a maximum of 100 TB. Large Project users are research groups with multiple members, represented by a lead scientist. A typical example would be a Research Programme hosted by the CHPC, under the leadership of a Principal Investigator approved by the CHPC, transferring data to and from the CHPC cluster with either command line tools or a web-based data deposit tool.

Exceptional cases for non-CHPC users requiring more than 100 TB of DIRISA/iRODS storage can be considered. In this instance, discussions will be conducted with the user to determine the feasibility and benefit of providing more than 100 TB of storage.

2. Technical Information

The Integrated Rule Orientated Data System (iRODS) is an open-source data management software system that is used to store and manage data. Data stored on the iRODS system can be accessed via command line or a GUI:

- iCommands is a collection of UNIX based utilities which allow users to interact with data on iRODS using a command line interface.
- Metalnx is a web application designed to work alongside the iRODS. It provides a GUI that can help simplify most administration, collection and metadata management tasks removing the need to memorize the long list of iCommands. An iRODS client stores files as data objects and categorizes them into collections and subcollections:
- A Data Object may refer to multiple Replicas. Replicas are exact copies of a file, located in multiple physical locations.
- Collections make no reference to the physical storage path. It is possible for two Data Objects in a Collection to be stored in different physical locations.
- A guide is provided for using the MetalNX interface to manage data.

Data Objects and Collections are stored in Storage Resources in an iRODS Zone. Each Storage Resource has a name (the Resource's logical representation), hostname and path (the physical representation of the Resource, where files are kept). The hostname is the network name of the device that serves the data, and the path is the local file system path or object storage bucket that holds the data.

The DIRISA storage system has geographic redundancy in the sense of being duplicated between Cape Town and Pretoria, hence a level of safeguarding is provided against disasters such as fires or floods.

3. Envisaged Use

This is a likely, but not definitive list of envisaged use cases:

3.1. Data sharing for co-operative projects

Research programs involved in co-operative research projects using other computing centres may apply to use the iRODS storage for exchange and safekeeping of data sets. This differs from the situation described in paragraph 3.3 by virtue of access to the data being controlled rather than open. The same limits as 3.2 apply: maximum of 100 TB total over 3 years.

3.2. Storage of large data sets for re-use at the CHPC

CHPC PIs may apply to store large (greater than 10TB and less than 100TB) input and output data sets on the iRODS storage, for longer than 1 year but less than 3 years, in order to avoid having to regenerate or re-download these data sets.

The Data Management Plan (DMP) will need to describe the method and purpose of the intended re-use. The South African Data Management Planning tool (DMP-SA, at https://secure.dirisa.ac.za/SADMPTool/) can be used to develop a DMP, but care must be taken to supply the information required in paragraph 4.3.1.

Data will only be accessible from the CHPC cluster and will have to be restaged to the CHPC's storage for use. Data movement will be by means of the iCommands iRODS command line interface.

3.3. Open Data and Data publishing

It is a standard practice that data underpinning peer-reviewed publications must also be freely published in order to permit other researchers to duplicate, check and build upon the published research.

DIRISA promotes the adoption of FAIR data practices i.e., the principle of "as open as possible, and as closed as necessary"). Recognising possible embargoes on some datasets, users are encouraged to share data where feasible. NICIS users may apply to upload data to the iRODS storage for this purpose, with a publicly accessible interface for downloading to be provided by DIRISA.

4. Access Application Process

4.1. Standard User

Standard users need to navigate to the DMP-SA tool at https://secure.dirisa.ac.za/SADMPTool/ to create a DMP using the Standard Data User template. The DMP would need to be downloaded and uploaded on the Data Deposit Tool (DDT Link) upon registration. Once application is completed and received, an administrator will process the user's request, and the applicant will be notified of the outcome.

4.2. Medium User

If a standard user needs extra storage above the provided 100 GB, they can complete an online application form (link to be provided). The user's request will be assessed and will be notified of the outcome. The user will need to submit a data retirement plan, detailing the anticipated lifetime of the data and details of what will happen to the data at the end of the lifetime. Failure to comply with the data retirement plan may result in the data being deleted.

4.3. Large Project User

Large users need to submit an online application form (link to be provided) containing the requirements described in paragraph 4.3.1. Requests will be evaluated and the applicant will be notified of the outcome.

4.3.1. Requirements

4.3.1.1. Motivation

The application must contain a motivation supplying the reasons for the application and technical requirements for the requested additional storage.

4.3.1.2. Data storage manager

Each research program with access to the DIRISA storage must designate an individual member as "Data Storage Manager" as well as a secondary Data Storage Manager. This designated person is responsible for uploading, downloading, monitoring and removal of data.

4.3.1.3. Data management plan

The principal investigator and data storage manager must prepare and submit a detailed data management plan containing at least the following:

- Quantity of storage required
- Data lifetime (up to 3 years)
- Plan describing what will be done with the data at the end of the data lifetime
- Access and security requirements
- Anticipated growth rate (up to a total maximum of 100TB)

- Description of a metadata system that will ensure that the members of the research program at all times know what they have, how much of it they have, where it is to be found, when it will be retired and who is responsible for a given set of data

4.3.1.4. Continuity plan

The application shall specify secondary designated persons who can take on the responsibilities of the principal investigator and the data storage manager if they are not available. These substitutes will be deemed to have the authority to make decisions on behalf of the research program. Both the primary and secondary designated persons shall be subscribed to the DIRISA storage mailing list, which will be used to inform them of any events, changes or developments affecting the users.

4.3.1.5. Security and access

The level of security and access to the data set will be under control of the users, who need to take note of the following anticipated access modes and set access accordingly:

- Upload access for members of the Research Programme only, download access for everybody
- Upload access for members of the Research Programme only, download access to specified credentials only
- Upload and download access for members with specified credentials only

4.3.1.6. Compliance with privacy legislation

The lead scientist or CHPC Principal Investigator shall take responsibility to ensure compliance with the POPI Act and provide a written confirmation that either (a) no personal information will be included in their data; or (b) that the programme is fully compliant with POPIA.

In the latter case the PI must provide their privacy policy and proof that consent was granted by each and every individual to the PI to collect, store and process their personal information, and the PI must provide full indemnity to the CSIR, as the third party “processor” as defined in the POPI Act, for any misuse or loss or leak of the personal information.

4.3.1.7. Monitoring

The data storage manager shall take responsibility to monitor the following properties of all data sets belonging to the research program:

- Age of data
- Use of data
- Size of data set
- Relevance of data set (is it still in use, or anticipated to be used?)

The data storage manager must be able to supply this information on request. Failure to do so may result in data being removed without the consent of the research program.

4.3.1.8. End of Interim Storage Period

The PI must provide, as part of the data management plan, a detailed plan for the end of data storage period. At the end of the storage period (maximum of 3 years) the data manager must arrange for the implementation of this plan and provide confirmation that this was done within 30 days. After that all remaining data may be permanently deleted.

4.3.1.9. Evaluation Process

The motivation for access to the storage must be clear on the needs of the programme, the lack of access to storage elsewhere, the importance of the project, etc.

NICIS is only able to evaluate eligibly based on reasonable requirements and genuine need. The historical record of data use at NICIS will be taken into consideration, along with past feedback provided by the PI.

The data management plan is essential and will be scrutinised for practicality and effectiveness. It must be reasonable to accommodate within the constraints of the DIRISA facilities.

4.3.1.10. Agreement

In the case of a successful application, the PI will enter into and sign a standard agreement with NICIS which confirms that the following requirements have been specified to the satisfaction of NICIS:

- Motivation
- Purpose
- Type of access
- Data management plan
- Nomination of primary and secondary designated individuals
- Amount of storage
- Data lifetime
- Data retirement plan

5. Special Circumstances

At the NICIS manager’s discretion, in case of emergency, special circumstances, or where deemed necessary to fulfil the mandate of the NICIS, exceptions may be made to the above conditions.

Priority projects: NICIS is mandated to support high priority and time-critical projects for weather and pandemic forecasting which need to be prioritised for usage of resources from time to time. In order to achieve timeous results for national decision-making, the removal of some data may be required at short notice when storage usage is high.