EPJ Web of Conferences **186**, 12002 (2018) https://doi.org/10.1051/epjconf/201818612002
LISA VIII

Progress with implementing the new research data management system at HartRAO

Glenda Coetzer[1,2,⋆], *Roelf* Botha[1], and *Lorette* Jacobs[2]

[1] *Hartebeesthoek Radio Astronomy Observatory, South Africa*
[2] *University of South Africa, South Africa*

Abstract. The Hartebeesthoek Radio Astronomy Observatory (HartRAO) participates in global radio astronomy and fundamental astronomy (space geodesy) research activities. Data and data products produced by HartRAO's expanding range of on-site and off-site instrumentation must be archived and stored at HartRAO and made accessible to the scientific community. The data management and storage systems currently being used for managing fundamental astronomy data are not capable of handling the large volumes of data and have become obsolete. This necessitated the design and implementation of a next-generation Geodetic Research Data Management System (GRDMS), which complies with internationally accepted data service standards. We present the top-level conceptual model of the GRDMS and progress to date with developments of various sub-systems, data structuring and organisation within the sub-systems.

1 Introduction

The Hartebeesthoek Radio Astronomy Observatory (HartRAO) participates in radio astronomy and fundamental astronomy (space geodesy) research activities. Radio astronomy activities focus predominantly on the use of HartRAO's 26 m radio telescope to conduct radio astronomy observations in single-dish mode and Very Long Baseline Interferometry (VLBI) observations in conjunction with other radio telescopes across the globe. HartRAO's fundamental astronomy activities include the use of HartRAO's 15-m and 26-m radio telescopes, in conjunction with other radio telescopes globally, to conduct astrometric and geodetic VLBI observations. Collocated on the HartRAO site are various Global Navigation Satellite Systems (GNSS) reference stations, two Satellite Laser Ranging (SLR) systems, a Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS) system, various meteorological instrumentations as well as a gravimetric and seismic instrument. In the near future, a Lunar Laser Ranging (LLR) system and VLBI Global Observing System (VGOS) radio telescope will become operational at HartRAO. We are also expanding our network of instruments with the installation of GNSS, seismic and meteorological instruments across Southern Africa. Together with HartRAO's participation in the African VLBI Network (AVN) these will expand HartRAO's fundamental astronomy data collection and research activities.

HartRAO is the only operational facility of its kind in Africa and provides a vital link to the global geodetic network of data providers [1]. HartRAO is responsible for the generation, correlation, storage

⋆e-mail: glenda@hartrao.ac.za ORCID: 0000-0002-0788-5660

and dissemination of technique-specific data and data products to international service providers, such as the Crustal Dynamics Data Information System (CDDIS), University NAVSTAR Consortium (UNAVCO), International Laser Ranging Service (ILRS) and International VLBI Service for Geodesy and Astrometry (IVS), to name but a few. Some data and data products, such as Global Positioning System (GPS) and single-dish observation data, are stored on-site. Other data, for example VLBI data, are shipped to international correlators and data centres [2].

HartRAO's current fundamental astronomy data management and storage system is not capable of managing the predicted additional large volumes and complexity of data resulting from the expansion. The current system has many drawbacks - it is outdated, segmented and distributed; it has limited functionality and capacity to handle different data types and different user requirements; different approaches are followed for handling datasets; data are stored on different servers; access to the data holdings are via ftp or http, which have many broken links and are not frequently updated. This necessitated the design and implementation of a next-generation Geodetic Research Data Management System (GRDMS) that will be able to cater for all of HartRAO's fundamental astronomy data, whilst complying with internationally accepted data service standards.

2 Overview of the GRDMS

The main objectives of the new system is to organise, structure, archive and store geodesy and geodynamics related data and data products in a central data bank; disseminate data and data products to global data service providers and the growing research community; maintain metadata and provide metrics for reporting purposes. Professionals from different disciplines - academic/scientific, information technology and communications (ITC) and library and information science (LIS) - are working together to design and implement a new GRDMS [1]. To meet the set objectives and not to 'reinvent the wheel', it was decided that the new data management system will feature components (see Figure 1) similar to that of the CDDIS and UNAVCO [3, 4]. The same data structures and file-naming convention will be used for all datasets. Each dataset will receive persistent interoperable identifiers. User access will be facilitated by an interactive web-based graphic user interface (GUI). The complete system will be divided into functional units, which can operate relatively independently of each other - this helps to avoid critical overall failures and also enables faster repairs of broken / damaged components. Security is also a major concern and the systems have been designed so that, from the public domain, only limited, read-only or no direct access is possible to mission-critical systems and the main data archives.

Redundancies on all functional units and processes are required to prevent data loss and system downtime, therefore a clone of the complete GRDMS will be housed at a different location. It will be transparent to the stations as well as to the end-user which data centre is being used.

2.1 Functional overview of the GRDMS

The overall functional concept of data collection and distribution is depicted by Figure 1. The main components are the functional units as follows:

- Virtual Private Networking is used to connect all stations and instruments to the Data Collection Virtual Machine (VM), Vector, which handles the data collection station monitoring.
- The Data Tools VM then pulls the raw data from the Data Collection system, Quality Check (QC) it and stores it in the Archive, in the correct folder and file format and naming convention.

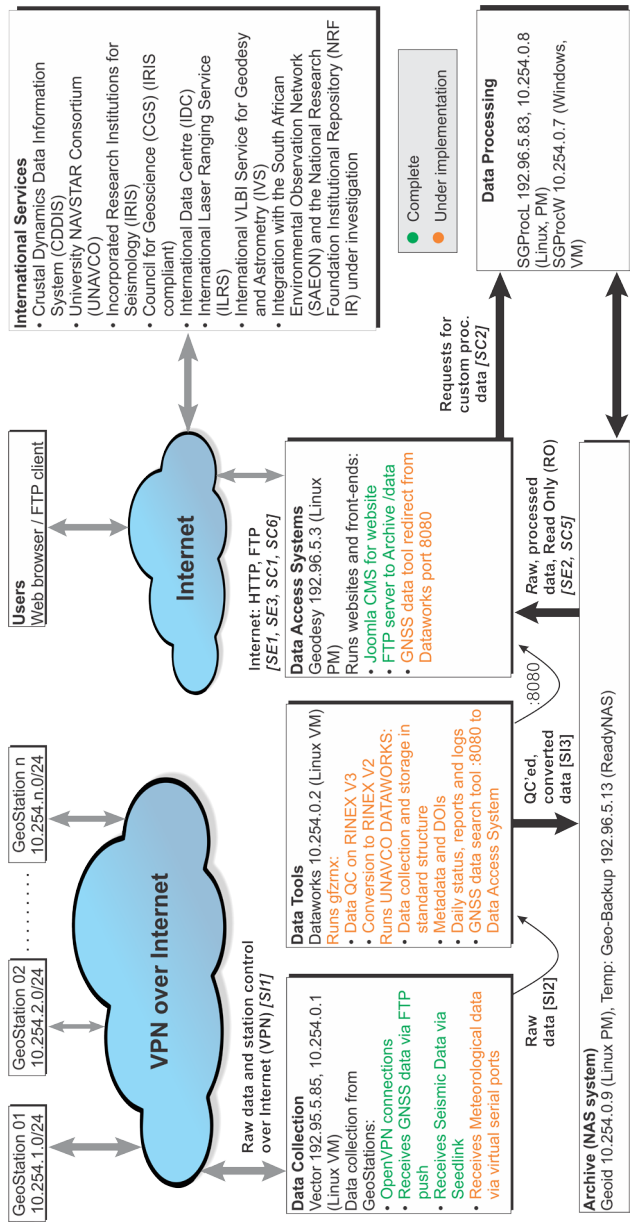

Figure 1. The GRDMS functional diagram (square brackets denote steps in the data cycle, see Section 2.2).

- The Data Access System provides various tools (websites, FTP access etc.) to let users gain access to the raw data, as well as request custom processed data products. This system also streams the data to the international partners and service providers.
- The Data Processing Systems consists of a Linux Physical Machine (PM) and a Windows VM. These have read-only access to the raw data in the Archive and can also store data products to the

archive. The processing can occur by requests via the Data Access System or from a user logged into one of the systems.

- The Archive is a storage cluster that can be expanded dynamically, it also has an off-site fail-over duplication.

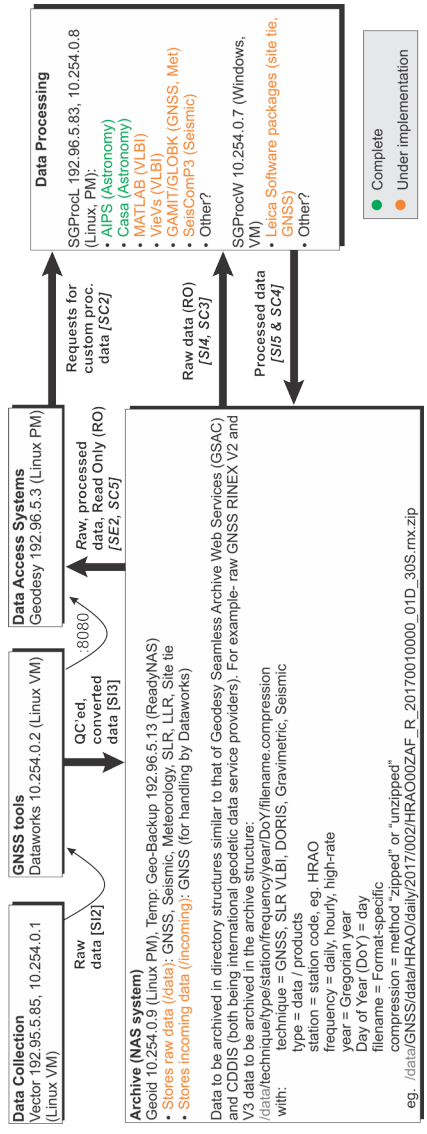

Figure 2. The GRDMS Archive internal data structure and the Data Processing software packages (square brackets denote steps in the data cycle, see Section 2.2).

The internal structure of the Archive and software on the Data Processing systems is depicted in Figure 2.

2.2 Logical processes of the GRDMS

The proposed logical processes are divided in three types: automated Internal Steps [SI], Steps for External requests [SE] and Steps for Custom data requests [SC] - these are depicted in Figures 1 and 2 as [SI], [SE] and [SC] and will be used in the filenames and procedures executing these steps.

Internal Steps [SI] data cycle:

- SI1: The Data Collection Unit controls different instruments at different stations and receives raw data from the stations.
- SI2: Raw data is collected by the Data Tools Unit to quality check it and then pre-process it into required file types, formats and filing structures (eg. the Rinex V3 data are quality checked and converted to Rinex V2 data).
- SI3: Checked and pre-processed data are sent and stored in specified formats and structures in the Archive Unit.
- SI4: Checked and pre-processed data are collected by the Data Processing Unit to process it to predetermined basic data products.
- SI5: The basic data products data are sent and stored in specified formats and structures in the Archive Unit.

External Steps [SE] data cycle:

- SE1: The scientific community can interact with the Data Access System and request data via a website (HTTP) and/or FTP.
- SE2: The Data Access System obtains the requested data from the Data Archive Unit.
- SE3: Requested data is packaged into a single compressed file and made available.

Custom Steps [SC] data cycles (for requests which are simple enough to be handled automatically by the system):

- SC1: Requests are submitted on the website via a special interface.
- SC2: Requests are translated to a script and pulled by the relevant Data Processing Unit.
- SC3: The Data Processing Unit obtains the required raw data and processes it.
- SC4: Processed data is sent to the Archive Unit for storage.
- SC5: The Data Access System retrieves processed data.
- SC6: Results are sent to the requesting user.

2.3 Data structure within the GRDMS

The GRDMS data structures will be similar to those of the CDDIS. An example of the structure of raw technique-specific Rinex data is explained in Table 1.

At least the “8.3.Z” file-naming convention will be used for all GNSS data for example, see Table 2.

Table 1. An example of the structure of raw technique-specific Rinex data in the format: /data/technique/type/station/frequency/year/DoY/filename.compression

Descriptor	Description	Example value(s)
technique	Technique abbreviation	gnss, slr, vlbi, doris, gravity, seismic
type	Data Type	Rinex
station	Technique specific station Code	HRAO, MATJ
frequency	File Frequency	daily, hourly, high-rate
year	Gregorian year	2017
DoY	Day of Year	028
filename	Technique-specific	see Table 2 for example
compression	Compressed? and type	.Z, .gz, .zip

Table 2. An example of the structure of a GNSS Rinex2 filename "8.3.Z" eg.SSSSDDD0#.YY.Z

Descriptor	Description	Example value(s)
SSSS	Site code	HRAO, MATJ
DDD	Day of Year (DOY)	028
0	Sequence number	0
#	session code	a
YY	two-digit year	17
Z	compressed?	.Z

3 Progress with the design and implementation of the GRDMS

The sections marked in green within Figures 1 and 2 are complete, the sections in Orange within these figures are under implementation.

The following processes and software implemented in different sub-systems have been completed:

- within the Data Collection unit, open Virtual Private Network (VPN) connections were established
- the system receives GNSS data via ftp push and receives seismic data via Seedlink
- Joomla Content Management System (CMS) for websites was installed on the Data Access System unit and ftp server to archive data
- Astronomical Image Processing System (AIPS) and Casa software were installed on the Data Processing unit

Functionality still to be implemented are:

- the receipt of meteorological data via virtual serial ports
- quality checks on Rinex Version 3 data and the conversion of Rinex Version 3 data to Version 2 data
- the installation and running of UNAVCO Dataworks software to manage the collection and storage of data in standard predefined structures
- the collection of metadata and assigning DOIs to datasets
- daily status reports and logs and the installation of GNSS data search tools to redirect data from the Dataworks port
- the installation of MATLAB, VieVs, GAMIT/GLOBK, SeisComP3 and Leica software packages

The new seismic network has a specially designed data processing and storage systems, which arrived at HartRAO during 2016. These systems are currently being integrated into the GRDMS. HartRAO has also obtained a Dell Westmere processing cluster, containing of 1152 CPU cores (14.8 Teraflops) and 3456 GB RAM. This cluster will form an additional functional section of the Data Processing system, focused towards processing of VLBI data. Implementation is expected to occur during 2018.

4 Progress with the design and implementation of the GRDMS

Data and data products resulting from HartRAO's expanded range of on-site and off-site instruments must be archived and stored at HartRAO and made accessible to the scientific community. We have highlighted the drawbacks of the current data management and storage system, which served as a motivation for the design and implementation of a new research data management system, currently under implementation. We presented progress to date on various sub-systems (units) as well as progress on establishing data structure and organisation within the GRDMS sub-systems.

Acknowledgement

The authors would like to acknowledge funding awarded by the National Equipment Programme (NEP) of the National Research Foundation (NRF) for funding the development of the co-located academic network.

References

[1] G.C. Coetzer, R.C. Botha, L. Combrinck et al., *A New Geodetic Research Data Management System at the Hartebeesthoek Radio Astronomy Observatory* in *Open Science at the Frontiers of Librarianship*, edited by András Holl, Soizick Lesteven, Dianne Dietrich, and Antonella Gasperini (2015), Vol. 492 of Astronomical Society of the Pacific Conference Series, p. 22–30

[2] Behrend, D., Data Science Journal **12**, WDS81-WDS84 (2013)

[3] Noll, C., CDDIS. IVS2015-2016 Biennial Report (2017)

[4] UNAVCO, UNAVCO GSAC WS: Web Services for Geodesy Data Repositories (2013year)