Designing a Geodetic Research Data Management System for the Hartebeesthoek Radio Astronomy Observatory

By

GLENDA LORRAINE COETZER

Submitted in accordance with the requirements for the degree of

MASTER'S OF INFORMATION SCIENCE

at the

UNIVERSITY OF SOUTH AFRICA

SUPERVISOR: DR L JACOBS

NOVEMBER 2018

DECLARATION OF ORIGINALITY

Name: Glenda Lorraine Coetzer

Student number: Student number: 4023-843-1

Degree: MASTER'S OF INFORMATION SCIENCE

DESIGNING A GEODETIC RESEARCH DATA MANAGEMENT SYSTEM FOR THE HARTEBEESTHOEK RADIO ASTRONOMY OBSERVATORY

I, declare that the above dissertation is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references. I further declare that I submitted the thesis/dissertation to originality checking software and that it falls within the accepted requirements for originality. I further declare that I have not previously submitted this work, or part of it, for examination at UNISA for another qualification or at any other higher education institution.

Signature ……………………………………………

Date: …………………………………………………

ABSTRACT

The radio astronomy and space geodesy scientific instrumentation of the Hartebeesthoek Radio Astronomy Observatory (HartRAO) in Gauteng, South Africa, generate large volumes of data. Additional large data volumes will be generated by new geodesy instruments that are currently under construction and implementation, including a lunar laser ranging (LLR) system, seismic and meteorological systems and a Very Long Baseline Interferometry (VLBI) global observing system (VGOS) radio telescope. The existing HartRAO data management and storage system is outdated, incompatible and has limited storage capacity. This necessitates the design of a new geodetic research data management system (GRDMS). The focus of this dissertation is on providing a contextual framework for the design of the new system, including criteria, characteristics, components, an infrastructure architectural model and data structuring and organisation. An exploratory research methodology and qualitative research techniques were applied. Results attained from interviews conducted and literature consulted indicates a gap in the literature regarding the design of a data management system, specifically for geodetic data generated by HartRAO instrumentation. This necessitates the development of a conceptual framework for the design of a new GRDMS. Results are in alignment with the achievement of the research questions and objectives set for this study.

ISIFINGQO

Umkhakha we-Radio Astronomy kanye ne-Space Geodesy wase-Hartebeeshoek Radio Astronomy Observatory (HartRAO) Gauteng, South Africa, ukhiqiza ulwazi oluningi (data). Olunye ulwazi oluningi luzokhiqizwa imishini emisha yokucwaninga ye-geodesy efana ne-Lunar Laser Ranging (LLR) system, imishini yokucwaninga ukuzamazama komhlaba (seismic and meteorological systems), isipopolo se-radio esisha se-Very Long Baseline Interferometry (VLBI) GlobalObserving System (VGOS), kanye neminye imishini eyakhiwayo e-HartRAO njengamanje. Indela yokugcina nokunakekela ulwazi e-HartRAO isindala futhi ayisakulungele ukumelana nolwazi olukhizwayo njengamanje, futhi ayisakwazi ukumelana nenani lolwazi. Lokhu kuyisizathu sokuthi kwakhiwe indlela entsha yokunakekela ulwazi, phecelezi i-Geodetic Research Data Management System (GRDMS). Umongo walesisifundo ukwakha umgogodla (framework) wendlela okuzokwakhiwa ngayo lendlela yokunakekela ulwazi entsha. Lokho kuhlanganisa i-criteria, ama-characteristics, amalunga (components), i-model ye-infrastructure kanye ne-data structuring kanye ne-organisation. Lesisifundo sisebenzise inhlobo yokucwaninga ebizwa nge-exploratory research kanye nezindlela zokucwaninga ezi-qualitative. Imiphumela etholwe ngokubuza abantu kanye nokuhlaziya imibhalo mayelana nalolucwaninga ihambisana nemibuzo kanye nezinhloso zalolucwaninga.

OPSOMMING

Die radio-astronomie en ruimtegeodesie wetenskaplike instrumentasie van die Hartebeesthoek Radio-Astronomie Observatorium (HartRAO) in Gauteng, Suid-Afrika, produseer groot hoeveelhede data. Die nuwe geodesie-instrumente wat tans gebou en geïmplementeer word – byvoorbeeld ‘n maan laser reikaftstandbepaler, seismiese- en meteorologiese stelsels en ‘n baie lang basislyn interferometrie globale waarnemingstelsel (VGOS) radioteleskoop – gaan verhoogde datavolumes genereer. HartRAO se huidige databestuur- en bergingstelsel is verouderd, onversoenbaar en beskik oor beperkte bergingsvermoë. Dit noodsaak die ontwerp van ‘n nuwe geodetiese navorsingsdatabestuurstelsel~~.~~ Hierdie verhandeling fokus op die lewering van ‘n kontekstuele raamwerk vir die ontwerp van die nuwe stelsel, insluitende maatstawwe, karaktereienskappe, komponente, ‘n infrastruktuur argitektoniese model, sowel as datastrukturering en -organisasie. ‘n Ondersoekende navorsingsmetodologie en kwalitatiewe navorsingstegnieke is gevolg. Resultate van die onderhoude wat gevoer is en die literatuur wat geraadpleeg is, dui daarop dat daar 'n leemte bestaan aangaande in die ontwerp van ‘n navorsingsdatabestuurstelsel, spesifiek vir geodetiese data wat deur HartRAO se instrumentasie gegenereer word, bestaan. Dit noodsaak die ontwikkeling van / dat 'n nuwe kontekstuele raamwerk vir die ontwerp van die nuwe GRDMS tot stand gebring word. Die resultate spreek die navorsingsvrae aan en volbring die doelwitte van hierdie studie.

KEY TERMS

Conceptual/Architectural model; data life-cycle; data management systems; scientific data management systems; data standardisation; data structuring; digital object identifiers (DOIs); file naming conventions (FNCs); geodetic data; metadata.

ACKNOWLEDGEMENTS

During the course of three years of research, fieldwork, data collection, analysis and discussions, I received valuable help, support and guidance from numerous people.

Special gratitude to my mentor and supervisor, Dr Lorette Jacobs (extraordinary lecturer of the University of South Africa), for unwavering support, guidance and for believing in me.

I am equally grateful to Dr Roelf Botha for his continuous support and guidance.

Appreciation is also extended to the Hartebeesthoek Radio Astronomy Observatory for the financial assistance provided to me to embark upon the study.

I also extend my gratitude and appreciation to Marisa for her support, assistance and motivation throughout my research.

I would like to thank my children, Charné and Divan, for their love, patience and support.

Thank you to my parents, Willem and Breggie, brother Norman and sister-in-law Chantelle for believing in me, and for all their encouragement.

Thank you to Leslie for his advice and support.

Thank you to Tewie, Golda and Genevieve for their support.

Finally, and most importantly, I thank my Lord for giving me the ability and strength to complete my studies.

LIST OF ABBREVIATIONS

ACID	Accuracy, Completeness, Isolation and Durability
AEOMN	African Earth and Ocean Monitoring Network
AIPS	Astronomical Image Processing System
API	Application Programming Interface
ASCII	American Standard Code for Information Interchange
AT-LBA	Australia Telescope – Long Baseline Array
AVM	Astronomy Visualization Metadata
AVN	African VLBI Network
CASA	Common Astronomy Software Applications
CDDIS	Crustal Dynamics Data Information System
CDS	Strasbourg Astronomical Data Centre
CHPC	Centre for High Performance Computing
CPF	Consolidated Prediction Format
CPU	Central Processing Unit
CRIS-INAF	Current Research Information System of the Istituto Nazionale di Astrofisica
CSIR	Council for Scientific and Industrial Research
DMS	Data Management System
DOI	Digital Object Identifier
DORIS	Doppler Orbitography and Radiopositioning Integrated by Satellite
DSO	Data Structure and Organisation
DSS51	Deep Space Station 51
EOSDIS	Earth Observing System Data and Information System
ESO	European Southern Observatory
EVN	European VLBI Network
FITS	Flexible image transport system
FNC	File-Naming Convention
FRD	Foundation for Research Development
FTP	File Transfer Protocol

GB	Gigabyte
GDMS	Geodetic Data Management System
GNSS	Global Navigation Satellite System
GPS	Global Positioning System
GRDMS	Geodetic Research Data Management System
GSAC	Geodesy Seamless Archive Centers
GUI	Graphic User-Interface
HartRAO	Hartebeesthoek Radio Astronomy Observatory
HTTP	HyperText Transfer Protocol
IAU	International Astronomical Union
ICRF	International Celestial Reference Frame
IERS	International Earth Rotation and Reference Systems Service
IGS	International GNSS services
ILRS	International Laser Ranging Service
I/O	Input/Output
IP	Internet Protocol
ISNI	International Standard Name Identifier
ISO	International Organization for Standardization
IT	Information Technology
ITRF	International terrestrial reference frame
IVS	International VLBI Service
JIVE	Joint Institute for VLBI in Europe
JPL	Jet Propulsion Laboratory
KB	Kilobyte
KBps	Kilobyte per second
LiFi	Light Fidelity
LLR	Lunar Laser Ranging
MB	Megabyte
MBps	Megabyte per second
NACSN	National Academic Co-located Seismic Network

NASA	National Aeronautics and Space Administration
UNAVCO GSAC	University NAVSTAR Consortium Geodesy Seamless Archive Centers
NCCS	New Computer Control System
NECSA	South African Nuclear Energy Corporation
NGS	National Geodetic Survey
NRF	National Research Foundation
NRAO	National Radio Astronomy Observatory
NSI	National System of Innovation
ORCID	Open Researcher and Contributor Identification
PWV	Precipitable Water Vapour
RAID	Redundant Array of IndepenDent [arrays]
RDA	Resource Description and Access
RDBMS	Relational DataBase Management System
RINEX	Receiver INdependent EXchange
SGProcL	Space Geodesy Processing Linux
SGProcW	Space Geodesy Processing Windows
SINEX	Solution INdependent EXchange
SKA	Square Kilometer Array
SLR	Satellite Laser Ranging
S/LLR	Satellite/Lunar Laser Ranging
SSD	Solid-State Disks
TB	Terabyte
UNAVCO	University NAVSTAR Consortium
UNISA	University of South Africa
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
URN	Uniform Resource Name
USA	United States of America
USB	Universal Serial Bus

VGOS	VLBI Global Observing System
VLBA	Very Long Baseline Interferometry Array
VLBI	Very Long Baseline Interferometry
VPN	Virtual Private Network
WiFi	Wireless Fidelity

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Introduction and background

The Hartebeesthoek Radio Astronomy Observatory (HartRAO) was established in 1961 by the National Aeronautics and Space Administration (NASA) of the United States of America (USA). The Observatory is situated in the foothills of the Magalies Mountain Range, 50 kilometers north-west of Johannesburg, Gauteng, South Africa (Davis & Coetzer 2010:69; Gaylard & Nicolson 2007:49).

A 26-metre diameter radio dish was constructed at Hartebeesthoek and was used by NASA, in collaboration with South Africa, to obtain data from and send commands to unmanned USA space probes (HartRAO 2001c). The Observatory, which was referred to as *Deep Space Station 51* (DSS51) at the time, was handed over to the South African Council for Scientific and Industrial Research (CSIR) when the USA withdrew its research activities from African soil. Following the handover, the 26-metre radio dish was converted into a fully functional radio astronomy telescope in 1975, which changed the focus of the Observatory and, thereby establishing the first and only radio astronomy observatory in Africa. In 1988, the Observatory became a national facility of the Foundation for Research Development (FRD), which is now known as the National Research Foundation (NRF), an independent government agency of the Republic of South Africa (Gaylard & Nicolson 2007).

During the post-NASA era, the original research and operation of the Observatory was mainly in radio astronomy. However, during the eighties, a new science developed at HartRAO, namely Space Geodesy. Although there are differences between these two subfields of astronomy, there is much synergy, in that they are mutually supportive of each other (Combrinck & Combrink 2004). Space geodesy relates to the measurement and representation of the earth, including its gravitational field, geodynamical phenomena such as crustal motion, tides and polar motion in a three-dimensional time-varying space, as well as the determination of the surfaces and gravity fields of celestial bodies, such as the moon and planets (HartRAO 2012; Torge & Muller 2012).

As an internationally recognised radio astronomy and space geodesy research facility, HartRAO is a vital link in a global network of high-quality scientific data service providers (De Witt, Gaylard, Quick & Combrinck 2013:45; Gaylard & Nicolson 2007:51–52; HartRAO 2001c; Mashaba, Combrinck, Botai, Munghemezulu & Botha 2016). Various co-located radio astronomy and space geodesy instrumentation at HartRAO makes this research facility the only fundamental geodesy station in Africa (Nickola 2012:5) and one of seven fiducial sites worldwide (Combrinck 2014). The station participates with global networks of radio astronomy and space geodesy instruments and service providers, such as the:

- European Very Long Baseline Interferometry (VLBI) Network (EVN);
- Australia Telescope – Long Baseline Array (AT-LBA);
- United States of America Very Long Baseline Array (VLBA);
- Space VLBI with orbiting radio telescope known as *RadioAstron*;
- African Earth and Ocean Monitoring Network (AEOMN);
- International VLBI Service (IVS) for Geodesy and Astrometry;
- International Laser Ranging Service (ILRS);
- International Global Navigation Satellite System (GNSS) Service (IGS);
- International Earth Rotation and Reference Systems Service (IERS); and
- African VLBI Network (AVN).

Data generated by HartRAO’s instrumentation are distributed to international data service providers, where data is correlated, analysed, archived, stored and made available to the scientific community. Examples of these international data service providers include the Joint Institute for VLBI in Europe (JIVE), the Pawsey Supercomputing Centre, the Crustal Dynamics Data Information System (CDDIS), the IVS and the IERS (HartRAO 2012; Coetzer, Botha, Combrinck & Fourie 2015).

Technological and scientific advances in astronomy and space geodesy research have fueled the design, development and application of new instrumentation in support of research activities (Appleby, Bianco, Noll, Pavlis & Pearlman 2016:23). HartRAO’s geodesy instrumentation and technique expansion encompasses the design, development, construction and implementation of new geodetic instruments, which comprise a Lunar Laser Ranging (LLR) system, VLBI Global Observing System (VGOS) telescope, as well as gravimetric, seismic and meteorological instrumentation (Coetzer, Botha, Combrinck & Fourie 2015; Mashaba 2014). HartRAO will offer additional data services and consolidated African geodetic datasets to the international scientific data service providers and the scientific user community (De Witt, Mayer, MacLeod, Combrinck, Petrov & Nickola 2016:118–122). This will contribute to South Africa's pursuit of national and international cutting-edge research, knowledge production, exchange, access to research infrastructure and human capacity development (NRF 2017:105–109). The expansion of research services offered by HartRAO align with the *National Development Plan, 2030* (Republic of South Africa 2011); the National System of Innovation's (NSI) strategic objectives of knowledge generation, exploitation and the creation of a globally relevant knowledge infrastructure (Manzini 2012); and the vision and mission of the NRF (NRF 2017).

Hartebeesthoek Radio Astronomy Observatory generates massive amounts of data on a daily basis, and its ability to manage the dynamic data is of critical importance to its operations. The recently constructed HartRAO VGOS radio telescope will be fully operational within the next few years (Mey 2017:6). Gravimetric, seismic and meteorological instrumentation has been successfully installed at HartRAO, the South African Nuclear Energy Corporation (NECSA) in Gauteng, the Nelson Mandela University in Port Elizabeth in the Eastern Cape and Matjiesfontein in the Western Cape. These geoseismic stations are operational and are streaming geodetic data to HartRAO servers. HartRAO's scientists and engineers envisage the new instruments will generate large volumes of additional technique-specific data after full commissioning. This will, in turn, increase the amount of data requiring proper management.

The current geodetic data management and storage system used for managing HartRAO's geodesy data will not be able to cater and store the additional large volumes of scientific data (Table 1.1). The current system was not designed to manage the additional new techniques-specific data types (Coetzer et al. 2015). The above are only a few of the drawbacks of the current system (a full discussion of the drawbacks are provided in section 1.2). It is imperative for HartRAO to design a new data management system that can cater for all of HartRAO's current and future geodetic data service needs.

Within the context of the foregoing information, the aim of this research is to investigate what has been reported on the topic in the body of published knowledge and to solicit information from people actively involved in geodetic research at HartRAO. Key theoretical information will be used to develop a contextual framework. The data/information obtained from this study will guide and contribute to the creation of a new geodetic research data management system (GRDMS) for HartRAO and the future implementation thereof.

1.2 Problem formulation

Research is a process of inquiry during which information is collected, synthesised, analised, interpreted and applied (Davis, 2014b:10). One of the core elements of the research process is the research problem and the investigation thereof. The research problem clarifies the purpose of the study. Identifying a research problem and providing a suitable description that reflects its precise nature are the most important steps in any research project (Bezuidenhout & Davis 2014:60).

The Space Geodesy Programme at HartRAO is expanding its geophysical research capabilities with the construction, installation and implementation of new geodetic and seismic instrumentation (Coetzer et al. 2015). Scientists predict that the new instruments will generate additional large volumes of data that will require proper data management to ensure effective and efficient distribution and usage. It is anticipated that GNSS and National Academic co-located Seismic Network (NACSN) instrumentation alone will stream approximately 300 megabytes (MB) of data per station per day. The current storage space available on the existing system for storing GNSS and seismic data is only 180 gigabytes (GB) (Coetzer et al. 2015). In addition to the storage requirements predicted for the next ten years (as set out in Table 1.1), data must be organised and structured according to international data standards and in such a manner as to facilitate and promote data retrieval and accessibility.

HartRAO's current geodetic data management and storage system has many drawbacks and inefficiencies. When the system was originally designed, some of the new techniques had not been implemented at HartRAO yet. The system was not designed to cater for the new technique-specific data types. The current system is outdated, fragmented over various systems (even for single data types) and segmented. This hampers management and monitoring of data and requires multiple storage upgrades (Coetzer, Botha & Jacobs 2018). Data is stored on different servers in different geographical locations (Coetzer et al. 2015) and a huge drawback is the current system's limited storage capacity. Little or no automated managing and monitoring are available. The designers used different programming languages and scripts which led to software problems and problems with contingencies. Different approaches to the organisation and structuring of data were applied. There is very little consistency or uniformity in the manner data is stored. Access to data holdings is limited due to the outdated terminology being used (Coetzer, Botha & Jacobs 2018) and the user interface is archaic with many broken links. Webpages are not updated regularly, which makes for inefficient searching. To add to the problem, the system has limited capabilities, e.g. it cannot capture metadata efficiently. When comparing the current available space to the space required over the next ten years (Table 1.1), the limited storage capacity of the system becomes evident (Coetzer & Botha 2016).

Table 1.1: Technique-specific data volumes and future storage requirements for HartRAO's expanding Geodesy Programme

Technique (Instrument)	**Storage frequency**	**Data type**	**Data products**	**MB/ station/ day**	**Current space (GB)**	**Required space GB) (~10 years)**
Satellite Laser Ranging (SLR)	Satellite pass	CPF file	Orbital data	10	1	37
Lunar Laser Ranging (LLR) (New)	Tracking session	Compressed text files	CPF and orbital data	10	0	37
Roscosmos (New)	Tracking session	Compressed text files	CPF and orbital data	10	0	37
VLBI (Products only)	Per experiment	NGS card files	ITRF, ICRF, station position and motions	10	1	300
GNSS	Daily 24 hour	RINEX and SINEX files	ITRF, station position and motions, PWV	300	150	45150
Seismic (New)	Daily 24 hour	Seedlink records	Seismic event data	300	30	45030
Gravimetric (New)	Daily 24 hour	Compressed text files	Gravimetric variations	3	0	12
Meteorological	Daily 24 hour	Compressed text files	Variations, trends	5	1	18
Tide gauge	Daily 24 hour	Compressed text files	Tidal periods, extremes	5	1	20
TOTAL				**653**	**184**	**90641**

(Coetzer & Botha 2016)

Gigabyte (GB) = 1000 MB; terabyte (TB) = 1024 GB) (Inspedium 2018)

As indicated in Table 1.1, and in light of the current research focus, there are several *new* techniques/instruments, data types and products, and additional data volumes for which the current data management and storage system is unable to cater. The current storage capacity is 1.84 terabytes (1840 GB), whereas the predicted storage capacity required over the next ten years is 90 terabytes (90641 GB) (Coetzer & Botha 2016). This presents a strong indication that HartRAO's current data management and storage system will be inadequate to cater for HartRAO's scientific needs. Continued use of the current system, with all its challenges and drawbacks, together with the anticipated amount and influx of data, will lead to an increase in management, monitoring and storage upgrades. A single integrated system with management, monitoring and combined storage will alleviate these problems.

1.3 Research purpose, objectives and questions

The research problem clearly defines what intrigues the researcher and focuses on what the researcher will study Jansen (2017). It is the beacon that guides research in finding answers to the problem within the context of the research. According to Neuman (2011), the research purpose refers to a broad topic or focus area that the research should address.

The main purpose of this study is to investigate what has been reported on the research problem in the global knowledge base, to provide a contextual framework which HartRAO's system designers could use to create a comprehensive geodetic research data management system, and the future implementation thereof. The data reported here is a subset of a larger study that comprises the design of a geodetic research data management system. The design of a new GRDMS should comply with current requirements and future technological expansion at HartRAO, so as to meet the needs and requirements of local and international scientific communities and data service providers (i.e. global data service providers such as the CDDIS and ILRS) as well as those of universities, researchers, students and members of the public.

Within the context of the research purpose, research objectives and questions are required to guide the focus of the research. According to Babbie & Mouton (2011:76–

77), *research objectives* refer to what is needed to be studied and what the unit of analysis should be. Aligned to the objectives, research questions provide guidance to find answers to research problems. There are two broad categories of research questions – i.e. empirical questions, which refer to what can be observed and measured – and non-empirical questions, which involve interpretation that often involves aspects and processes that cannot be observed. Research objectives and questions related to this study are presented in Table 1.2.

Table 1.2: Research objectives and questions

Research objectives	**Research questions**
Identify information related to geodesy on data management, data management system and data management system design, by means of data collection techniques.	What type of information, related to geodesy data management, data management systems and data managements system design, should be considered for design of the GRDMS?
Identify conceptual models to be considered for the design of the new GRDMS for HartRAO by means of data collection.	Which conceptual models should be considered for the design of a GRDMS for HartRAO?
Suggest geodetic data management system components, characteristics and criteria as well as data structuring and organisation relevant for the design of the new GRDMS.	Which geodetic data management system components, characteristics, criteria, data structuring and organisation can be proposed for the design of the GRDMS?

The alignment of the research objectives with the research questions are discussed in Chapter 6 (see Figure 6.1).

1.4 Motivation for the study

Astronomy and geodesy instrumentation hosted by HartRAO generate large volumes of raw, high-quality astronomical, astrometric and geodetic data. HartRAO provides the collected data to global data service providers. Members of the scientific user community access HartRAO's data through these international data service providers (Coetzer et al. 2015).

HartRAO is expanding its geodetic capabilities with the construction, installation and implementation of new geodetic and geo-seismic instrumentation. The new instrumentation will produce additional large data volumes and different data types. As is evident in Table 1.1, the current HartRAO data management and storage unit cannot manage the different types of technique-specific datasets and the large volumes of additional data.

The situation has become critical, because the completed geo-seismic installations at HartRAO, the South African Nuclear Energy Corporation (NECSA) in Gauteng, the Nelson Mandela University in Port Elizabeth in the Eastern Cape and Matjiesfontein in the Western Cape are already streaming geo-seismic data to HartRAO, overloading the existing servers. Furthermore, in the near future, HartRAO will not only supply raw data and data products but will also become a geodesy data correlator and analysis centre for African geodetic datasets. The design of a new GRDMS will ensure that HartRAO maintains its mandate of being a regional data service provider of high-quality publicly funded data and products in a network of international data service providers.

1.5 Benefits of the study

The conceptual framework proposed in this study highlights the information that can be considered in the design of a new GRDMS. This study aims to contribute towards the design of the GRDMS and not the implementation of thereof. Knowledge and expertise gathered will form part of the theoretical background for designing a new GRDMS. The designers envisage an automated system that will handle externally and internally structured data – i.e. it will automatically process incoming raw data into data products and automatically generate data holdings summaries, with less human interaction and, therefore, less opportunity for human error and will allow for a more efficient system, independent, yet interoperable and compatible with international data service providers (e.g. the CDDIS, IVS, etc.). Also, it will provide all of HartRAO's geodetic data at a central location with multiple access points. The researcher predicts that this study will contribute to the global geodetic data management service and library and information science knowledge base. In Chapter 6, more detail is given on additional benefits.

1.6 Research design and methodology

In line with the relevant objectives and research questions, the research design selected for the study provides a set of guidelines and instructions to be followed in addressing the research problem. It can also be seen as the strategy that the researcher uses to implement the research design or a systematic way to investigate, explain, describe, predict and resolve the research problem (Rajasekar, Philominathan & Chinnathambi, 2013).

The research methodology process followed in this study, as illustrated in Figure 1.1, is outlined here, with a comprehensive discussion provided in Chapter 4.

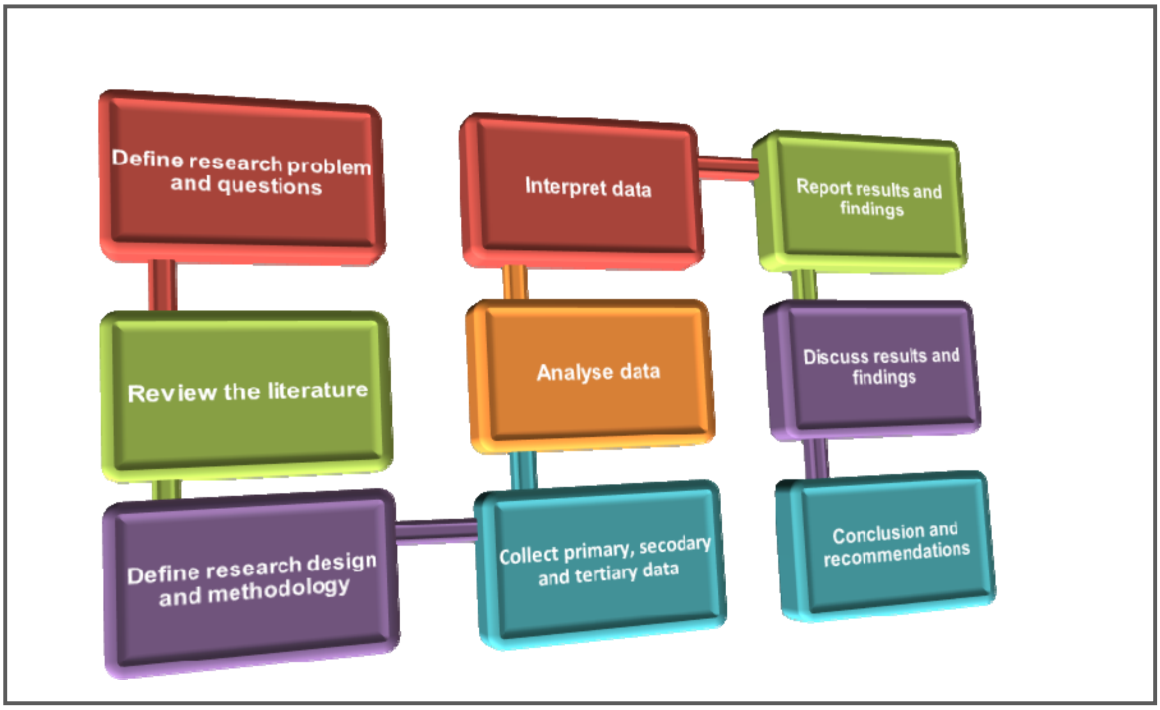

Figure 1.1: The research methodology process

(Adapted from Gurav 2017)

As displayed in Figure 1.1, the research methodology process comprises of various activities, which are outlined in the following sections.

1.6.1 Defining the research problem and question

The first step in the methodology process is defining the research problem and questions. The research problems and questions relating to the research topic were discussed in section 1.2.

1.6.2 Reviewing the literature

The literature review relates to interrogating authoritative resources produced by experts in the field of data management, data management systems and design and data structuring and organisation. The literature consulted in the research process is discussed under appropriate headings in Chapters 2 and 3.

1.6.3 Defining the research design and methodology

The research design and methodology involve the identification of the research paradigm, approach and strategies and data collection methods and tools. The research philosophy utilised in this study is that of critical realism, with an exploratory research approach. Qualitative research techniques with a deductive approach were applied. A detailed discussion is provided in Chapter 4.

1.6.4 Collecting the data

Primary, secondary and tertiary data, linked to the research problem, were collected by means of an in-depth literature review and one-on-one interviews. Linked to the case study method used, the target population was researchers in the field of space geodesy and the target population included geodesists, post-doctoral, doctoral and Master's students employed to conduct research at HartRAO. A systematic purposive sampling method was used. The sample selection was based on the ability of those selected to provide data relevant to the study.

1.6.5 Analysing the data

This study followed constant comparative methods of data analysis, which entailed comparing categories and concepts as they emerged with those already in existence. Interpretative qualitative content analysis was conducted, based on the content of text data gathered by means of the literature survey. Data collected by means of one-on-one interviews were also analysed via a coding process. A detailed discussion of the analysis process can be found in Chapter 5.

1.6.6 Interpreting the data, reporting and discussing results and findings

In this study, the interpretation and reporting of findings were presented and discussed according to theme and category in Chapter 5. These themes relate to the following themes: Management of scientific data (Theme 1); Scientific data management systems (Theme 2; Data management systems used for geodesy data (Theme 3); Data standardisation (Theme 4); and Data management system architecture (Theme 5). Subcategories, directly linked to the main themes and categories, were reviewed and discussed in the context of the interviews and the literature reviewed.

1.6.7 Conclusions and recommendations

The conclusions of this study are based on findings presented in Chapter 5. Conclusions and recommendations were formulated concerning the design framework, components, characteristics and criteria of research data management systems, data standardisation (structuring and organisation of data) and conceptual models that should be considered in the design and development of a data management system. Conclusions and recommendations are presented in Chapter 6, according to themes and categories.

1.7 Unit of analysis

According to Babbie & Mouton (2011:85), the unit of analysis in the social sciences can be individuals, groups, organisations and social artefacts. In this study, the unit of analysis is the design of a GRDMS. Individuals within the organisation provided their input and they were asked to comment on and make suggestions about a prototype GRDMS for HartRAO to ensure that the GRDMS complies with institutional needs and demands.

Interpretive qualitative analysis of empirical data collected by means of one-on-one interviews with reference to the literature reviewed was conducted. In this study, the qualitative data analysis method consisted of a three-stage analysis, which included data reduction, data display and drawing of conclusions. Data were transformed into findings through qualitative analysis and interpretation.

1.8 Scope and delimitations of the study

The scope or domain of a study refers to the relevancy of components to a particular study. Enslin (2014:275) is of the opinion that the scope of a study is defined by its focus. Delimitations are parameters or borders set for a study and can be applied to any part of the research scope, any area of the study, or to challenges encountered during the implementation of the research design. The scope and delimitations assist in focusing and managing a study and in supporting the validity and reliability of the results and findings. Delimitations result from the decision made by the researcher and must be noted, declared and justified to ensure rigor.

In this study, the focus is on geodetic data management and data management systems, including architectural/conceptual models, components, characteristics and criteria, as well as data structuring and organisation. The study also focuses on exploring the global knowledge base for relevant resources by conducting an in-depth literature search. Results from the literature review will be used to provide a contextual framework, which includes components, characteristics, criteria, data structuring and organisation methods, as well as a conceptual model to be considered for the design of a new GRDMS. The delimitation of the study is influenced by the assumption that the definition of what constitutes a good data management system varies between stakeholders' perspectives and system purposes. No system is effective in an absolute sense and, because of this assumption of relativity, a managerial perspective was chosen, implying that an effective data management system will requires most stakeholders being reasonably satisfied.

According to DBMS Intervals (2018), there are three levels in the design of a system – the conceptual, logical and physical design levels. For the purpose of this study, only the conceptual design level of a system was investigated. The findings and recommendations will be presented to the system developers for future consideration. The construction, implementation, testing, revising, integration and management of the

GRDMS fall outside the scope of this research. However, cognisance is taken of these variables and concepts, as they inform the recommendations for future research presented in the final chapter.

1.9 Key definitions

Concepts are defined in terms of their theoretical meaning and application (Bezuidenhout (2014:42). Keyton (2011:41) observes that concepts can be objects, events, relationships or processes and, in this way, they are words that can have many different meanings.

Within the data management systems field, there is an implicit language used, which contains terms such as: *data, data type, data structuring, data management, data management systems, repositories, system design*, etc. When studying data management systems, these terms can be converted into points of observation (Palmius 2007). In this section, definitions of key theoretical concepts, drawn from various information resources, are provided to define terminology, in an attempt to avoid possible misconception and misunderstanding surrounding the problem being investigated.

1.9.1 Astronomy

Astronomy is the scientific study of celestial objects, such as stars, planets, comets, gas, galaxies, gas, dust and other non-earthly bodies and phenomena that originate outside the atmosphere of the earth (ScienceDaily 2018; Taylor Redd 2017).

1.9.2 Data

Depending on its constitution, data can be characters or symbols on which operations are performed by a computer, being stored and transmitted in the form of electrical signals and recorded on magnetic, optical, or mechanical recording media (*English Oxford living dictionary* 2018).

1.9.3 Data management

Data management, which involves the action of managing information and data after it has been generated, includes the management of the receipt, structuring, organising, archiving and storage of data for future access and use (i.e. the management of data throughout the entire data life-cycle). It also includes the management of the development of a system (i.e. architectural design, implementation, functioning, testing, debugging, fine-tuning, revising, integration and enhancement of a system) (Holl, 2010:201; Penn State University Libraries, 2015; TechTarget, 2015).

1.9.4 Data management system

A data management system is a computerised software system that facilitates the creation, maintenance and use of electronic databases. It is designed to perform data management activities, such as the overall management of the data life-cycle, and it provides facilities for users to either manipulate data in databases or manage the structure of the databases (*The Free Dictionary. Thesaurus* 2018; IBM Knowledge Centre 2010).

1.9.5 Digital data repository

A digital data repository is a digital platform on which data is deposited and preserved. It is also a platform on which data is analysed and assessments are shared. The aim of a digital data repository is to support the discovery, use, manipulation, management and re-use of primary and secondary data (University of Minnesota Libraries 2015).

1.9.6 Data organisation

Data organisation is the act of classifying, sorting, naming, structuring and organising data, datasets, files and folders to allow and support easy location/retrieval, accessibility of data and to ensure data integrity (Georgia Technical Library 2018; University of Cambridge 2018).

1.9.7 Data standards

Data standards refer to protocols that facilitate compatible communication and interoperability between instrumentation and computer systems. Hypertext transfer protocol (HTTP) is an example of a protocol that is used worldwide (University of Minnesota Libraries 2015).

1.9.8 Data structuring

Data structuring involves the organisation of information/data into structures, such as queues, stacks, linked lists and trees, etc. (Black 2014). It is a system where random, unstructured data can be taken as input and a number of operations executed on it (Williams 2015).

1.9.9 Data type

A data type is a particular kind of data item defined by the values it can take; the programming language it can use; or the operations that can be performed on it (Shaffer 2008:9). Within a computer system, it determines the type of data permitted in a field, e.g. numerals only or a combination of numerals and letters or symbols within a field (Watt 2015).

1.9.10 Geodesy

Geodesy, also known as *geodetics*, is a branch of applied mathematics and earth sciences that deals with the measurement and representation of the earth (or any planet), including its gravitational field, in a three-dimensional time-varying space (Torge & Muller 2012:1).

1.9.11 Radio astronomy

Radio astronomy is a subfield of astronomy that focuses on the study of celestial objects at radio wavelengths/frequencies (Kraus 1989). Astronomers use radio telescopes to explore the Universe by detecting radio waves emitted by celestial objects (SKA 2018).

1.9.12 Space geodesy

Space geodesy is the branch of geodesy that studies methods of determining the relative position of points on the earth's surface, dimensions and shape of the earth, and the parameters of the earth's gravitational field using space techniques (Seeber 2003).

1.10 Thesis structure

The dissertation consists of six chapters, which are outlined below.

Chapter 1 introduces the research. The problem statement, aim, objectives and research questions are briefly introduced. Details pertaining to the motivation for embarking on the research, the significance, scope and delimitations of the research and the research methodology underpinning the study are also presented.

Chapter 2 involves the theoretical background to the research. Constructs are discussed to contextualise the theories and applications. Different views and theories are discussed under appropriate subheadings. Details on conceptual models used for the design of data management systems (DMS) are provided, with the emphasis on relevancy to the HartRAO scenario.

Chapter 3, which relates to the literature review, focuses on technological advances and needs at HartRAO and on conceptual frameworks, components, characteristics, criteria and data structuring and organisation to be considered in the design of the GRDMS, as determined from an in-depth literature analysis.

Chapter 4 provides the research methodology that underpins the study, which consists of a case study design using multiple data collection methods. Details pertaining to the justification of the chosen approaches are also presented and discussed.

Chapter 5 focuses on the qualitative analysis of empirical data collected by means of face-to-face interviews, supported by findings from the literature reviewed. Findings based on the participants' views of the proposed design and the literature surveyed, are discussed.

Chapter 6 provides a brief reiteration of the research objectives and provides recommendations for the design of the new HartRAO GRDMS. These recommendations are based on conclusions drawn from the literature reviewed and participants interviewed. A proposal of a design model that can be used as a blue-print for the development of a new GRDMS for HartRAO is presented. The chapter concludes with suggestions for future related research in South Africa and Africa.

1.11 Conclusion

In this chapter, an introduction and background to the research topic were presented. The research problem was formulated, and the research purpose, objectives and questions were addressed. This chapter contains a discussion of the motivation for conducting the study, as well as a summary of the importance and benefits of the study to science and society.

In line with the research objectives and questions, the research design selected for this study was indicated and the research methodology was illustrated and discussed. The unit of analysis employed was briefly discussed. A detailed discussion of the data analysis, interpretation and synthesis process followed in this study are presented in Chapter 5.

In Chapter 1, the scope and delimitations that guided and supported the study were highlighted and discussed. The variables and concepts that fell outside the scope of this study were also identified. Key definitions that clarify, support and underpin the research topic were introduced. The chapter concludes with an explanation of the structure of the dissertation. In the next chapter (Chapter 2), the theoretical framework to the study is presented.

CHAPTER 2: THEORETICAL FRAMEWORK

2.1 Introduction

Major advances in the understanding of geophysical phenomena are built on the scientific foundation of high-quality data, including knowledge and understanding of the origins and evolution of the universe, the rotational dynamics of the earth and its possible correlation with earthquakes and plate motions (Eubanks 1993; Fox & Harris 2013). According to Koopman & de Jager (2016:1), high-quality data have intrinsic and commercial value that, if managed properly, can enhance and support the expansion of science and human capital development. Utilised properly, it can be beneficial to socio-economic development.

In the research process, data is generated at different stages, which is referred to as the *flow of data* (Ray 2014:10). Once data have been collected, it needs to be structured, organised, archived and stored to ensure future access and usage. This is referred to as *data management*, which includes the development of systems, policies, practices and procedure for the management of data (Holl 2010:201; Penn State University Libraries 2015). For data to be effectively utilised, users need to know what is available; where it is stored; and how it can be retrieved. In an attempt to make publicly funded data discoverable, accessible, citable, reviewable, reproducible and reusable, an integrated data management system is required (De Waard, Cousijn & Aalbersberg 2015).

This chapter provides the theoretical framework of the study, linked to data management, and focuses on the data life-cycle and data management system design. The chapter starts with the data life-cycle theory of Behrend (2013), as applied to astronomy and geodesy data. This is followed by an investigation of architectural/conceptual models of data management systems involved in the research problem of this study.

2.2 The data life-cycle model

Data is the core of any research activity and in any research domain, be it historical research, experiments, measurements or surveys (RWTH Aachen University 2010). Data, which are generated in many different formats, are heterogeneous in both format and content (Trauth & Sillmann 2013).

In space geodesy, geodetic / astrometric data obtained from various equipment/instruments require effective and efficient management. To assist institutions/observatories in making publicly funded scientific data accessible, networks of operational, regional and international geodesy data centres serve as repositories for raw and calibrated scientific data (Behrend 2013). For example, data related to VLBI activities are organised through the International VLBI Service for Geodesy and Astrometry (IVS), which is an international collaboration of institutions/observatories operating to support the management of data from VLBI technologies (Behrend 2013). Radio astronomy/geodetic observatories in multiple countries host various geodetic technique-specific instrumentation. Data management systems are required to manage the extensive volumes of geodetic, astrometric and geophysical data generated by the various instruments. To manage their scientific data holdings, observatories perform activities that focus on managing data throughout its life-cycle (Trauth & Sillmann 2013).

The data life-cycle is a high-level overview of the different stages involved in the flow of data. According to Ray (2014:10), the life-cycle concept assists with focusing system designers' attention on issues such as data quality, documentation, preservation and data sharing at the time of creation. The general stages of a data life-cycle are: data creation, processing, synthesis, analysis, preservation, access, use and archival (Chisholm 2016). In this study, the work of Behrend (2013) was consulted, in order to reach an understanding of these data life-cycle stages. Behrend's (2013) illustration of the data life-cycle (Figure 2.1) depicts the flow of geodesy and astrometry data through its life-cycle.

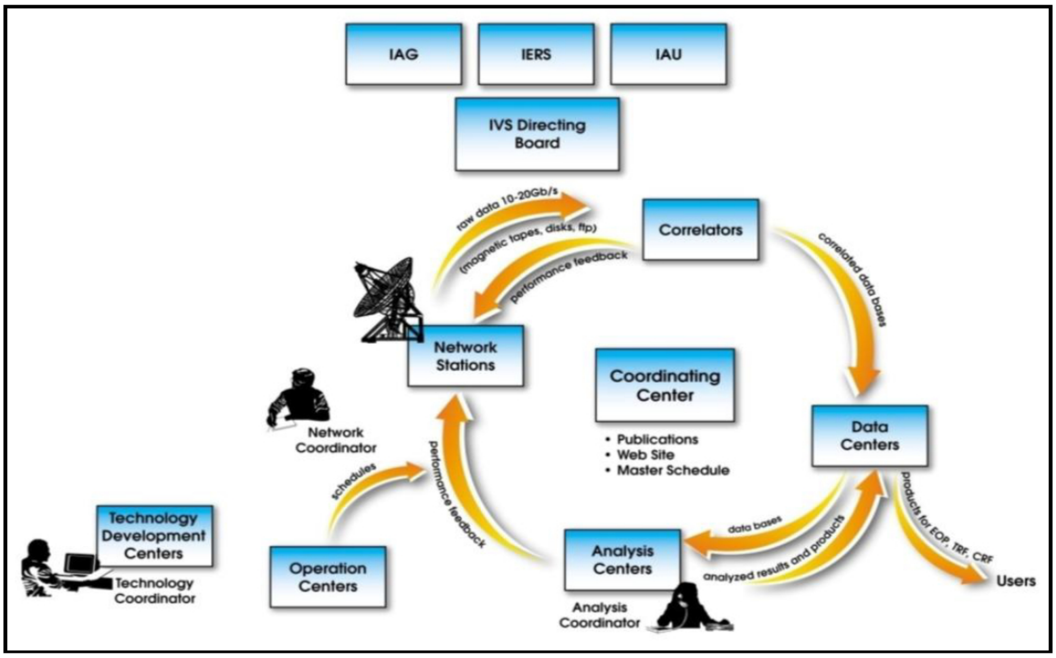

Figure 2.1: Geodesy and astrometry data life-cycle

(Behrend 2013:n.p.)

According to Behrend (2013), data creation, as the first stage in the flow of data, involves observation of celestial and terrestrial objects with the aid of astronomical and geodetic instrumentation at network stations. The second stage is the shipment or transfer of recorded raw data to pre-determined correlators, where it is correlated and reduced to database formats (Noll 2010). During the third stage of the life-cycle, databases are uploaded to regional or global data centres, where the data is organised, archived, stored and preserved in databanks. During the fourth stage in the life-cycle, analysis and technique combination centres download the databases from the data centres. Data is analysed and the results, together with technique-specific and combined data products, are uploaded to repositories and data centres. During the fifth stage of the life-cycle, data centres provide analysed data, technique-specific data and combined data products to data service providers and the scientific user community. The last stage of the cycle consists of performance feedback provided by the correlators and analysis centres to network stations and government bodies (Behrend 2013:n.p.). The last stage may result in the continuation of research and the re-use of data, which can lead to new research activities.

HartRAO follows a similar (but not exact) approach to the data life-cycle as proposed by Behrend (2013:n.p.) and illustrated in Figure 2.1. An illustration of HartRAO's VLBI data life-cycle is shown in Figure 2.2.

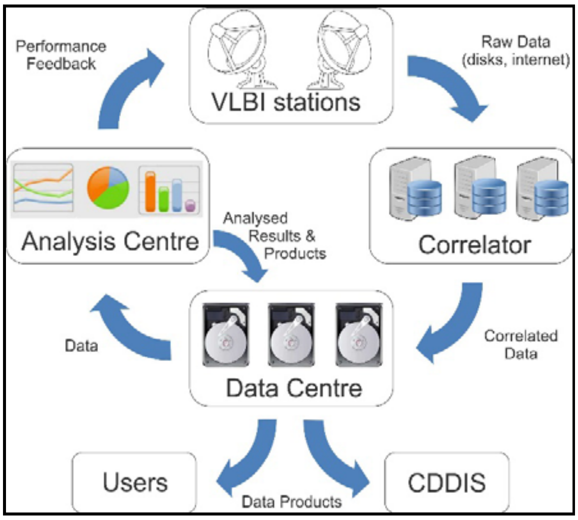

Figure 2.2: Data life-cycle model for HartRAO's VLBI data

(Coetzer et al. 2015:24)

With reference to Figure 2.2, geodetic/astrometric VLBI data obtained by the HartRAO radio telescopes are recorded to storage media and transferred to pre-determined international data correlator facilities and data service providers (HartRAO 2012). At the correlators, the data is correlated and reduced to database formats. The database files are then uploaded to data banks at various international data centres. Analysis centres extract the data from the data banks and analyse the data. In return, the analysis centres provide the products derived from the data to the data centres. Users can access the analysed data and products hosted at the various international data centres. Analysis centres provide station performance feedback and other statistics to stations, such as HartRAO. Station reports are then submitted to the governing funding agencies (Coetzer et al. 2015).

With the expansion of HartRAO's geodetic instrumentation and techniques, the HartRAO data life-cycle requires revision. HartRAO will not only produce data in future but will also become a correlator and analysis centre for African geodetic datasets. Current and additional data types from new technique-specific data, such as geo-seismic data, including that of instrumentation currently under development, will be correlated, reduced, analysed, archived and stored at HartRAO and supplied to regional and international data service providers and made accessible to members of the scientific user community. Subsequently, the need has developed for a new GRDMS for HartRAO, which considers the data life-cycle and activities that should be included to ensure effective management of various data types. In Chapter 6, the proposed steps within the data cycle of the new GRDMS are discussed in more detail.

2.3 Data management system design

Data management consists of a broad range of stages, from research and technical activities to administrative aspects of managing data (Martin & Ballard 2010:1). In a broad sense of the word, data management stages include activities that range from data collection, storage, retrieval, data policies, data ownership, custodianship, data quality control, documentation and organisation (Martin & Ballard 2010:1). According to Rob & Coronel (2007:640), data addition, deletion, modification and listing can be added to the aforementioned activities.

Data management activities in the geodetic scientific community include data sharing among researchers and secondary users; data preservation; data re-use over long periods of time; and the storage of large volumes of data (Nelson 2009; Ray 2014:2). In addition, Noll (2010), Behrend (2013) and Ray (2014:2) identify the following data management activities that are performed in the geodetic scientific community: the creation, correlation, provision, dissemination, archiving, analysis and use of raw and semi-processed geodetic data; the creation of products derived from the data, the re-use of data; and the storage and preservation of large volumes of data over long periods of time. In recent years, more activities have been added to the list of existing data management activities, such as the structuring and organising of data according to file naming conventions (FNCs); metadata schemes; assigning of digital object identifiers (DOIs); and open researcher contributor identifications (ORCIDs) to datasets. The aim of including such identifiers is to assist with the structuring and organisation of data and to identify, locate/retrieve and cite data (Clark 2012; Fox & Harris 2013:WDS3; Tenopir, Birch & Allard 2012; Riley 2017).

To accommodate the data needs of the geodetic scientific community, centralised data management systems aligned to the data management activities mentioned above are used by the Crustal Dynamics Data Information System (CDDIS) and University NAVSTAR Consortium Geodesy Seamless Archive Centers (UNAVCO GSAC) systems to facilitate geodetic data management services (Wier, Rost & Boler 2016:4). The CDDIS system was designed to store all geodetic data (raw and analysed data) and data products in a central data repository. The aim of the system is to archive data and disseminate it to the scientific community. The CDDIS system design includes a relational database management system (RDBMS) and within this system, data sets received from network stations are organised and catalogued for easy retrieval (Noll, 2010). Multiple accesses are provided through the online interactive menu-driven interface of the system. The UNAVCO GSAC system provides a range of capabilities to connect data users to the data repository via the Internet (Wier, Rost & Boler 2016:4). The UNAVCO GSAC system provides information / metadata regarding GPS stations / sites instruments and data files. Users can download GPS receiver data files by using file uniform resource locators (URLs). Users can also request stations, instruments and instrument data files via the system. As Wier, Rost & Boler (2016:4) observe, the UNAVCO GSAC system reads data from the database and sends the results to end-users.

It is important to consider these system activities when planning a new GRDMS for HartRAO, as they will influence the outcome of the design. Up until now, HartRAO has used an adapted data management system (DMS) for the collection and distribution of geodesy data. The current DMS is hosted on computers connected to local servers running on Linux Operating System software. The new GRDMS should be designed to collect, archive, organise and store raw and processed data according to data standards (e.g. NGS standards). In addition, it should provide rapid access to data in a user-friendly manner; preserve data integrity; promote data independence; avoid data redundancy; ensure data security; and provide procedures for data maintenance (Orsborn 2013; Valle 2013; Watt 2015).

The design of a system should be done at different levels – i.e. the conceptual, logical and physical design levels (Wier, Rost & Boler 2016:4). For the purposes of this study, the conceptual design level of a system was investigated, because the focus of this study is on presenting a contextual framework for the design of an RDMS and not the actual construction of the system. The logical and physical design levels are more suitable and relevant to the physical construction, implementation, testing, revising, integration and management of a system, which does not form part of this study.

A tool that may assist and guide the design of the system is the architectural/conceptual design model by Wang, Shen, Xie, Neelamkavil & Pardasani (2002:981). The aim of a model is to provide a formal description and conceptual representation of the design of a system, used to support reasoning related to the structure of the system (Jaakkola & Thalheim 2011). It involves the formulation of abstract ideas with approximate representation, which are subsequently evaluated against system requirements. The importance of the latter is that they yield a description of the system components, architecture and characteristics (Rob & Coronel 2007:363). The application of the architectural/conceptual model can affect the overall system effectiveness and usability. It is extremely difficult – even impossible – to compensate or to correct the shortcomings of a substandard design concept (Wang et al. 2002:981). Therefore, extreme care should be taken when designing a DMS to ensure that all components of the system have been considered.

In terms of what constitutes an architectural/conceptual design model, the research by Wier, Rost and Boler (2016:4) related to the UNAVCO GSAC conceptual model, Noll's (2016 & 2017) research related to the CDDIS conceptual model and the research by Mashaba, Combrinck, Botai, Munghemezulu and Botha (2016) on the GNSS DMS model, were reviewed. The conceptual design models of these authors are particularly relevant to the geodetic environment and the scope of this study, in that they provide detailed information on the construction of a well-organised and user-friendly DMS.

2.3.1 The UNAVCO GSAC Conceptual Model

Wier, Rost and Boler (2016:4) opine that the UNAVCO GSAC data management system focuses primarily on managing GPS datasets and the provision of access to raw data, ancillary data and metadata. The system consists of various components, including an application server, subsystems within the application server (e.g. a query and output handler), repository with subsystems and servers for data transfer (Figure 2.3). The system uses Dataworks for GNSS open source software modules (discussed in Chapter 6) to handle data acquisition, downloading to local repositories, data management within the repositories and the transfer/sharing of data with the scientific community.

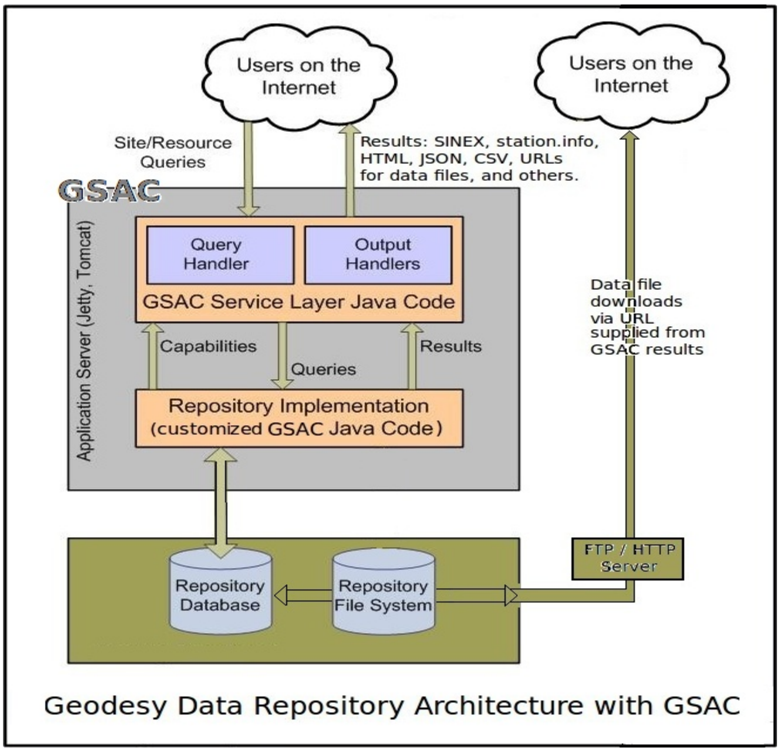

Figure 2.3: Conceptual model of the UNAVCO GSAC system

(Wier, Boler & McWhirter 2012:1)

The model by Wier, Boler & McWhirter (2012) (Figure 2.3) illustrates the entity-relationship, which can be used to access the application server, where both the *GSAC Service Layer* and the *Repository Implementation Layer* are stored. The *GSAC Service Layer* serves as the query and output handler. Queries are communicated via the

Repository Implementation Layer to the *Repository Database*. Query results are returned to the *Repository Implementation Layer*. The output handler in the *GSAC Service Layer* provides the results in different file types and formats (e.g. SINEX) as requested by the user. The *Repository File System* can also provide data via the *ftp / http Server*. Data files can be downloaded by the user via the URL supplied from the UNAVCO GSAC results.

2.3.2 CDDIS Conceptual Model

The CDDIS is a central facility that provides users with access to geodetic data and derived products to facilitate scientific investigation. The system archives and stores VLBI, GNSS (GPS, etc.) Satellite/Lunar laser ranging (S/LLR) and Doppler Orbitography and Radiopositioning Integrated by Satellite (DORIS) data online and has been supporting the archiving and distribution of VLBI data since its inception in 1982 (Noll 2016). The CDDIS archive of VLBI data and products is accessible to the public through anonymous ftp (at ftp://cddis.nasa.gov) and on the Web (at https://cddis.nasa.gov/archive) (Noll 2016).

The CDDIS conceptual model (Figure 2.4) is based on a distributed server environment (Noll 2017). Distinct servers, with online backup components, handle incoming, archiving and outgoing operations.

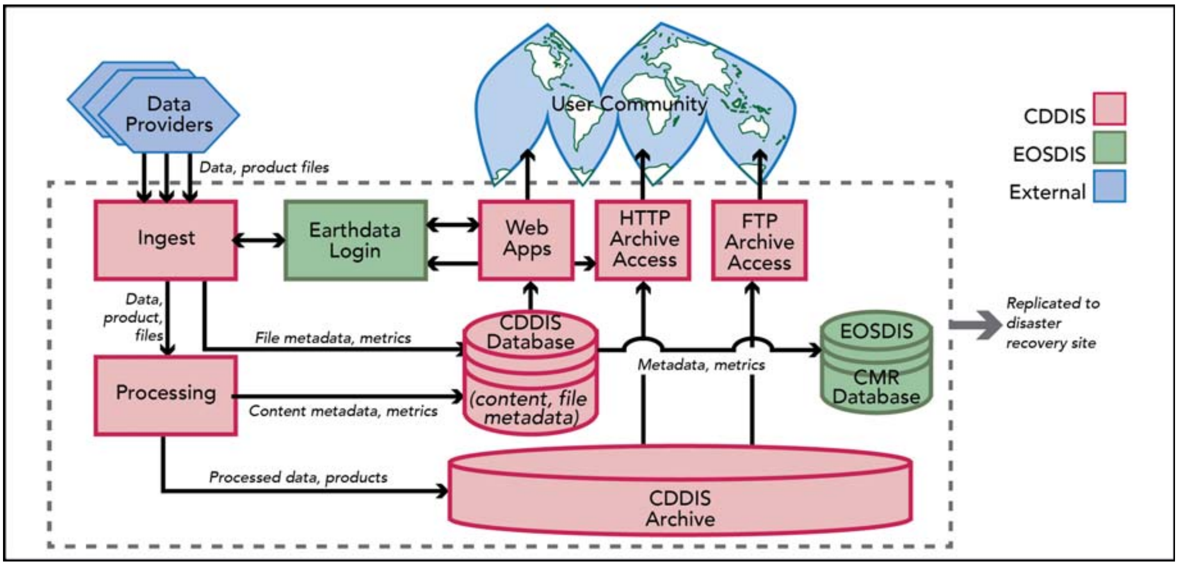

Figure 2.4: Conceptual model and system architecture of the CDDIS facility

(Noll 2017:3)

As shown in Figure 2.4, incoming data are on the host computer, *cddisa.gsfc.nasa.gov*, receives data and product files from data providers, consisting of network stations (of which HartRAO is one) and data centres. By using specified file names, data files are transferred to appropriate directories. From the *Ingest* unit, deposited data and product files are transferred to the *Processing* unit. File metadata and metrics are sent to the CDDIS database, which retains the content and file metadata. The CDDIS database provides metadata and metrics to the Earth Observing System Data and Information System (EOSDIS) database (Noll 2017). Processed data and data products are stored in the *CDDIS Archive* unit. Automated archiving routines peruse the directories and migrate any new data to the appropriate public disc area, based on the filename / directory (e.g. data, product, project and other directories), depending on the type of data. Members of the scientific user community can access the *CDDIS Archive* unit via the http or ftp *Archive Access* units (Noll 2017).

The CDDIS was designed to provide multi-user access via an interactive graphical user-interface (GUI) (Noll 2010). The system has a mechanism to set up and control user accounts and has access restrictions on proprietary data, in order to protect the producer of the data and the data integrity. Users have to register on the system to gain access to the data archive. Internet Protocol (IP) range and password authentication are also required as security measures.

2.3.3 GNSS Data Management System Model

Another conceptual model to consider is that of Mashaba et al. (2016), as illustrated in Figure 2.5.

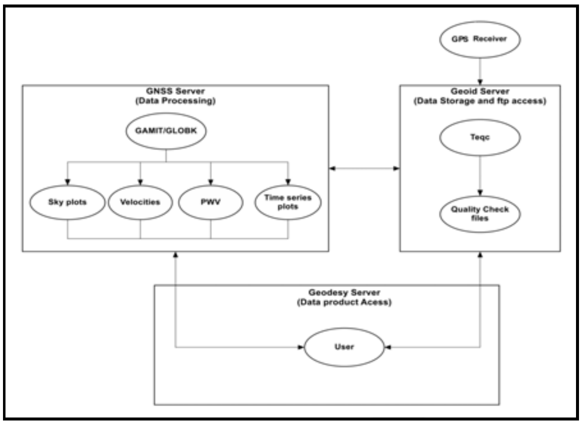

Figure 2.5: Conceptual model of the GNSS data management system

(Mashaba et al. 2016:119)

The model of Mashaba et al. (2016) consists of the following three units: the *Geoid Server, GNSS Server* and the *Geodesy Server*. The conceptual model is based on a distributed server environment, similar to that depicted by Noll (2017). According to the model of Mashaba et al. (2016), GPS data received from GNSS stations via the Internet are received and stored on the *Geoid Server* (the data storage and ftp access unit). From there, the data is transferred to the *GNSS Server* (the data processing unit). These servers interact to make the final data and data products available to the end-users. The *Geoid server* (a data storage repository) archives raw RINEX data files. The *GNSS server* processes the RINEX data files and produces GNSS data products (e.g. sky plots, velocities, time series plots, etc.). The *Geodesy Server,* which hosts the website, is the front-end, where users can access the data of the stations via interactive web maps.

2.3.4 Design components to consider for the GRDMS

Based on the foregoing discussions of the three conceptual/architectural design models, it is possible to make a number of conclusions in terms of what should be considered for the GRDMS design. The design components proposed to be considered for the new GRDMS are a combination of components featured in the conceptual/architectural design models of the UNAVCO GSAC (Figure 2.3), the CDDIS (Figure 2.4) and the GNSS Data Management System model of Mashaba et al. (2016) (Figure 2.5). These three models share similar components, in that all of them have an application server unit with subsystems to handle queries from users and provide access and data to users (via Web applications, http or ftp). In the case of the UNAVCO GSAC model, these components reside in what is called the *GSAC service layer* and in the case of the GNSS model of Mashaba et al. (2016), it is referred to as the *Geodesy server*. Both the CDDIS and GNSS models feature an ingest server unit and a processing server unit. In the case of the GNSS model, the ingest unit is referred to as the *Geoid server* and the processing unit as the *GNSS server*. Both the UNAVCO GSAC and the CDDIS models illustrate a repository/archive server unit and databases.

To conclude: the following design components need to be considered for the conceptual/architectural design model of the new GRDMS: ingest, processing, repository/archive and application/access server units, which should contain subsystems and databases, together with all the necessary network components to support connectivity and data access. These and other relevant components to be considered are discussed in Chapters 5 and 6.

2.4 Conclusion

Data management is a crucial activity for any organisation whose mandate is to supply cutting-edge data and research. Once data have been created, it should manage to support and promote usage. Data management systems facilitate effective data management. The aim of this chapter was to present a theoretical framework for the study. Key information to be considered for the design of the new GRDMS includes the general flow of astronomy and geodesy data.

Relevant information regarding data management activities and key conceptual models to be considered – such as the UNAVCO GSAC (Figure 2.4), the CDDIS (Figure 2.5) and the GNSS Data Management System model of Mashaba et al. (2016) – were investigated, analysed and discussed. Aligned to these conceptual models, various design components/units that should be considered, such as application/user access units, ingest and processing units, archive/repositories, storage units and databases, were also discussed. Chapter 3 subsequently contains a summary of the literature reviewed during this study.

CHAPTER 3: LITERATURE REVIEW

3.1 Introduction

A literature review, which is an important step in the research process, involves the searching, reading, evaluating and summarising of available literature that relates directly or indirectly to the research topic (Howard 2014:101). In this study, a comprehensive literature review was conducted. The term "comprehensive" does not suggest that the literature review is exhaustive toward a totality of literature on any given subject, but comprehensive inasmuch as it should involve the use of rigorous techniques to search, retrieve and collect resources (Onwuegbuzie, Frels & Hwang 2016).

During the literature review, searches were conducted to retrieve secondary and tertiary literature (i.e. peer-reviewed journal articles, conference proceedings, books and grey literature in paper and electronic format) through bibliographic searches in online databases, repositories and websites. Printed monographs within HartRAO's literature holdings were also searched. Hofstee's (2006) literature review funnel method was applied during review of the literature sources.

3.2 The literature review funnel

The literature review funnel entails the review of resources at a broad level. It uses key ideas and themes as a guide for reviewing the literature. A pre-defined review protocol was followed throughout the study. A review protocol specifies the methods that will be employed to undertake a systematic review (University of Durham. Department of Computer Science 2007). For the purposes of the literature review, the protocol includes elements of the rationale for the study and research questions that define the search terms and resources applicable to the research.

In this study, primary and secondary research relating to data management, data management systems and data management system design architecture were reviewed by means of the literature review funnel method proposed by Hofstee (2006) (see Figure 3.1). This model was selected to ensure that literature on data management and data management systems (with specific focus on geodetic data management system design) was condensed from a broad to a narrow focus.

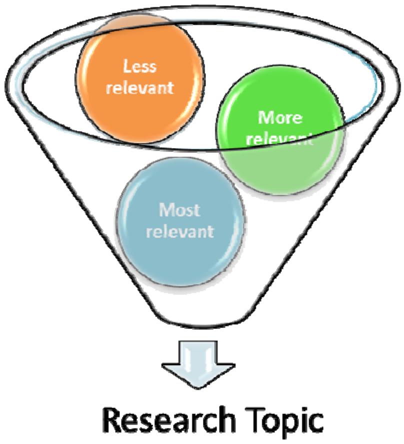

Figure 3.1: The literature review funnel method

(Hofstee 2006:96)

The funnel method of structuring a literature review is designed to ensure that all the objectives of the literature review are met. If the funnel method is applied properly, the credentials and originality, as well as the theory base, context and significance of the study will all emerge with less effort (Hofstee 2006). The author describes the literature review funnel method as grouping works by commonality, followed by the ordering of the constructs into categories and the internal ordering of the categories.

The ball (Figure 3.1) in the top of the funnel (titled "Less relevant") represents categories of works that are relevant to the study, but that do not specifically address the research problem (Hofstee 2006). These tend to contain numerous works per category. There will be more works that are generally relevant than there will be works that are specifically related to the study. According to Levy and Ellis (2006), researchers should consider and become familiar with generally relevant works but should focus on the literature most relevant to the study. This leads to the narrower theory base with fewer topics/concepts that are more applicable to the topic being investigated within a particular research construct.

In this way, the narrow theory base relates to works that are closer to the research topic (Hofstee 2006). There will be fewer works per category and such literature should be dealt with in more detail, as they relate more closely to an aspect of the study (e.g. ball titled “More relevant” (Figure 3.1). As the review process commences down the literature review funnel, the categories should get closer and closer to the topic of the research being undertaken (e.g. ball titled “Most relevant” (Figure 3.1). Researchers may find that by narrowing down the review process, a category may contain only two or three, or possibly even one, literature source/s that relate/s to subthemes of the research (Hofstee 2006).

In this research, an in-depth literature review was conducted to identify literature that relates to the design of data management systems with specific reference to geodetic data management. Categories aligned to the theme of the research, which could be identified from the literature as part of the narrow theory base, related to data management system criteria and characteristics, data management system components as well as data structure and organisation. More details on literature related to these categories are given in the sections that follow.

3.3 Data management system characteristics and criteria

Designers of electronic systems are constantly confronted with expectations to improve on the system. Before designing a system, designers need to know what the goal of the system is and identify characteristics and criteria that should be used for assessing the achievement of the intended goal. The characteristics and criteria could serve as a checklist before, during and after the design phase to determine if the system performs as it is intended to.

In the global knowledge base, there are vast amounts of normative literature on how to design data management systems and how to structure and organise data within these systems. Most of these sources resided in the broad theory base of the funnel model. The researcher identified components, characteristics and criteria, as well as data structuring and organisation of astronomy/geodetic data management systems most often mentioned in the literature to limit the selection of literature towards the narrow theory base. Two matrices were composed to present the characteristics and criteria identified as well as relevant sources within the narrow theory base.

Table 3.1: Matrix of main data management system characteristics and relevant literature

Characteristics	Rob & Coronel (2007)	Noll (2010)	CDDIS (2015)	Thakur (2016)
Store any kind of data				✓
Manage processed data of homogeneous and heterogeneous nature in databases and repositories		✓	✓	
Support the data life-cycle from ingest to deletion		✓	✓	
Support accuracy, completeness, isolation and durability (ACID) properties	✓			✓
Ensure data integrity through query validation and optimisation	✓			✓
Allow concurrent use of various databases to access data				✓
Application interface to interact with end-users	✓	✓	✓	✓
Ability to determine data location and fragmentation	✓			
Able to prepare data for presentation related to applicable software used by end-user	✓			
Ability to backup and recovery data in the event of database failure	✓	✓	✓	
Maintain data consistency through comparison and linkages via various databases				✓
Manage transactions to ensure that data move from one consistent state to another (support BlockChain)			✓	✓
Manage integration and storage of large volumes of raw data		✓	✓	
Provide multi-user remote access via the Internet		✓	✓	✓
Control access to data to protect secure data		✓	✓	
Allow user registration and account management		✓	✓	

It is evident that there are a number of characteristics that should be considered for the design of a new DMS. Based on the literature consulted and the matrix of main data management characteristics relevant to the study (see Table 3.1), the characteristics most often mentioned to be considered and applied when designing the GRDMS include: an application interface; backup in event of failure; and multi-user remote access via the Internet. These characteristics are discussed in work of authors such as Rob and Coronel (2007), Noll (2010), CDDIS (2015) and Thakur (2016).

According to these authors, users must be able to interact with a DMS. To cater for the interaction, the system needs to have an application interface for users. Users can apply and request data, as well as receive data via the application interface. This characteristic can also give an indication of the use of the system and its return on the investment (Pearce 2015).

Another characteristic to be considered is the system's ability to backup and recover data in the event of system failure. Designers of the new GRDMS need to consider proper backup protocols to ensure that regular and reliable backups are conducted. The system must be able to retrieve data and restore it on the system, in the event of data loss due to hardware failures, software bugs, human action or natural disasters. In practice, more than one user usually accesses a system from local and remote sites, laboratories, observatories, offices, universities, homes, etc. To make this possible, systems are designed to cater for multiple user access. Therefore, the designers of the new GRDMS need to consider multi-user local and remote access via the Internet to ensure a system able to cater for both local HartRAO users and users in geographically dispersed locations.

In addition to the characteristics that should be considered during the design of a new GRDMS for HartRAO, sources such as Martin and Ballard (2010:1), Tenopir, Birch and Allard (2012), Behrend (2013), Fox and Harris (2013: WDS3) Wier, Rost and Boler (2016) indicate the criteria that should also be considered. Including these criteria during the design will enable the new GRDMS to manage the vast volumes of geodetic data and to provide the required access to prospective end-users.

Table 3.2: Matrix of main data management system criteria and relevant literature

Criteria	Nelson (2009)	Martin & Ballard (2010)	Noll (2010)	Tenopir, Birch & Allard (2012)	Behrend (2013)	Fox & Harris (2013)	Science HQ (2013)	Valle (2013)	Ray (2014)	CDDIS (2015)	ZDNet (2015)	Wier, Rost & Boler (2016)	Thakur (2016)
Data management life-cycle		✓		✓	✓	✓			✓			✓	
Multi-user access		✓	✓	✓	✓	✓		✓	✓			✓	
Software that receive and interpret user requests		✓	✓	✓	✓	✓			✓			✓	
Transfer ftp and http data		✓		✓	✓	✓						✓	
Structure and organise data for easy access		✓	✓	✓	✓	✓			✓			✓	
Manage large volumes of data		✓	✓	✓	✓	✓		✓	✓			✓	
Supply automated data capture frameworks	✓								✓				
Peruse and organise incoming data from various instruments automatically	✓			✓					✓				
Efficiently manage all files, even at remote sites				✓				✓	✓	✓	✓	✓	✓
Handle different categories of data		✓			✓			✓	✓	✓	✓	✓	✓
Conduct data quality checks			✓			✓	✓			✓		✓	✓
Apply metadata standards and sharing (Internet of Things)								✓		✓	✓	✓	✓
Assign DOIs to			✓							✓			

Criteria	Nelson (2009)	Martin & Ballard (2010)	Noll (2010)	Tenopir, Birch & Allard (2012)	Behrend (2013)	Fox & Harris (2013)	Science HQ (2013)	Valle (2013)	Ray (2014)	CDDIS (2015)	ZDNet (2015)	Wier, Rost & Boler (2016)	Thakur (2016)
datasets													
Ensure database consistency, security and integrity	✓						✓		✓	✓	✓	✓	✓
Support the re-use of data towards research data management			✓	✓	✓	✓				✓		✓	✓
Store large volumes of geodetic data indefinitely	✓		✓				✓		✓	✓	✓		✓
Support user registration and management of accounts	✓	✓	✓		✓	✓	✓			✓		✓	✓
Provide online interactive menu-driven system interface			✓							✓	✓	✓	✓
Supports data independence			✓							✓	✓	✓	✓
Conduct regular backups					✓	✓			✓	✓	✓	✓	✓
Apply different methods and tools to standardise and organisation datasets in support of easy retrieval			✓		✓					✓	✓	✓	✓
Display data holdings to assist users in deep data recovery			✓							✓	✓	✓	✓

From the above it is evident that there is a number of criteria to be considered for the design of a new DMS. Based on the literature consulted and the matrices of main data management criteria relevant to the study (Table 3.2), the criteria to be considered and applied when designing the GRDMS and that are most often mentioned by authors include: multi-user remote access; user registration support; user account management; software that receives and interprets user requests; structure and organisation of data for easy access; management of all files; provide database consistency; security and integrity; support of the re-use of data towards research data management; indefinite storage of large volumes of geodetic data; and conducting regular backups.

3.4 Data management system components

Different systems have different system requirements. According to TechTerms (2018), system specifications can include the following: operating system (i.e. Linux, Windows XP, Mac OS, etc.); processor speed (i.e. Pentium 4, etc.); memory (i.e. 512 MB); graphics card (i.e. Radeon 9800 w/256 MB video memory); hard disk space (i.e. 80 GB); and input/output ports (i.e. Universal Serial Bus (USB), serial, parallel), etc.

According to Noll (2010) and Austin (2016), a data management system consists of sets of interacting smaller systems known as *subsystems* or *functional units*. Each subsystem performs its own defined tasks within the data management system and each subsystem works in cohesion with other subsystems to achieve the overall objectives of the system. Components of these systems are hardware, software, data, processes, procedures and human resources (Noll 2013; LIMSWiki 2016). Aligned to the funnel model, a considerable amount of literature, related to both the broad and narrow theory bases, was consulted to obtain information relevant to this topic. Literature from the broad theory base provided a sound theoretical foundation for the identification of literature from the narrow theory base pertaining to categories relevant to the development of the HartRAO GRDMS. Knowledge of data management system components is important to minimise the risk of omitting vital components during the design phase. Based on the literature consulted, the components to be included in the design of a data management system are briefly discussed in the following sections.

3.4.1 Hardware

Hardware relates to components such as electrical and mechanical parts of a computer, including servers, storage devices, printers, etc. (Rob & Coronel 2007:18). Hardware components work as per instructions from software: without software, hardware simply cannot function. For example, in the case of the CDDIS (Figure 2.4), a dedicated server handles the delivery/ingest, processing and archiving operations of the CDDIS (Noll 2017). The UNAVCO GSAC system components (Figure 2.3) include an application server, ftp / http server, repository and Internet interface (Wier, Rost & Boler 2016). Mashaba et al. (2016) propose a data management system for GNSS data (Figure 2.5), which includes a data processing server (GNSS Server), a data storage and ftp access server (Geoid Server) and a data product access server (Geodesy server).

3.4.2 Software

Software relates to all programs running/functioning on computers, including operating system software, database management system software, application programs and utility programs (Rob & Coronel 2007:19). According to Amuno (2018), software can be categorised into two main types – system software and application software. System software includes programs that are dedicated to managing the computer itself, such as the operating system, file management utilities and disc operating system. Application software/productivity programs and end-user programs are software that enables the user to complete tasks, such as creating documents, conducting online searching, etc. There is also software called *malicious software* or *malware*. This software is designed to damage and interrupt computers deliberately and is to be avoided at all costs (Tanase, 2014). Software is required to provide the interface between the user and hardware through graphical user interfaces (GUIs). It also manages and allocates memory space for applications and processes the management of applications, input/output devices and instructions (Amuno 2018). Some software is designed to configure and manage internal and peripheral devices and single or multi-user storage in local and network computers; to manage files and applications; monitor system performance; and to produce error messages and troubleshooting options. Software is also designed to function as an interface for network communication and to manage single or multi-user systems (Amuno 2018).

The CDDIS and UNAVCO GSAC systems use commercially available relational database management system software to manage their databases (Noll 2017), for example DSpace and Dataworks for GNSS (Wier, Rost & Boler 2016), discussed in more detail in Chapter 6.

3.4.3 Data

All types of data collected or generated by an instrument or system should be stored for future retrieval. Data types include:

- Observational data (captured in real-time and irreplaceable);
- Experimental data (produced by laboratory equipment);
- Simulated data (generated from test models);
- Derived or compiled data (data mining); and
- Reference or canonical data (conglomeration or collection of smaller peer-reviewed datasets) (Tenopir, Birch & Allard 2012; Fox & Harris 2013: WDS3).

Technique-specific geodetic instrumentation generates space geodesy data (i.e. observational VLBI, GNSS, SLR and DORIS data). Some raw geodesy data is pre-processed to form data products and others kept in their raw form for other scientific applications (Noll 2013; Coetzer et al. 2015).

3.4.4 Processes and procedures

Processes and procedures relate to critical components of a system that govern the design and functionality of the system. These include operational elements of the system, such as inputs, transformation and outputs, as well as the monitoring and auditing of the operations of a system (Rob & Coronel 2007:19). For example, in the data management system of the CDDIS, the data archive is divided into four structural components, each performing their own functions, such as data deposit, download, operations and supplementary support (Noll 2017). The division by functions allows for effective processing and augmenting of the contents of the CDDIS archive. All processing of incoming files takes place in the operation area within the data management system and metadata is extracted and automatically loaded into a relational database. Noll (2017) indicates that the supplementary support component of the CDDIS is designed to support information; particularly files that summarise the contents (metadata) of the download area.

3.4.5 Human resources

Human resources relate to all users of the system (e.g. system administrators, designers, end-users, etc.) (Harley, Acord, Earl-Novell, Lawrence & King 2010). According to Antell, Bales Foote, Turner and Shults (2014:557) and Perret et al. (2015:13), professionals from different scientific backgrounds, such as scientists, information specialists and IT specialists, can work together in designing and developing data management tools and systems. Holl (2010:202) explains that collaborative efforts between experts are required to set up observatory organisational structures inclusive of data management activities. Examples of systems that exist due to fruitful collaboration are those of the Current Research Information System of the Istituto Nazionale di Astrofisca (CRIS-INAF), the Centre de Données astronomiques de Strasbourg Astronomical Data Centre (CDS), the National Radio Astronomy Observatory (NRAO), the European Southern Observatory (ESO) and the CDDIS (Konomi & Marra 2015:38; Perret et al. 2015:13).

3.5 Data structuring and organisation

A data structure involves a particular way of organising and storing data in a computer (Cormen, Leiserson, Rivest & Stein 2009). Data structures provide ways for collecting and organising data to support data usage (Davis 2016). The goal of data structuring and organisation is to create the necessary technical environment to assist users in identifying, retrieving, extracting and using data (University of Washington Libraries 2014). Data structuring and organisation are important and should be considered during the design phase of a new system. Without proper data structuring and organisation, data management activities and systems will not be successful.

One of the main functional requirements of the new GRDMS for HartRAO is that the system must be able to archive, store and supply geodetic data effectively to the scientific community. Therefore, the system designers need to decide at the start of the design phase how data will be structured and organised within the system. To obtain an understanding of what constitute data structuring and organisation, the work of Noll (2010), Tenopir, Birch and Allard (2012), Fox and Harris (2013:WDS3), Trauth and Sillmann (2013) and the Boston University Libraries (2015b) was consulted. According to the Boston University Libraries (2015a), research data is generated for different purposes through different processes and can be categorised in different ways, as indicted in Section 3.4.3. The CDDIS and UNAVCO GSAC data management systems follow the IGS standards guidelines for the structuring and organisation of geodetic data. *The Standards for GNSS data Submitted to CDDIS. Version 1.0 of January 2017* provides details of the types and standards of GNSS data to be managed (CDDIS 2017). The standard is applied to keep incoming data streamed/transferred to the CDDIS in formats consistent with those existing in the CDDIS archive, so as to minimise processing and improving archival efficiency (CDDIS 2017).

According to Davis (2016), the management of geodetic data within data management systems may be structured in a variety of ways, which include:

- arrays (fixed-length lists made up of a collection of objects or data values; allow for determining the position of each object or value by using mathematical formulae);
- queues (data structured in a first-in-first-out order);
- stacks (data structured in a last-in-first-out order); and
- trees (data structured in a hierarchical manner, consisting of one or more data nodes; the first node is called the root/parent; each node can consist of zero or more sub/child nodes).

According to Trauth and Sillmann (2013), data is generally stored in hierarchical order in directories. This provides for an effective, flexible and user-friendly data management system. Directories can usually be identified by a unique directory name. Files of the same nature are usually stored under the same directory. Directories can be classified as root directories, which contain subdirectories/parent directories or workspace directories. Workspace directories store data files relating to a particular project. Within these directories, subdirectories store data in various folders, subfolders and data files (Trauth & Sillmann 2013). Data files are created by using specifying paths. An access path defines the physical file and/or the logical views of that file (CA Technologies 2018). Access paths are made up of device name (database, disc, etc.), one or more directory names and file names. When constructing access paths, designers normally specify the order in which the data record needs to be stored – i.e. the file, the fields and the select/omit criteria for deciding which records from the files will be retrieved by the access path (CA Technologies, 2018).

A typical example of an access path for GPS data hosted by the CDDIS is as follows: *Root:ftp://cddis.gsfc.nasa.gov/pub/gps/data/daily/year/day/type* (Noll 2010). The CDDIS (2015) and Noll (2010) indicate that geodesy VLBI, SLR and DORIS data is stored in subdirectories in format type SINEX. GNSS data is stored in the format type RINEX. The files are available in different formats designed for use by different processing packages and binary databases. An example of the hierarchical structure for a particular type of technique-specific data hosted by the CDDIS is illustrated in Figure 3.2.

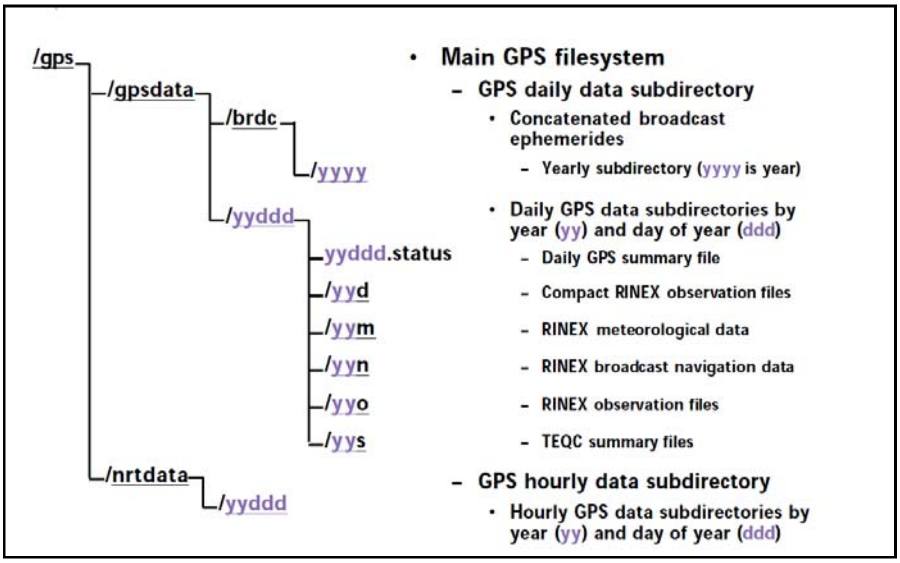

Figure 3.2: Directory structure for GPS data hosted by the CDDIS

(Noll 1999)

As reflected in Figure 3.2, GPS data is stored in a directory in the *Main GPS filesystem*, */gps*. The main directory is divided into subdirectories */gpsdata* and */nrtdata* containing daily and hourly GPS data, respectively. The subdirectory for daily GPS data is further divided into subdirectories for broadcast ephemeris (*/brdc*) and daily GPS data by year and day of year (*/yyddd*). The subdirectory for hourly GPS data is divided into data by year and day of year (*/yyd*). The broadcast ephemeris subdirectory contains a further 'Yearly subdirectory', while the daily GPS data by year and day of year subdirectory contains several further subdirectories related to summary and RINEX files.

Once data and files, as indicated in Figure 3.2, have been created, gathered or manipulated, they can very easily become disorganised (University of Cambridge 2018). To support and enhance the retrieval of data stored in a system, file naming conventions (FNCs), metadata schemas, digital object identifiers (DOIs) and open research and contributor IDs (ORCIDs) can be applied in conjunction with ordinary data structuring techniques to structure and organise data. The following sections summarise the information gathered from the literature pertaining to retrieval methods to enhance data access.

3.5.1 File naming conventions

File naming conventions (FNCs) are frameworks for naming files in a manner that delineate their content and the way in which they relate to other files (Purdue University Libraries 2017). FMCs are developed through a process of identifying the key elements of, for example, a research project and the important differences and commonalities between files. Key elements include aspects such as date of creation, author's name, project name and the version of the file (Purdue University Libraries 2017).

Consistent and descriptive file naming conventions serve different purposes, often related to information and data management and the usability of data (Antin 2016). It is essential to establish a FNC at the onset of data collection to prevent against encountering a backlog of unorganised files that may result in retrieval problems, misplacement or loss (Illinois University Library 2018). File names can be meaningful or non-descriptive. The University of Leicester (2015) advises that users benefit form secure storage of data and the ability to locate and access the data without difficulty where logical standards are applied consistently. Meaningful file names consist of numbers (roman numerals) and words or abbreviations that have some relation to the item being captured. They may be an abbreviated title, the accession number of the physical item, or some other descriptive identifier.

Meaningful file names work best for structuring and organising medium to small collections of digital files. According to University Libraries of Colorado (2008), meaningful file names are not ideal for large collections, because it is difficult to create unique names for all items. In large data collections, the meaning of the name may be lost or it may change connotation over time. Non-descriptive file names are usually sequential numbers and they work well for medium to large collections, as they are easy to assign. It should be kept in mind that non-descriptive file names provide no identifying information. Therefore, the files are harder to manage outside of the database that contains the associated metadata. The decision to use meaningful or non-descriptive file names should be based on the characteristics of the collection and the project specifications.

In space geodesy, FNCs are used to identify observations conducted at the various network stations; to identify the participating network stations; to identify the file format; and to provide enough information in the file name, so that data files can be transformed from one file format to another with no additional information (VLBI.org 2009). An example of an FNC used by geodesy data service providers such as the IERS and the CDDIS is the file naming convention *8.3.Z FNC* (Federal Agency for Cartography and Geodesy 2013). The *8.3.Z FNC* is synonymous with short file names and is used for SINEX and RINEX Version 2 and 3 data description.

An example of the access path for VLBI data hosted by the CDDIS is: *ftp://cddis.gsfc.nasa.gov/vlbi/ivsdata/* followed by the subdirectories and file name (as per *FNC 8.3.Z*): *db/yyyy/yymonddbb_v###.gz* where *db* indicates the directory, *yyyy/yymonddbb_v###* the file name and *.gz* the compression format (CDDIS 2015).

Figure 3.3 shows the *8.3.Z FNC* file naming definitions used by the CDDIS for RINEX Version 2 data.

Code	Description
DD	Two-digit day of month (01, 02, …31)
DDD	Three-digit day-of-year (001, 002, …366)
HH	Hour of day (00, 01, …24)
MM	Two-digit month (01, 02, …12)
SATNAME	Satellite name (legeos1, gracea, etc.)
SSSS	Four-digit station identification number
T	technique-specific file type (M =meteorological)
WWWW	Four-digit GPS week number (0649, 0650, …)
YY	Last two-digits of year (76, 77, …)
YYYY	Four-digit year (2018, 2019, …2025)
#	Sequence / version number

Figure 3.3: Code and description for naming CDDIS technique-specific data files

(CDDIS 2015)

For example, making use of the *Code* corresponding to the *Description*, an IGS ionosphere total electron content product file would be given as: */gps/products/rinex/YYYY/DDD* where the *YYYY* represents the actual year (e.g. *2018*) and *DDD* a three-digit day of year, e.g. *270*. In this way, the actual file will appear as: */gps/products/rinex/2018/270.* The file naming conventions for RINEX Version 3 technique-specific data is discussed in Chapter 6.

3.5.2 Metadata schemas

Metadata schemas are logical plans that indicate the relationship between metadata elements, normally through establishing rules for the use and management of metadata – specifically with regards to semantics, syntax and optionality of data values (National Information Standards Organisation 2006). Metadata standards are uniform sets of ground rules for tagging information (Higgins 2007). They are high-level documents that include principles for implementation, such as agreeing on language, spelling and date format for metadata (UNC University Libraries 2017). They describe the way in which the metadata is set up and usually address standards for common components of metadata, such as dates, names, and places.

Linked to data structuring and organisation, Riley (2017) found that metadata schemas are used to support data structuring and organisation within data management systems and, if properly applied, can enhance and ensure data retrievability There are different types of metadata schemas, such as descriptive, structural and administrative metadata schemas. Using any of these three schemas in data organisation provides flexibility to allow for the description of data at any level of aggregation. Metadata schemas also make it possible to search across networks (Riley 2017). Formal metadata standards enable interoperability among similarly formatted databases on local and global scales, which greatly enhances the sharing of scientific research.

According to Hogrefe & Stocks (2011), data collected in accordance with the quality and organisational guidelines set out in established metadata standards are not only more easily retrieved but can also be more easily shared and funded (UNC University Libraries 2017). Metadata schemas also provide direction for the documentation of data processing; file naming conventions and formats; and a glossary of precise definitions for applicable terms. Various metadata schemas and standards exist to assist with data discovery in a variety of scientific disciplines, e.g. *Dublin Core,* a general standard that can be adapted for different disciplines; *NASA Standards* used by NASA Earth Science Data Systems; and the *Data Document Initiative Alliance* metadata schema for social, economic and behavioural sciences.

A popular schema used in Astronomy is that of the International Astronomical Union (IAU) Astronomy Visualization Metadata (AVM) standard (Noll 2015a:2). In geodesy, there appears to be a knowledge gap in metadata standards for geodetic data. According to Noll (2015a:2), there are no current standards in place that aid in the discovery of geodetic data and products; nor is there a machine-readable standard that allows for the efficient automated transfer of geodetic metadata. The schemas available can only assist with basic metadata retrieval or transfer. There is a need for a new standard that will cover the sharing and collation of large volumes of diverse geodetic data and metadata captured and stored by any number of custodians using proprietary software (Noll 2015a:2).

3.5.3 Digital object identifiers (DOIs)

Digital object identifiers (DOIs) are unique alphanumeric identifiers associated with a specific piece of intellectual property, particularly one presented in an online environment (Sciencemag.com 2017). These identifiers do not specify the location of an online object, but rather its content. Therefore, a DOI is a "persistent" identifier and remains associated with the object, irrespective of changes in the object's Web address (Sciencemag.com, 2017). An example of a DOI for accessing registered research datasets for an earthquake event and its accompanying metadata is as follows: *doi:10.1594/GFZ.GEOFON.gfz2009kciu.* The location of a dataset may be indicated with the use of a specific file name/uniform resource name (URN), a uniform resource locator (URL), uniform resource identifier (URI) or a DOI. However, the use of persistent identifiers remains the preferred method for location specification, as objects have a tendency to change locations, making the assigned standard URL invalid and the data irretrievable.

Persistent identifiers are used in various scientific disciplines to assist with organising and structuring data. In geodesy, DOIs are becoming increasingly important for the retrieval of large volumes of dispersed data. In recent years, the CDDIS have implemented DOIs to assist in the selection of datasets for technique-specific data and products (Noll 2016:136). These identifiers provide easy, effective access to the CDDIS data holdings and assist scientists in citing data holdings in their publications.

3.5.4 Open research and contributor ID (ORCID)

Open research and contributor identifier (ORCID) is a non-proprietary alphanumeric code used to uniquely identify academic and scientific authors and contributors (Butler, 2012). This is a subset of the International Standard Name Identifier (ISNI) under the auspices of the International Organization for Standardization (ISO) (ISNI 2017). These identifiers aim to attribute research outputs reliably to their true author or originator by assigning a machine-readable, 16-digit unique digital identifier to each author or originator. An author’s ORCID may appear as follows on the ORCID platform: *0000-0002-1825-0097* and *https://orcid.org/0000-0002-1825-0097* (Open Research and Contributor ID 2017). Researchers are connected to their research by embedding unique IDs in manuscript submissions, grant and patent applications (Open Research and Contributor ID 2017). Including or creating an ORCID at the time of submission enables authors to track and distinguish their research output easily. This creates opportunities for new synergies between different interest groups and a growing interest in interoperation (ISNI 2017).

Datasets are an important scientific output and a vital resource for researchers of all scientific disciplines. These identifiers allow for tying datasets to publications. For the scientific community and data service providers to access and re-use datasets effectively, an understanding of data storage and attribution is fundamental. From this knowledge and understanding flow the development and use of standard data exchange protocols.

According to Bryant (2013), scientists of today have numerous options for storing their data. Most research universities support data management activities, including the publication and archival of data through an institutional data repository. These identifiers are embedded in data repository workflows. By capturing a wide variety of research objects, including figures, tables, spreadsheets and flat files in a repository, many scientific community platforms make it easy for individuals, scholarly societies, publishers, research institutions and funding agencies to promote data preservation and re-use. It is becoming increasingly important and possible to manage attribution with the assignment of DOIs to datasets. The assignment of DOIs for datasets signals an increased acceptance of research data as an essential and legitimate part of the research record.

The literature of Rob & Coronel (2007), Noll (2010), CDDIS (2015) and Thakur (2016) provides several data management system characteristics relevant to the research topic, which should be considered for the design of the GRDMS for HartRAO. These include characteristics such as (but not limited to) having an application interface; the ability to backup and recovery data; providing multi-user remote access, etc. The literature of Nelson (2009), Martin and Ballard (2010:1), Tenopir, Birch and Allard (2012), Behrend (2013), Fox and Harris (2013:WDS3), Science HQ (2013), Valle (2013), Ray (2014), CDDIS (2015), ZDNet (2015), Wier, Rost & Boler (2016) and Thakur (2016) provides information on the criteria that should be considered for a DMS. Key criteria to be considered and applied when designing the GRDMS include (not limited to): support of user registration and management of accounts; software that receives and interprets user requests; structuring and organisation of data for easy access; management of all files, even at remote sites; ensuring database consistency, security and integrity; storage of large volumes of geodetic data indefinitely; and conducting regular backups.

The foregoing is only some of the characteristics and criteria to be considered for the design of the GRDMS. Additional key characteristics and criteria are discussed in Chapter 6.

According to Noll (2010), Austin (2016) and LIMSWiki (2016), a DMS consists of components such as hardware, software, data, processes, procedures and human resources. Various authors indicate that data should be structured in arrays, queues, stacks and trees. For the design of the GRDMS, it is important to follow a hierarchical data structure, as this will contribute towards an effective, flexible and user-friendly data management system. Noll (2010) indicates the access paths and directory structures used at the CDDIS to store SINEX and RINEX data. For the enhancement of data retrieval, the literature of the University Libraries of Colorado (2008), the Federal Agency for Cartography and Geodesy (2013), CDDIS (2015), Hogrefe and Stocks (2011), Noll (2015a:2 & 2016), Sciencemag.com (2017) and Bryant (2013) provided important information relevant to data structuring and organising tools, such as DOIs, FNCs, metadata schemas and ORCIDs, which needs to be considered in the design of the new GRDMS.

3.6 Conclusion

In this chapter, important literature relevant to the research topic was discussed, including information on data management system characteristics and criteria, data management system components, as well as data structuring and organisation. Secondary and tertiary literature was retrieved through bibliographic searches in online databases, repositories and websites. Hofstee's (2006) literature review funnel method was followed to distil the review results. Key characteristics, criteria and components, data structuring and organisation and retrieval tools were discussed. It was evident from the literature consulted that a gap exists in the literature regarding the design of a data management system specifically for geodetic data generated by HartRAO instrumentation. Very little has been reported and documented. This necessitates the development of a new conceptual framework.

The information in this chapter is supported by the results of the dissertation, which are discussed in Chapters 5 and 6. The research design and methodological approach followed in this study is described in the next chapter.

CHAPTER 4: RESEARCH METHODOLOGY

4.1 Introduction

When researchers investigate a particular phenomenon, they try to find solutions to a particular problem, or they try to answer a specific question (Davis 2014a:10; Du Plooy-Cilliers 2014:19). Chapter 4 focuses on the research paradigm, approach, method, data collection and tools applicable to finding solutions towards developing a new GRDMS for HartRAO. The validity and reliability as well as ethics of the research are also treated in this chapter. The aim of the chapter is to provide detailed information on the methodological approach followed during the execution of the research.

Research methodology is an important component of research to enhance understanding of the wider field of discussions related to theories, approaches, methods and inter-relationships required to ensure the effective execution of the actual research. Du Plooy-Cilliers (2014:19) indicates that methodology is the way in which one collects and analyses data.

To describe the research process followed in this study, the metaphor of the *research onion* of Saunders, Lewis and Thornhill (2009) will be used to identify the core of the research (the central part of the research onion) in relation to other design elements (the outer layers of the research onion). Figure 4.1 provides a summary of the research design and methodology applicable to this research, in line with the research onion proposed by Saunders, Lewis and Thornhill (2009).

When viewed from the outside, each layer of the ‘onion’ describes stages in the research process, starting with the research philosophy, followed by the approach, strategy, data collection and analysis.

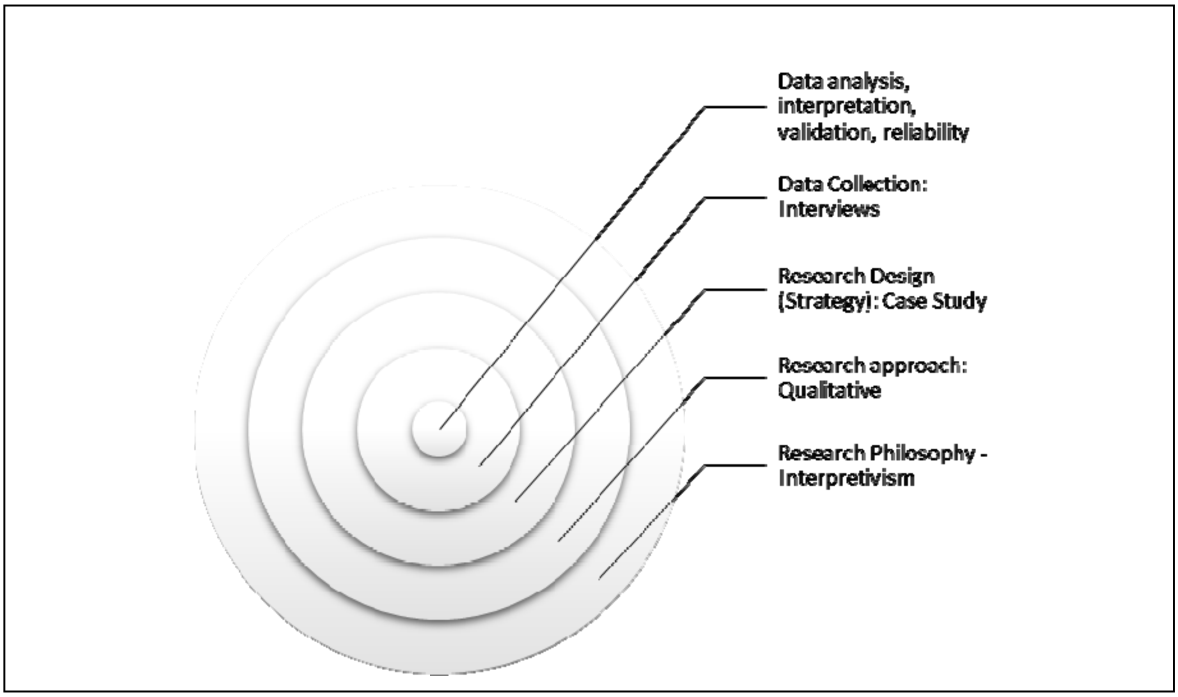

Figure 4.1: Research design aligned to the 'research onion'

(Adapted from Saunders, Lewis and Thornhill 2009:108)

4.2 Research philosophy

The first layer of the research onion – and the most critical one – is the research philosophy. A research philosophy is a belief or an idea about the collection, interpretation and analysis of data (Bryman 2012). According to Flick (2011), the assumptions created by a research philosophy provide the justification for the way in which the research will be conducted. Research philosophies can differ with regard to the aims and goals of the research and the way in which to achieve these goals (Goddard & Melville, 2004). In social science, the research philosophy or paradigm is often referred to as the research tradition or worldview (in other words, what we think about the world we live in) (Du Plooy-Cilliers 2014:19).

The research philosophy applied in this study is that of an interpretivist worldview. Nieuwenhuis (2007:58) asserts that interpretivism is practical and that the aim is not merely to accumulate knowledge for the sake of knowing, but for understanding a phenomenon in-depth. The motivation for the choice of philosophical position taken in this study is based on the interpretivist's belief that reality is interpreted or constructed by people. Because interpretivists believe that truth is dependent on people's interpretation of facts, interpretivists are not interested in generalising their results (Du Plooy-Cilliers 2014:29). In seeking the answers for research questions, the researcher following an interpretive paradigm uses those experiences to construct and interpret an understanding from collected data (Thanh & Thanh 2015:24). The research methodologies employed by interpretivists are sensitive to the specific context and never generalised beyond the context in which the study has been conducted. According to Du Plooy-Cilliers (2014:21–30), interpretivists collect information, analyse and interpret it, after which they formulate a theory, based on the information Willis (2007a). Researchers are often required to spend time in direct contact with the subject being studied, in order to gain a thorough understanding of what is meaningful and relevant (Du Plooy-Cilliers 2014:28).

The methodological position of interpretivists is also relevant to this study – facts are fluid and embedded within a meaning system. Facts are fragile and change as people's experience and perception change. According to interpretivists, facts are not objective and neutral. Instead, what is factual depends heavily on the context of people's interpretation of information (epistemology) (Nieuwenhuis 2007:59–60).

Another position of interpretivism, which is of particular significance to this study, is the view that social reality is in a constant state of flux and that, depending on circumstances, cultures and experiences, people may experience reality in different ways. Interpretivists value the complex understanding of multiple realities and do not attempt to conduct value-free research (ontological view). Instead, they discuss the values that shape their research, including their own interpretations and those of others (Du Plooy-Cilliers 2014:28). Since the aim of interpretivists is to obtain an in-depth understanding of multiple realities, they depend on qualitative research (Nieuwenhuis 2007:50). Researchers, who are using interpretivist paradigms and qualitative research methods, often seek understanding and perceptions of individuals for their data, so as to uncover reality, rather than rely on statistics (quantitative research methods).

4.3 Research approach

The second layer of the research onion of Saunders, Lewis and Thornhill (2009) relates to the research approach. Creswell (2012) and Datt (2017) describe the research approach as a plan and procedure that consist of broad assumptions leading to detailed methods of data collection, analysis and interpretation. The research approach is based on the nature of the research problem being addressed. In support of this view, authors such as Willis (2007b) and Silverman (2013) believe that the interpretivist paradigm is applicable when research is conducted to explore richness and depth of a specific research topic.

A qualitative research approach was followed in this study, with the aim of obtaining data and key information to be considered when developing a contextual framework for the design of HartRAO's new GRDMS. The qualitative approach involves reasoning from general to more specific assumptions. It allows the progression from a general (major) to a narrower (minor) premise, to a specific conclusion or claim (Davis 2014b:121).

The researcher moves from the general to the specific in several different cycles of analysis and interpretation (Bezuidenhout & Cronje 2014:234). In this study, the researcher moves towards exploring inherent structures and mechanisms that may have a positive influence on the accessibility of geodetic data. In support of the choice of research approach, according to Du Plooy-Cilliers (2014), qualitative research can be centred on one or more purposes, namely to obtain information about a topic that has not been researched before.

4.4 Research design (strategy): Case study

Aligned to a qualitative approach, Saunders et al. (2009) indicate that researchers can use one or more strategies within their research design to address a research problem and answer a research question. Research design is a strategic framework for action that serves as a bridge between research questions and the execution of the research (Terre Blance, Durreheim & Painter 2006:34). Linked to the view of Nieuwenhuis (2007:75), a case study is an empirical enquiry that investigates a contemporary phenomenon within its real-life context, when the boundaries between phenomenon and context are not clearly evident. Case study research relies on multiple sources of evidence; is multi-faceted; and can be used in a variety of ways.

The aim of this study is to obtain data and key information to be considered when developing a contextual framework for the design of a geodetic data management system. The case study research approach proved to be the most suitable approach for solving the research problem set for this study. The HartRAO case forms the basis of the study, as the new geodetic data management system is designed to cater for HartRAO's geodetic data service requirements.

The case study research is well-suited to provide an understanding of the interactions between information technologies, related innovations and organisational context (Williamson & Johanson 2013). Also, case study designs are useful when extensive exploration of an area of which little is known is involved, or when a holistic understanding of the situation and/or a phenomenon is required. Case study research is associated with description and theory development and can be used for the exploration of areas in which existing knowledge is limited. Acroyd (2010) observes that the case study is the best approach to use for exploring the interaction of structure, events, actions and context to identify and illustrate causal mechanisms.

Case study research is flexible and can be employed in various ways within different philosophical paradigms. According to Williamson and Johanson (2013), case study research is most suitable for investigating the design, development, implementation and use of information systems within organisations. However, case study research does not claim to make any generalisations to a population beyond cases similar to the one being studied (Wynn & Williams 2008).

The case study process begins with the *a priori* identification of constructs (or concepts) from the literature, which guides the research process. This is followed by the formulation of the research question(s) (Williamson & Johanson 2013). Emergent themes (from fieldwork) are then compared and contrasted with the literature. Theory and data is systematically compared and contrasted through a number of iterations, until the theory accurately reflects the data. Pattern matching is used to compare empirical data collected from the case study to outcomes predicted by the propositions.

An advantage of case studies is that they can be combined with other approaches. Gable (1994), for example, first used case study research to define constructs and to develop theory, which was subsequently tested by utilising survey research. The researcher of this study followed a case study strategy in combination with a literature review and one-on-one interviews, in an attempt to understand a particular phenomenon from a wider population. According to Saunders, Lewis and Thornhill (2009), the case study strategy is useful for generating answers to the 'what', 'why' and 'how' questions of the research.

In this study, the researcher seeks answers to the following questions:

- Which conceptual models are suitable and should be considered in the design of the GRDMS for HartRAO?
- What components, characteristics and criteria, data structuring and organisation should be considered when designing the GRDMS?
- To what extent does the design of the GRDMS for HartRAO comply with technological expansions and institutional needs?

The case study offers a multi-perspective of one or two participants in a situation, but also provides the views of other relevant groups of role players and the interaction between them (Nieuwenhuis 2007:75). This was particularly significant for this research, because it allows for a deeper understanding of the dynamics of the situation. The case study strategy followed in this research was developed in conjunction with the literature review. Exploring the sources assisted in the careful selection of examples of research data management systems, components, characteristics and criteria, as well as data structuring and organisation for such DMS, which helped to establish necessary boundaries related to the design of a GRDMS for HartRAO.

Multiple data management system resources in Geodesy were explored, with the aim comparing and extracting the important constructs necessary for the design of the GRDMS for HartRAO. Examples of some of these sources consulted include that of the

CDDIS data management system, the UNAVCO GSAC software system and the GNSS data management system model by Mashaba et al. (2016). Each example was treated individually and each of the examples related to a geodetic data management system. The information obtained from each example contributed to the identification of key components, characteristics and criteria, as well as data structuring and organisation to be considered in the design of the proposed GRDMS.

4.5 Data collection

Case study research generally includes the use of multiple data collection techniques via the inquiry of multiple sources (Williamson & Johanson 2013:181). The use of multiple data collection techniques and multiple sources strengthens the validity and credibility of outcomes. This enables different interpretations and meanings that can be included in data analysis. Linked to the research problem set for this study, data were collected by means of an in-depth literature review and one-on-one interviews. Primary, secondary and tertiary sources were consulted, in order to obtain a deep objective understanding of the phenomenon.

4.5.1 Literature review

A comprehensive literature review of the body of knowledge regarding the design of a data management system was conducted. It is not suggested that the literature review can be exhaustive towards a totality of literature on any given subject. A literature review should be comprehensive inasmuch as it should involve the use of rigorous techniques to search, retrieve and collect resources (Onwuegbuzie, Frels & Hwang 2016). In this study, searches were conducted whereby primary, secondary and tertiary literature (i.e. peer-reviewed journal articles, conference proceedings, books and grey literature in paper and electronic format) were retrieved through various bibliographic searches in online databases, repositories and on websites by using Hofstee's (2006) literature review funnel method (Chapter 3: Figure 3.1).

The researcher initiated the literature review by starting with a broad theory base (i.e. works that are relevant to the study but that do not address the research problem). As the review process unfolded, the baseline between broad and narrow (works relevant to an aspect of the study) theory bases were systematically condensed (selecting, simplifying, abstracting and/or transforming data that appear in the full scope of resources (Miles, Huberman & Saldaña 2014:12). Experts in the field of geodetic data management were consulted to identify additional literature. This provided supporting information on characteristics and criteria of data management systems, design components and organisation and structures to be considered in the design of the new GRDMS for the effective and efficient management of different datasets.

The literature searches yielded approximately a hundred salient articles. The research topic was approached from a broad perspective (less relevant information) and narrowed down to the most relevant information. The literature was interrogated to identify themes, which formed the basis for data analysis. The data collected were systematically organised and sorted according to themes and concepts to allow converging lines of inquiry and patterns to be discovered.

The researcher examined the data by employing various interpretations, in order to find linkages between the research objectives and the outcomes with reference to the original research questions and objectives. During the familiarisation process, *a priori* themes and categories on the research topic were identified by distilling it from the literature, which formed the basis for subsequent interviews and data analysis. Themes and categories relevant to the research topic were selected from the literature consulted and discussed under appropriate subheadings in Chapter 3. Themes and categories identified included: the management of scientific data (components, characteristics and criteria, as well as data structuring and organisation); scientific data management systems used for geodesy data; standardisation of data and data management system design (conceptual/architectural design models). These key themes and categories formed the foundation for the development of a coding scheme (discussed in Chapter 4: Sections 4.6.3–4.6.7).

Regarding DMS components, characteristics and criteria and structuring and organisation, literature from the narrow theory base was selected. The researcher identified components, characteristics and criteria and data structuring and organisation most often mentioned in the literature relating to DMS and evaluated the relevance to geodetic data management systems. To present the components, characteristics and criteria, data structuring and organisation themes were identified from the literature, two matrixes were compiled (see Chapter 3: Table 3.1 and Table 3.2).

Linked to Hofstee's (2006) funnel model (discussed in Chapter 3: Section 3.2), a considerable amount of literature from both the broad and narrow theory bases was consulted to obtain information relevant to data management system components and data structure and organisation. The key DMS components identified included hardware, software, data, processes and procedures and human resources. Each of these components was discussed in Chapter 3 (see Chapter 3: Sections 3.4.1–3.4.4).

Information pertaining to data structure and organisation, distilled from the literature consulted, included aspects such as data structures (e.g. arrays or hierarchical structures, directories, subdirectories, folders, subfolders and data files), FNCs, metadata schemas, DOIs and ORCIDs. These aspects were discussed in Chapter 3 (see Chapter 3: Sections 3.5.1–3.5.4). Because the funnel model was explained and applied in the literature review chapter, details regarding the data collection process, focusing on the interview method, are provided in the section to follow.

4.5.2 Interview method of data collection

Interview strategies are used in qualitative research to obtain the opinions, perspectives and attitudes of respondents towards an issue, product, system, programme or service (Kumar 1999). Interviews are valuable sources of information that allow for the interpretation, understanding and meaning of participants' responses (Strydom & Bezuidenhout 2014:188). The interview also allows the researcher to ask for clarification of a point of view and a more detailed explanation which, in turn, allows for more flexibility in the research process.

According to aforementioned authors, there are three types of interviews – informal conversational interviews (no specific predetermined questions are asked, allowing the interview to progress naturally as a conversation); general interviews (following a conversational style although certain themes are covered by asking predetermined questions) ;and standardised open-ended interviews (that focus on asking the same set of open-ended questions to all the participants and allow for comparison of views and opinions of the participants in an organised manner).

In this research, a combined informal conversational and general interview approach was followed. The aim was to solicit as much information as possible from the participants without making them feel threatened or uncomfortable. The interviews started with an informal discussion that led to the general interview approach of asking relevant questions. An interview schedule/interview guide containing sets of open and closed-ended questions was used to guide the interview process (see Appendix A).

The aim of the open-ended questions was to encourage the participants to provide full and meaningful answers. The open-ended questions were phrased as statements that required responses. The responses were then compared to information already known to the interviewer. The aim of the closed-ended questions was to encourage a short or single-word answer and not to lead the participants in answering the question. Interview questions were adjusted to focus the interview. The interviews lasted for approximately an hour and were based on twenty-five questions set relevant to the research topic.

The responses were drafted while the participants were being interviewed. Depending on the responses from the interviewees, the researcher probed further in an attempt to enhance the richness of the data collected. The responses were later transcribed by means of word processing software. To strengthen the credibility and transcription of the responses, the participants received a copy of the questions and responses and were asked to comment on the validity of the transcriptions. Where errors were observed, corrections were made in consultation with the relevant participants. Each participant received a final copy of the transcribed document.

4.6 Target population and sampling method

To ensure that the data collected via interviews applied to the aim and objectives of the research, a target population that could provide insight into the design of a data management system in a geodetic environment had to be selected. The target population consists of all units, people or things, possessing the attributes or characteristics of interest to the researcher (Keyton 2011:121). The shared attributes or characteristics and the number of people in a population are referred to as the *population parameters* of the study (Pascoe 2014). The population parameters of this study related to the nature of the population (scientists in the research field of space geodesy); the population size (scientists employed by the NRF and actively participating in geodetic research at HartRAO); and the unique characteristics of the population (geodetic scientists and students). The target population identified for the study included geodesists, Master's and doctoral students and post-doctoral scholars at HartRAO.

Once the population of the research was determined, the next was to select a sample from the identified target population. The sample selection was based on the ability of those selected to provide data relevant to the study. Within this context, the principle of saturation was applied. Saturation of sample selection occurs when the selection process reaches a point where there was nothing or nobody else to select. To determine the number of participants to include in the sample size (up to the point of achieving saturation), Pascoe (2014:136) indicates that probability sampling should be applied. *Probability sampling* refers to whether or not each element/unit in the population has an equal opportunity to be part of the sample.

Within this sampling method, a non-random method of selecting participants for this research was deliberately chosen, because the aim was not to select a random sample from the population, but to select participants that could provide as much information as possible, so as to understand the phenomenon in its totality (Kumar 1999). In other words, the sample group was selected to provide as much detail as possible on the design of a data management system and is, therefore, not necessarily representative of each of the groups of individuals that may be involved in the design, development and utilisation of a new GRMDS for HartRAO. The selected sample consisted of a total of ten participants (research geodesists, doctoral and Master's students) specialising and studying in the field of space geodesy. The researcher was confident that the modest number of participants selected could provide a rich enough source of data required to address the research questions in full. A breakdown of the number of participants per group is presented in Table 4.1.

Table 4.1: Participant numbers

Participant type	Number
Research geodesists	3
Doctoral and Masters students	7

4.7 Data analysis

Qualitative data analysis is the process of bringing order, structure and meaning to the mass of collected data. Data analysis typically involves the process of reducing the volume of raw data; sifting significance from trivia; identifying significant patterns; and constructing a framework for reporting the essence of what the data reveals (De Vos et al. 2011).

In case study research, qualitative data analysis is usually highly interactive (Williamson & Johanson 2013:181). The unit of analysis in case study research is, for example, the phenomenon, case and/or situation that being studied. Miles and Huberman (2014) define a *case* as a phenomenon of some sort that occurs in bounded context. The case is, in effect, the unit of analysis. In information systems, a case may be a potential user, a management practice, or a particular technology/system (Williamson & Johanson 2013). The case identified for this study is that of HartRAO's geodetic data management.

According to Bezuidenhout & Cronje (2014:233), there are several qualitative data analysis methods available, namely content/textual, discourse, conversation, multimodal conversation and semiotic analysis. For the purpose of this study, qualitative content/textual data analysis was conducted. The aim was to provide a detailed description of the reality mirrored in the text and interviews. When content/textual analysis is used as a method to analyse text, it is used for both the content of text and for all transcribed data.

The most sensible manner of discussing the process of data analysis is to present it systematically and in the form of sequential steps. The choice of steps to follow will be determined by the aim of the study. Hence, the research goal and research questions will determine the choice of steps taken in the analysis of the data (Bezuidenhout & Cronje 2014:233). The steps followed in this study are presented in Figure 4.2.

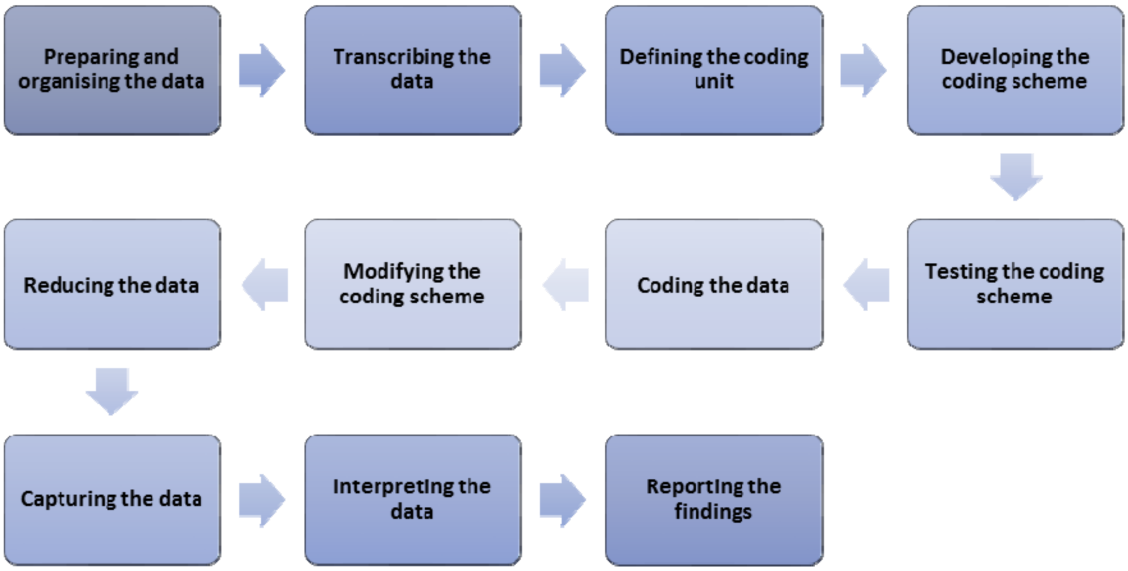

Figure 4.2: Data analysis process

4.7.1 Data preparation and organisation

The first step in this process was the compilation of a master list of all the participants. The master list consisted of the initials, surname, title, job title, interview status and the date of the interview. This metadata is indispensable and strengthens the credibility and validity of the research. The master list serves as proof of the participant's existence and participation in the study.

After the master list had been compiled, each interviewee was given an alphanumeric code that served as individual identification. For example, the first participant's code was *P1*, the second *P2*, etc. Alphanumeric codes were assigned to protect the participants' anonymity.

4.7.2 Data transcription

The next step in the data analysis process was the transcription of the interviews, which is a way of preparing the data for analysis. Data transcription was accomplished with the use of computer word processing software. The raw data collected from each interview were transcribed. The researcher decided to include the full range of responses to prevent the omission of relevant and significant data.

To increase the trustworthiness of the study and to serve as evidence of the processes followed, the researcher conducted memoing. – Reflective notes (see Figure 4.3) were compiled as a form of memoing to serve as reminders of the researcher's observations, thoughts and insights during the interviews, reading and transcription during data analysis. The reflective notes contain the opinions and views of the researcher.

According to Shuttleworth (2018), memos can provide an outlet for the researcher to contemplate additional data to be collected in order to fully present the phenomenon. Memoing assisted the researcher in making conceptual leaps from raw data collected to abstractions. It also enhanced data exploration; enabled continuity of conception; and facilitated communication throughout the research process. The researcher could refer back to the notes to channel discussions and lead conversations relevant to the research.

4.7.3 Defining the coding unit

According to Bezuidenhout & Cronje (2014:236), this step in the analysis process refers to defining the basic coding unit or text intended for analysis. The coding unit can be entire documents, paragraphs, phrases, sentences or symbols (Bezuidenhout & Cronje 2014:236). The purpose of defining the basic coding unit is to assist the researcher in organising the data into manageable chunks. In other words, by assigning codes to text chunks describing a particular concept, the researcher is grouping and organising the data in a sensible manner that will simplify analysis. After examining the text (data), the researcher decided to use sentences as the coding units.

4.7.4 Developing a coding scheme

The coding scheme was developed deductively – from *a priori* concepts emerging during the familiarisation phase of the research – inductively from the raw data. Most themes and categories already emerged while compiling the theoretical background (see Chapter 2) and the literature review (see Chapter 3). The coding scheme was, therefore, constructed primarily from concepts identified during these stages. Alphabetical codes were assigned to each theme, category, subcategory and sub-subcategory (see Table 4.2 and Figure 4.3) and organised into a hierarchical conceptual framework.

Figure 4.3 depicts the code *DM-SGD-U's* deconstruction and demonstrates the hierarchical structure followed in this study.

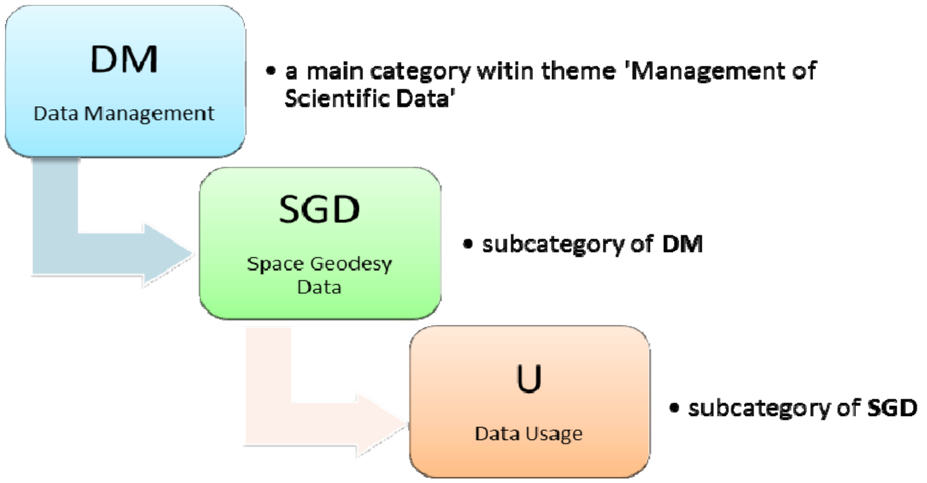

Figure 4.3: Hierarchical conceptual framework

The hierarchical framework for organising codes was useful, as it clearly illustrated the propagation of codes within the coding scheme.

Table 4.2: Themes, categories and coding scheme

Themes	Main category and code	Sub-category and code	Sub-sub category and code
1. Management of scientific data	Data Management (DM)	Space Geodesy Data (DM-SGD)	Data Usage (U)
			Data Type (T)
	Data Management Systems (DMS)	Design (DMS-D)	–
		Internet (DMS-I)	Search (S)
			Interface (In)
			Access (A)
			Retrieval (R)
2. Scientific data management systems	Research Data Management Systems (RDMS)	Components (RDMS-Ca)	–
		Characteristics (RDMS - Cb)	
		Criteria (RDMS-Cc)	
3. Data management systems used for geodetic data	Geodetic Data Management System (GDMS)	–	–
4. Data standardisation	Data Structure and Organisation (DSO)	Digital Object Identifier (DSO-DOI) and File Naming Convention (DSO-FNC)	–
5. Data management system architecture	Conceptual Models (CM)	–	–

The text (data) was, therefore, assigned these “combined” codes, depending on the theme, category and sometimes subcategory and sub-subcategory to which it belonged. Where text (data) belonged to more than one theme, category, subcategory and/or sub-subcategory, an ampersand logogram (‘&’) was used.

4.7.5 Code scheme testing

This step involved testing the consistency of the themes, categories and the coding scheme by using a sample of the data collected during the interviews. To test the coding scheme, the researcher extracted random specimen from the data (i.e.

sentences from the interview responses); matched concepts in the text to corresponding themes, categories, etc.; and assigned the appropriate codes.

If the consistency level was low, the researcher 'redefined' the coding definitions and conducted further tests to assess consistency (i.e. applied an error correction method). These steps were repeated, until the results showed a high level of consistency.

4.7.6 Coding the data

Coding the data is the process of applying a thematic framework to the data by using numerical or textual codes to identify specific pieces of data corresponding to various themes (Lacy & Luff 2007:14). In preparation for the text-based qualitative coding and analysis of the data, the researcher used word processing software in the transcription of the interviews. The researcher studied the transcripts thoroughly to determine dominant trends and patterns and to gain an overall impression and understanding of the text. The researcher identified concepts in the text and compared these to the themes and categories that had emerged from the theoretical background and literature review (see Table 4.2). The text (data) was deconstructed, compared, sorted and grouped in terms of similarities, dissimilarities and relationships of concepts, and corresponding notes and coding were added in the margins. The form of coding applied in this study was open/substantive coding with the use of colours to show differences and similarities in the raw data collected.

Colour coding also assisted with the ordering concepts gleaned from the text (data) into themes, categories etc., which were then coded for data analysis with the corresponding alphabetical codes. More than one code could be assigned to a concept, if it spanned themes, categories etc.

Figure 4.4 provides a snapshot of a transcribed interview featuring the reflective notes captured during the interview, interview questions and responses and the coding applied. Full transcriptions (i.e. the raw data) can be viewed in Appendix B.

Own reflective notes and observations	Transcript: P1	Code
	Section A: Data management	
	1. Do you use space geodesy data for your research? *Yes, I use mostly SLR and meteorological data.*	DM-SGD-U
Respondent P1 has vast experience in geodetic data and systems	2. Where do you search for space geodesy data? *CDDIS mostly, IGS, sonel.org, local repository*	DMS-I-S
	3. How do you access geodetic datasets? *FTP and HTTP*	DMS-I-A & GRDMS
	4. … *n*	

Figure 4.4: Snapshot of a transcribed interview

4.7.7 Modifying the coding scheme

The transcriptions of the participants' answers provided in the interviews were examined in detail, in order to refine the coding scheme and to determine whether there were any other themes, categories etc. that should be added. New and unexpected categories emerging during the reading, transcription and coding process were added to the coding scheme. Alphabetical codes were adjusted, realigned and reassigned to the data.

4.7.8 Data reduction

Data reduction is the process of selecting, focusing, simplifying, abstracting, transforming and removing of the data that appear in written-up field notes or transcriptions (Miles & Huberman 2014). Data need to be condensed to make it manageable and transformed to make it intelligible in terms of the research problem being addressed. In this study, the data was cleaned or reduced during the transcription phase. Aspects of the assembled data were emphasised, minimised or set aside. During this part of the analysis process, extreme care was taken not to omit relevant and important data.

4.7.9 Data capturing

Data capturing is the process of transforming data from one form of media to another (Datanet n.d.). Data, codes and/or text from the original material are entered/imported into a computer and stored. Files can be exported from one computer to another (Datanet n.d.). During this step of the analysis process, data were captured in a *Microsoft Office Excel®* spreadsheet. The data were then reviewed / monitored for accuracy. The captured data were stored on the HartRAO server to prevent data loss and to ensure its preservation.

4.7.10 Interpretation of data

The analysed data were interpreted by considering theoretical constructs (general) and the researcher's own sense of the meaning embedded in the text (specific), as well as the augmentations and amplifications of the researcher's understanding by referring to the broader context of the study (Bezuidenhout & Cronje 2014:236). At this stage of the analysis process, the researcher inferred and presented the reconstructions of meaning derived from the data. The data collected were interpreted by drawing inferences from existing theories and previously conducted studies. The researcher also relied on her own sense of meaning of the text. Since the interpretation of the data was context-sensitive, the researcher considered and described the way in which the reconstructed meaning from analysis of the text linked to the broader context. The researcher considered extraneous information and relevant factors that might affect the research problem and lead to false interpretation. The researcher assessed phrases and words in the text to discover intonations and inference, which led to further thematic linkages.

4.7.11 Findings

According to Welman & Kruger (2003), the results or findings of the research should be released to expand scientific knowledge. The reporting of the research findings is usually the last step in the research process (Jacobs 2014). Such findings can be presented in many formats. In this study, the findings are presented in the form of a dissertation. Evidence-based data obtained from authoritative sources were used to provide sufficient evidence in support of the findings. The research findings are presented as tables, figures and in discussions in Chapter 5. The researcher will use the findings to support and guide the designers in designing the new GRDMS.

4.8 Validity and reliability

Validity involves determining whether the research measured what it was supposed to measure (Koonin 2014:252). *Reliability* refers to the consistency, repeatability, stability, accuracy, credibility and precision of the measurement of the phenomenon (Koonin 2014:253). Content, sampling, criterion-related and instrument validity are some of the types of validity that may be encountered in research (Koonin 2014:256). Construct validity requires the use of correct measures for the concepts being studied. Internal validity (which is particularly important with explanatory or causal studies) demonstrates that certain conditions lead to other conditions and requires the use of multiple pieces of evidence from multiple sources to uncover convergent lines of inquiry (Yin 2013). According to Koonin (2014:257), external validity indicates whether or not the findings are generalisable beyond the immediate case(s). For example, the more variations in places, people and procedures a case study can withstand and still yield the same findings, the better the external validity Koonin (2014:257).

In this research, techniques such as cross-case and within-case examination, along with literature reviews, assisted in ensuring external validity and confidence in the research findings. The researcher strived to achieve and maintain all types of validity throughout the study. Emphasis was placed on the researcher's perspectives, preconceptions, assumptions and worldview, in order to increase credibility and to demonstrate validity of the research. The researcher conducted an interpretive analysis of the interviews and identified common themes and categories. The researcher did not attempt to influence participant responses during the interviews. All data collected were analysed in an objective manner and the findings were based on the actual data obtained from participants. The researcher collected data over a two-week period, which allowed sufficient time for themes and categories to emerge from the data.

History, attrition and maturation of data were taken into account, but it did not affect the study. *History* refers to the possible negative effects of the propagation of time on the study (Seliger & Shohamy 1989). Participants were not engaged in this study for an extended period of time; only for sufficient time to collect the necessary data. *Attrition*, which refers to changes in the composition of the sample being studied over time (Seliger & Shohamy 1989), did not affect this study either. *Maturation* refers to cognitive and developmental changes in the sample being studied (Seliger & Shohamy 1989). Participants involved in the study comprised geodesists, post-doctoral, doctoral and Master's students whose knowledge of and expertise in geodesy were unlikely to be affected during the two-week period.

Regarding interaction of participant selection and research, Seliger & Shohamy (1989) identifies the use of paid or volunteer participants as a threat to external validity, because they may not be representative of the wider population. Participants in this study were all employed by the NRF and received a monthly income. Participants were neither paid nor asked to volunteer for the study. According to Seliger & Shohamy (1989), participants who are aware that they are part of a research study may change their behaviour, which leads to distortion of results. The researcher's opinion is that none of the participants distorted their responses either before or after the study was conducted.

Reliability refers to the consistency of the results obtained from a study (Koonin 2014:254). Data obtained from the literature reviewed and the interviews conducted were unambiguous and not open to subjective interpretation. The researcher believes that independent researchers, on re-analysing the data, would come to a similar conclusion regarding the data. Data obtained from the literature and the interviews were qualitative data with low explicitness.

The researcher used the following two techniques to guard the integrity of the data:

- consensus reliability (Seliger & Shohamy 1989) – Data obtained from the literature and interviews were given to a colleague, who examined the data to determine whether he/she could identify similar themes and categories in the data as the researcher had originally identified; and

- regrounding (Seliger & Shohamy 1989) – Data obtained from the literature and interviews were analysed a second time and themes and categories were compared with the results obtained the first time, in order to confirm the original results.

Data obtained from the literature review and the interviews with a sample of geodesists, post-doctoral, doctoral and Master's students in geodesy revealed similar themes and categories. The researcher is of the opinion that, should independent researchers replicate this study, using the same data collection instrument and methods, selecting a similar sample of participants and adhering to the processes of inter-rater reliability and regrounding, the researchers would be likely to reach similar conclusions.

Saunders, Lewis & Thornhill (2016) identify observer and subject bias as well as error as threats to reliability. Objectivity versus subjectivity in research is another important consideration and the researcher needs to be sure that his/her personal biases and opinions do not get in the way of research (Saunders, Lewis & Thornhill 2016). Other biases to guard against include publication bias (which refers to the problem that positive results are more likely to be published than negative results); selection/sampling bias (e.g. omission and inclusion bias); performance measurement; procedural, bias; interview bias; and response and attrition bias (Shuttleworth 2018).

Interviews are open to bias and distortion (Seliger & Shohamy 1989). The researcher controlled this by means of inter-rater reliability and by discussing the results of the study with the head of the GRDMS project and the designers. The researcher ensured that the participants understood the questions and noted down the participants' exact response to the interview questions. At no time did the researcher attempt to shape responses.

Another form of bias to guard against is related to the number of participants involved in the study. According to Seliger & Shohamy (1989), the smaller the number of participants, the more likely the study is to be susceptible to biases created by an over-representation of some participant characteristics. The researcher acknowledges that the size of the group was relatively small. However, the group was large enough to rule out the influence of individual variability.

To demonstrate rigour is to provide participant validation. According to Lacy & Luff (2007:27), participant validation can range from sending feedback of findings to participants to the supply of transcripts or quotations to participants to check accuracy or consent for use. In this research, the participants were asked to comment on the interpretation of the data. Where participants provided feedback, any important issues arising from the feedback were considered. If there were competing interpretations, the researcher considered the competing interpretations, assumptions or particular views of the participants.

4.9 Research ethics

Research is required to be ethical; particularly when it involves the following: people, animals, genetic material, agriculture, living organisms, etc. (UNISA 2013). Ethics is a matter of integrity, but the implications reach much further than the individual (Louw 2014:262). Driscoll & Brizee (2012) and Saunders, Lewis and Thornhill (2016) assert that the key ethical issues that the researcher should are the privacy of respondents (possible or actual); the right of the respondent to withdraw partially or completely from the research process; informed-consent; confidentiality where necessary; permission from participants; intellectual property rights; and restricted use by granting access through a data enclave environment.

The importance of research ethics cannot be over-stated, because it may affect all stakeholders involved in a positive or negative way. A researcher acting with integrity adheres to ethical principles and professional standards that are essential for practicing research responsibly (Louw 2014:263). Researchers are required to adhere to the research ethics policy and guidelines pertaining to responsible ethical practice and behaviour in respect of copyright infringement, plagiarism, intellectual property, research methods and procedures, social and environmental interaction and representational concerns (UNISA 2017).

Before this study commenced, ethical approval was obtained from HartRAO's Managing Director, the resident Geodesist and IT specialist. Intellectual property rights and confidentiality issues were addressed as part of the informed-consent process. As this study is mostly exploratory in nature, the need for strict confidentiality is low. At the onset of the study, ethical clearance was also obtained by means of a formal application at the University of South Africa (UNISA) Ethics Review Committee. During the research process, caution and care were taken to ensure that the application of the research and the data collection methods complied with all the requirements of the *Ethics Policy of Unisa, 2013* (UNISA 2013).

Due to the nature of the study, it not only involved a review of the literature relevant to the research problem, but also human participation in the form of one-on-one, in-depth interviews. To guard against any form of ethical wrongdoing, basic ethical principles for research (e.g. moral and general ethics principles) were applied.

The perusal of moral principles (UNISA, 2013) included:

- autonomy (respect the freedom, rights and dignity of research participants);
- justice (benefits and risks should be fairly distributed among people);
- beneficence (make a positive contribution towards the well-being of people); and
- non-maleficence (not cause harm or injury to the research participants).

In this study, the perusal of general ethics principles (UNISA, 2013) included:

- Research that was essential and relevant - consideration was given to existing literature on the subject under study as well as alternatives available, demonstrating that the research is contributing to the pursuit of knowledge.
- Research that maximised public interest and social justice - research was carried out for the benefit of society and with the motive of maximising public interest and social justice.
- A competent researcher with ability and commitment to the research - the researcher executed the research in the required period of time to the best of her ability.
- A researcher respecting and protecting the rights and interests of the participants and institution - the researcher strived to respect and protect the dignity, privacy and confidentiality of participants and, where relevant, institutions, and strived not to

expose them to procedures or risks not directly attached to the research project or its methodology.

- Informed and non-coerced consent - informed consent (Appendix C) was given freely by the participation. Direct or indirect coercion and undue inducement of the participants in the name of research were avoided.
- Respect of cultural diversity - participants was treated as human beings within the context of their community systems. The researcher respected what is sacred and secret to the participants, including tradition, culture and religion.
- Justice, fairness and objectivity - participant selection was conducted in a fair and un-biased manner.
- Integrity, transparency and accountability - the researcher was honest about her own limitations, competencies, belief systems, values, etc. and guarded against using her position or knowledge for personal gain.
- Risk minimisation - the researcher strived to ensure that the actual benefits to be derived by the participants or society from the research clearly outweigh possible risks, and that participants were subjected to only those risks that are clearly necessary for the conduct of the research.
- Non-exploitation - care was taken not to exploit participants, researchers, communities, or institutions.

4.10 Conclusion

The rationale and research methodology of the study were described in this chapter, which included a discussion of research philosophy, approach, strategies, data collection and analysis. The study utilised primary, secondary and tertiary data sources. The most significant data collection techniques employed in the research process were a comprehensive literature review and in-depth, one-on-one interviews. The data were presented in descriptive form. Qualitative content analysis was conducted on the collected text (data). Concepts extracted from the data were compared, sorted and grouped into themes, categories etc. Various interpretations and constant comparative methods of data analysis were followed to find linkages between concepts.

The validity and reliability of the study were strengthened by cross-case and within-case examination, which was conducted in tandem with the literature review and the interviews (see Section 4.8). Validity was further strengthened by the perspectives, preconceptions and assumptions and the worldview of the researcher being declared. To guard bias, the results and findings were discussed with the head of the project (resident Geodesist), system designer and Information Technology specialists at HartRAO. To increase the confidence in the findings, respondents were asked to comment on the interpretation of the data.

Ethical approval was obtained from HartRAO's Managing Director, the resident Geodesist, IT specialist and the UNISA, Department of Information Science. Intellectual property rights and confidentiality issues were addressed as part of the informed consent process. Care was taken to ensure that the application of the research and data collection complied with all the requirements of the *Ethics policy of Unisa, 2013*. The resulting data analysis, interpretation and findings are presented in Chapter 5.

CHAPTER 5: DATA ANALYSIS, INTERPRETATION AND FINDINGS

5.1 Introduction

In the previous chapter, the research design and methodology applied to this study were described. According to Nieuwenhuis (2007:99), there are numerous approaches, traditions, processes and procedures associated with qualitative data analysis whereby researchers extract some form of explanation, understanding or interpretation from the collected qualitative data. In this chapter, the focus is on the qualitative analysis of the empirical data collected by means of one-on-one interviews, supported by findings from the literature, as reviewed in Chapter 3.

What is important to note is that an interpretivist worldview was followed, implying a manner of subjectivity in the review and analysis of collected data. From an epistemological viewpoint, the experience of the researcher as the librarian at HartRAO, responsible for the management of current and new data, is to be considered. Accessibility and the easy retrieval of surface and deep data is of the essence to comply with user information needs and demands.

In addition, the theoretical framework was also considered, in order to ensure that various data types, streamed from all geodetic tools, are managed through their life cycle to ensure that they are captured, saved, made accessible and archived for future use in either original or reworked formats. Details from existing data management systems related to CDDIS, UNAVCO GSAC and GNSS were, therefore, considered in the interpretation of data provided by participants. Literature from these and other sources were aligned to findings to support or, in some instances, contradict viewpoints of participants. The aim of presenting the findings is to provide a context within which key proposals can be made as to the design of a model related to the GRDMS for HartRAO.

The process of data analysis followed in this study was based on the qualitative data analysis method of Bezuidenhout & Cronje (2014:233) as well as Creswell (1994:155), comprising:

- data reduction, where Appleton (1995:995) refers to data reduction as "the process of selecting, focusing, simplifying, abstracting and transforming the data as the researcher elicits meanings and insights from the words of the respondents";
- data display, where the content of each interview was studied, transcribed, coded, re-analysed and divided into themes and categories that were displayed in table format (see Table 4.2); and
- drawing of conclusions in the form of descriptions and explanations, in line with existing literature.

A qualitative research method known as *interpretative qualitative analysis* was followed in this study. Qualitative data analysis is generally based on an interpretative philosophy that aims at examining meaningful and symbolic content of qualitative data in an attempt to approximate construction of the phenomenon (Bezuidenhout & Cronje 2014:232). Through qualitative analysis and interpretation, data is transformed into findings. Data gathered during the interviews, as transcribed and categorised in Appendix B, were analysed according to the layout of themes and categories as depicted in Table 4.2. Findings from these transcribed interviews were correlated with the findings from the corresponding data collected during the literature review. Linked to the proposed three-stage analysis, details per theme, as presented in Table 4.2, are presented in this chapter.

The key themes identified, include:

- Theme 1: Management of scientific data.
- Theme 2: Scientific data management systems.
- Theme 3: Data management systems used for geodetic data.
- Theme 4: Data standardisation.
- Theme 5: Data management system architecture.

Details on these themes, aligned to data obtained during the empirical study and aligned with existing literature, are provided in the sections below.

5.2 Theme 1: Management of scientific data: explanation, analysis and interpretation

Scientific data have intrinsic value, which requires proper management (Koopman & de Jager 2016:1). Knowledge and information regarding the management of scientific data is necessary for the design of the GRDMS, in order to ensure the design of an effective system. As indicated in Table 4.2, two main categories associated with the management of scientific data, “data management” (DM) and “data management systems” (DMS) were identified and included under Theme 1. Key data relevant to these categories are grouped and discussed in Section 5.2.1 and Section 5.2.2. Details per main category of the theme “management of scientific data” are further broken down into subcategories and presented in this section. The main theme, categories, subcategories and sub-subcategories as, aligned to Table 4.2, are presented in Table 5.1 again.

Table 5.1: Theme 1: Category breakdown

<table>
<tr><th>Theme</th><th>Main category and code</th><th>Subcategory and code</th><th>Sub-subcategory and code</th></tr>
<tr><td rowspan="7">1. Management of scientific data</td><td rowspan="2">Data Management (DM)</td><td rowspan="2">Space Geodesy Data (DM-SGD)</td><td>Data Usage (U)</td></tr>
<tr><td>Data Type (T)</td></tr>
<tr><td rowspan="5">Data Management Systems (DMS)</td><td>Design (DMS-D)</td><td>–</td></tr>
<tr><td rowspan="4">Internet (DMS-I)</td><td>Search (S)</td></tr>
<tr><td>Interface (In)</td></tr>
<tr><td>Access (A)</td></tr>
<tr><td>Retrieval (R)</td></tr>
</table>

An analysis and interpretation of data collected relating to the management of scientific data, and further broken down into data management (DM) and data management systems (DMS), and their respective subcategories and sub-subcategories, are presented in the sections to follow.

5.2.1 Main category: Data management (DM)

The focus of this main category is on managing information related to the acquisition, storage, protection, processing, dissemination and accessibility of data, as well as on obtaining insight into the management of raw data generated by various technique-specific geodetic instruments. Two sub-subcategories related to data use and data type were explored in this sub-section. A summary is presented in Table 5.2, with a description of each sub-subcategory following.

Table 5.2: Data management (DM) breakdown

Main category	Subcategory	Sub-subcategory
Data management (DM) is the process of managing information and data generated during the research project and includes processes such as data acquisition, storage, protection, processing, dissemination and accessibility (Trauth & Sillmann (2013).	**Space geodesy data (DM-SGD)** are raw and/or processed data relevant to the scientific discipline Geodesy. Examples of geodesy data include (but are not limited to) VLBI, GNSS, SLR, LLR, DORIS, seismic, gravimetric and meteorological data generated by technique-specific geodetic instrumentation (Coetzer & Botha 2016).	**Data usage (U)**, with reference to this study, “data usage” involves the application of data in research activities.
		Data type (T) is a particular kind of data item, as defined by the values it can take; the programming language used; or the operations that can be performed on it (Shaffer 2008:9), e.g. raw and processed data (Coetzer et al. (2015), Noll (2010 & 2017)).

Subcategory: Space geodesy data (SGD) and sub-subcategory: Data use (U) and Data type (T)

Related to the data usage of space geodesy data, data is generated in many different formats suitable for a specific purpose. In space geodesy, data is used to understand the fundamental properties of the earth, such as its geometric shape, orientation in space, and gravitational field, as well as the way in which these properties change over time (Torge & Muller 2012). Related to the data usage category, three researchers (P1, P5 and P9) and seven postgraduate students (P2–P4, P6–P8 and P10) were asked to answer two questions that related to:

- whether they use space geodesy data for their research. Feedback from this question will allow for the design of a system that will cater for the needs of the users; and
- identifying relevant stakeholders that will make use of a DMS. This question was asked to determine possible users of the new GRDMS in order that these users' needs and requirements may also be considered when designing the new GRDMS.

In terms of the first question, seven participants (P1–P6, P8) indicated that they make use of space geodesy data and one (P5) indicated that he/she does not use it as much as other colleagues do. Two participants (P9 and P10) do not make use of space geodesy data. One of these participants (P9) indicated that he/she does not use space geodesy data but is aware of and familiar with the data. One participant (P7) indicated that he/she does not currently use space geodesy data but has used it in the past.

Participants not using space geodesy data is HartRAO researchers and students currently employed to perform supporting roles, such as geodesy software development and IT management. Technique-specific data mentioned by the participants who do use geodetic data included VLBI, SLR, LLR and meteorological data. The fact that participants make use of various types of data is supported by the findings of Coetzer and Botha (2016), in that the scientific community is interested in a variety of geodetic data, based on their specific subject interests.

With reference to the second question, participants identified various potential DMS users, as presented in Figure 5.1. All participants considered researchers to be potential users of the DMS. Six participants (P2, P3, P5, P8, P9 and P10) expected students to make use of the DMS as well. Participants P2 and P3 were of the opinion that specialised individuals, laymen and curious people will also be interested in the use of the DMS. One participant (P2) indicated that research institutions, end-users, managers of systems and lecturers may also become potential users. Two participants (P4 and P10) saw data service providers as potential users. One participant (P4) also identified governing bodies, the scientific community and lecturers to be possible users of the DMS. Two participants (P7 and P8) indicated that geodesists and meteorologists would be users of geodetic data. Potential users of the DMS identified by participants align well with the findings of HartRAO (2012) and Coetzer et al. (2015), which indicates that geodetic VLBI, GNSS and SLR data generated by HartRAO instrumentation are distributed to global data service providers such as the CDDIS, IVS and the IERS, which provide data to users identified by the participants.

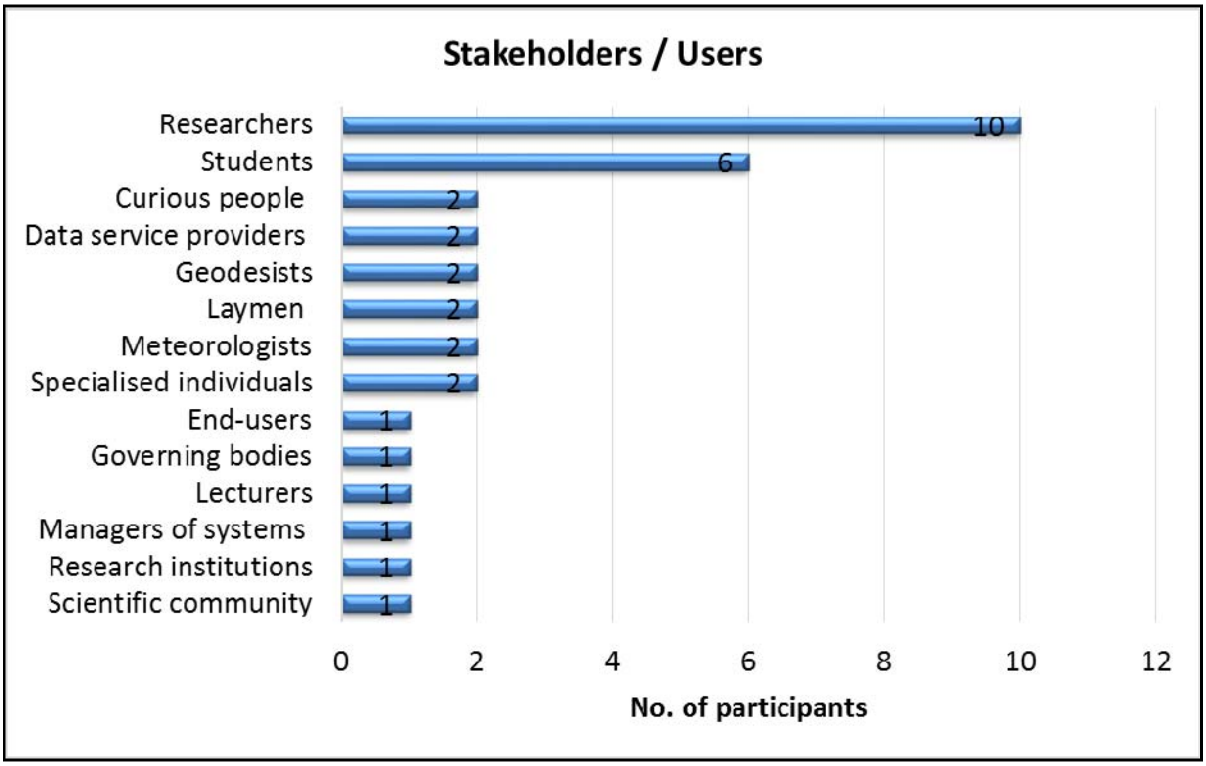

Figure 5.1: Possible DMS users as identified by research participants

Geodetic data is generated by various types of technique-specific geodetic instrumentation, e.g. VLBI, GNSS, SLR, LLR, and DORIS. The formats of raw and processed data delivered by technique-specific geodetic instrumentation are also of various types (Noll 2013). Participants were asked to provide detail on the type of data they would normally gather. This information is important to guide the designers of the GRDMS in catering for the various data types obtained and used by the scientific research community. The various data types of interest to the participants are presented in Figure 5.2.

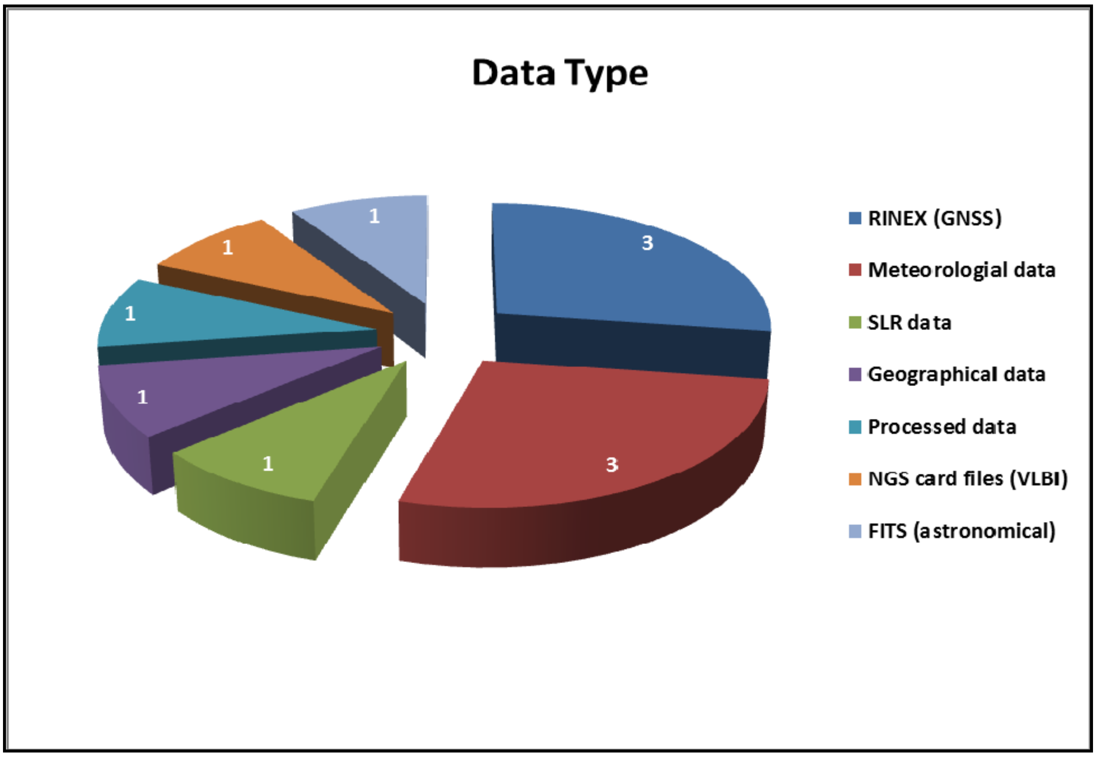

Figure 5.2: Type of geodetic data obtained and used by participants

Linked to Figure 5.2, three participants (P1, P5 and P10) use RINEX data (American Standard Code for Information Interchange (ASCII) format for raw GNSS data), with one participant (P5) indicating an interest in SLR data. Three participants (P2, P7 and P8) make use of meteorological data and two of these participants (P2 and P8) specified the type of meteorological data they use as including air pressure, humidity, wind direction and speed, atmospheric data and precipitable water vapour (PWV), and geographical data of the area surrounding HartRAO. One participant (P3) uses processed data, while another (P4) makes use of NGS card files (experiment files in ASCII format). One participant (P6) indicated that he/she uses geodetic data converted into flexible image transport system (FITS) files (astronomical data format).

Participant P9 did not consider this question to be applicable to him/her. Further questioning of participant P9 revealed that, although he/she had previously used geodetic data, he/she is not currently using geodetic data for his/her research, as the focus of the research moved outside the scope of geodetic data. Another participant

(P10) indicated that he/she is in the process of developing programs for the use of RINEX data and is, therefore, required to access as many data types as possible.

Data types identified by the participants align well with the literature reviewed. Technique-specific instrumentation at HartRAO produces the following types of data: raw VLBI data, NGS card files containing data, RINEX and SINEX GNSS data files and compressed tide gauge text files, SLR consolidated prediction format (CPF) data files, compressed LLR text files (new), compressed Roscosmos text files (new), Seismic files (new), Seedlink records, compressed gravimetric text files (new) and compressed meteorological text files (Coetzer & Botha 2016). The CDDIS manage pre-processed and analysed data, ancillary and metadata, as well as project management data (Noll & Michael 2016).

5.2.2 Main category: Data management systems (DMS)

"Data management systems" (DMS) as a second main category of the "Management of scientific data" theme, focus on computer system requirements. This is to ensure the effective and efficient design of the proposed GRDMS. Details on each subcategory and sub-subcategory are presented in Table 5.3.

Table 5.3: Data management system (DMS) breakdown

Main category	Subcategory	Sub-subcategory
Data management systems (DMS) are computer systems designed to perform data management functions (McLeod 1983).	**Design (DMS-D)** relates to the actions of planning, designing or modelling a data management system for a particular purpose (Saffer 2006).	
	Internet (DMS-I) is a large network of linked computers with a worldwide scope, which facilitates data communication (Rouse 2014). Internet service providers connect networks of computers to the Internet.	**Search (S)** is the act of searching for something in particular (Merriam-Webster 2018). With reference to this study, it is the purposive searching for data types, datasets, etc.
		Access (A) involves the act of obtaining information/data from a

Main category	Subcategory	Sub-subcategory
		computer system database or repository (Techopedia 2018). With reference to this study, it refers to the act of entering a data management system and obtaining data from the system's data holdings.
		Interface (In) is a boundary across which two independent systems meet and act on or communicate with each other (Beal 2018). In the context of this study, it refers to the software designed to assist the user in accessing a data management system and obtaining data from the system's data holdings.
		Retrieval (R) refers to the act of finding or extracting stored information or data from a system's database/repository (Techopedia 2018). With reference to this study, it refers to finding and/or extracting geodetic technique-specific data from a geodetic data management system.

Details related to the foregoing, according to participants and aligned to the relevant literature, are given in the following sections.

5.2.2.1 Subcategory: Design (D) and Internet (DMS-I) and sub-subcategory: Search (S), Access (A), Interface (In) and Retrieval (R)

Participants were asked to explain whether they have experience in data management system design. This information is important to the designers of the GRDMS, as they could tap into existing knowledge and experience. This finding aligns with the literature of Altintas, Berkley, Jaeger and Mock (2004) and Lim, Bosch, Dubois-Felsmann, Jenness, Kantor, O'Mullane, Petravick, Comoretto and the DM Leadership Team (2017), who state that experience in DMS design is important for the design of a new fully functional system. Four of the participants (P2, P5, P9 and P10) indicated that they do have experience in DMS design, while two participants (P1 and P8) indicated having some experience. Four of the participants (P3, P4, P6 and P7) have no experience in designing a DMS. Participant P2 elaborated, indicating experience in Oracle, MySQL and Microsoft Access. This information can be utilised during the design of the new GRDMS as part of testing the new design.

The focus of subcategory *Internet (DMS-I)* was to determine the data and communications facilities required and to be considered in the design of the GRDMS. It relates to how participants would want to search, access, retrieve and engage with the data via the proposed GRDMS. Four sub-subcategories were explored during the questioning of participants. These related to the search ability of the GRDMS, its proposed interface, access points and acceptable retrieval speed.

People use different methods and systems to search for data. One of these methods is via the Internet. Participants were asked to answer the following three questions relevant to this sub-subcategory:

- The first question focused on where participants search for space geodesy data online. This information is important, as it will allow designers to investigate and compare DMS provider sites, which will inform the design of the new GRDMS.
- The second question focused on the databases that participants prefer to use when searching for space geodesy data, as it may highlight and support alignment of the new GRDMS with other existing databases and repositories.
- The third question focused on search strategies used by the participants to search for geodesy data. Knowledge as to the search types/terms familiar to and used by participants can provide a basis for a taxonomy for the GRDMS, which will increase the retrieval rate and ensure that the required data is retrieved.

With regards to the first question, as shown in Table 5.4, participants identified the HartRAO repositories, servers and websites, remote repositories, servers and websites as the key sites on which they search for space geodesy data. The fact that eight of the participants indicated that they search digital platforms align to the findings of the University of Minnesota Libraries (2015) and Core Informatics (2016) that digital platforms, such as repositories, are used by researchers to deposit and preserve data generated during research projects. Data service providers, such as the CDDIS, IVS IGS, ILRS and IERS, provide remote access to their data holdings/repositories via their websites (Behrend 2013; Appleby et al. 2016 & MacMillan et al. 2016).

According to Coetzer et al. (2015), members of the scientific user community can access HartRAO data through the aforementioned international data service providers. However, there are some technique-specific data, such as GPS data, that can only be accessed directly via the URL of HartRAO's Space Geodesy Programme. The importance of this for the design of the GRDMS is to make the designers aware of the fact that people are currently using different methods and systems to search for HartRAO data.

Table 5.4: Space geodesy data Storage location / place

Location/Place	Number of participants	Participant codes
Local (HartRAO) repositories / servers / websites	4	P5, P7, P8, P10
Remote repositories/servers/websites	4	P2, P3, P6, P8
Both: local and remote repositories/servers/websites	2	P1, P4
Not applicable	1	P9

In terms of the second question, according to Table 5.5, most of the participants use data service provider databases when searching for data. Two of the participants (P5 and P10) use the HartRAO geodesy database, Geoid, exclusively. One participant (P8) uses all three mentioned databases; while participant P7 indicated that he/she is not sure which database he/she is using, as he/she accesses data via proxy from the HartRAO server. The participants (P6 and P9), who indicated 'Not applicable', did not elaborate on their answers.

The fact that participants use subject-specific databases align to the views of Wier, Boler and McWhirter (2012), in that databases should offer access and retrieval options related to specific space geodesy themes. The CDDIS and UNAVCO GSAC systems make use of commercially available relational database management systems to manage the data contained in their archives. According to Noll (2017), the main benefit of using such a system is its flexibility, as it eliminates the need for the continued development of software packages for the retrieval and display of data to the users. What this means for the design of the new GRDMS is that the new system needs to mirror current systems being used, but with a stronger focus and emphasis on HartRAO's geodesy datasets, including data generated by geodetic instrumentation in Africa.

Table 5.5: Databases used by the participants to search for data

Databases	Number of Participants	Participant Codes
Data service providers databases	5	P1, P2, P3, P4, P8
HartRAO Geodesy database 'Geoid'	3	P5, P8, P10
Meteorological databases	1	P8
Not sure	1	P7
Not applicable	2	P6, P9

Regarding the third question, participants listed eight different search types/terms, as indicated in Table 5.6. Most of the participants (P2, P4, P6, P7, P8 and P9) use more than one search type/term when searching for data, while two participants (P1 and P3) use only one search type/term. Participants P5 and P10 use URLs provided by the HartRAO Space Geodesy Programme and did not indicate which other search types/terms they use.

Table 5.6: *Search types/terms used by HartRAO researchers and students*

Search types								
Keywords	**Station name**	**Experiment names**	**Observation names**	**Session names / Codes**	**Year**	**Author search**	**Principal investigator (PI)**	**None**
P2	P1	P4	P6	P4	P4	P8	P9	P5
P3	P2	P6	–	–	–	–	–	P10
P7	P7	P9	–	–	–	–	–	–
P8	P8	–	–	–	–	–	–	–
–	P9	–	–	–	–	–	–	–

Linked to the findings in the literature review, space geodesy data management systems allow users to search for data via the database by using keywords, station names, experiment or observation names, session codes and year or day-of-year search strategies (Noll, Michael & Pollack 2015). The implication for the development of the GRDMS is that the system needs to be designed in such a way that it aligns with other existing databases and repositories familiar to the users. The GRDMS should allow for the use of search types/terms familiar to the users. Therefore, it is important for the designers to incorporate a taxonomy consisting of familiar search terms in the design of the system.

Geodetic data stored at various data service providers, such as the CDDIS and UNAVCO, can be accessed and retrieved by means of a graphical user-interface (Noll et al. 2012). Participants were asked whether data via these data service providers were displayed / presented in an easily understood manner. This question was asked to determine whether the system interface used by these providers presents data in an easily understandable and user-friendly manner. This information is required to guide the designers in developing a modern interface, tailored to the needs of the user community.

As can be seen in Table 5.7, analysis of the data revealed that none of the participants responded in the negative, but that participant P1 responded with “Medium” and participant P5 with “Okay”. Three participants (P6, P7 and P9) responded with a “Yes” only. Four participants (P2, P3, P4, P8 and P10) expanded on their answers as follows (also Table 5.7):

- *Yes, the system displays the data in a format that can be understood. It is user-friendly. (P2).*
- *… users can specify how they would like the results displayed (P3);*
- *… easy if you know what the data is all about, but if you don’t know, then it is not easily understood (P4).*
- *… the format in which the data is displayed appears old and does not grab my attention (P8).*
- *Yes, but the filenames given for the datasets are difficult to be understood by a novice (P10).*

All participants found the manner in which the data is displayed/presented easy to understand. However, some participants indicated that prior knowledge of the subject discipline may be required to understand the results displayed, and that the manner in which the data is displayed appears archaic.

Linked to the literature, Thakur (2016a) indicates that data management systems must be able to provide user-friendly graphic user interfaces. The CDDIS system provides an interactive graphic user interface, which can be used by both researchers and novices (Noll 2010).

Table 5.7: Interface user-friendliness assessment

Participants	Interface user-friendliness assessment
P1	*Medium*
P2	*Yes*
P3	*… users can specify how they would like the results displayed*
P4	*… easy if you know what the data is all about …*
P5	*Okay*
P6	*Yes*
P7	*Yes*
P8	*… the format in which the data is displayed appears old ...*
P9	*Yes*
P10	*"Yes, but the filenames given for the datasets are difficult to be understood by a novice"*

The GRDMS must be designed in such a way as to allow easy access via user-friendly graphic user interfaces and to present the requested data in an easily understandable manner.

Participants were asked to answer three questions relevant to access. The first question asked of the participants was how they access geodetic datasets, as the designers need to understand the way in which users' access data on current systems, in order for the new GRDMS to be able to accommodate the current and various alternative access methods.

- The second question enquired from participants whether they know where to find HartRAO datasets and how they go about accessing the data, as the designers need to assess whether users know where to find data and how to access it.
- The third question enquired about the level of access experienced by participants when accessing data hosted on servers. This question was asked so that the designers of the new system may have an idea of the level of access that will be required from the new GRDMS.

With regards to the first question, nine of the participants (P1–P8 and P10) indicated that they obtain data by logging onto websites, such as those of the IVS and CDDIS by using URLs, ftp and http. They also access data by logging onto severs such as HartRAO's Geoid server. One participant (P9) indicated that this question is not applicable to him/her.

The fact that the majority of the participants obtain data by logging onto websites aligns with the views of Coetzer et al. (2015), Noll (2015b) and Wier, Boler and McWhirter (2012) in that method for accessing geodetic data include access to data repositories via the Internet by means of ftp, http and URLs. The importance of this for the development of the GRDMS is that the system must be designed in such a way that it will allow users to log in using URLs, ftp and http. This method of access is what the users are familiar with and what they have been using to access data stored in different systems.

In terms of the second question, all participants knew where to find HartRAO datasets. – Participant P1 uses ftp / http to gain direct access to HartRAO's data holdings. Participant P2 indicated that he/she gains accesses to SLR site/station data from the ILRS repository by following the "Data and Product" link provided. Participant P3 uses the NASA/Jet Propulsion Laboratory (JPL) Horizon platform to find the data. He/she did not indicate how he/she go about obtaining the data. Two of the participants (P4 and P7) find data on the IVS website, with P4 specifying that he/she navigates the website by following the tabs and links. Participant P7 did not indicate how he/she goes about to obtain the data. Five participants (P5, P6, P7, P8 and P10) indicated that they use the space geodesy programme webpage to access HartRAO datasets. These participants did not indicate specifically how they go about accessing data. Participant P9 did not specify where he/she finds HartRAO's datasets and did not indicate how he/she goes about obtaining access to space geodesy data. A summary of the access points preferred by participants is provided in Table 5.8.

Table 5.8: Access to HartRAO data holdings

Participants	**Access points**	**Method of access**
P1	HartRAO data holdings	Use ftp / http to gain direct access
P2	ILRS repository	Following *Data and Products* link
P3	NASA/JPL Horizon platform	Not indicated
P4	IVS website	Navigate the website by following tabs and links
P7	IVS website	Not indicated
P5, P6, P7, P10	Space geodesy programme webpage	Not indicated
P9	Not indicated	Not indicated

Participants' use of websites such as the IVS, ILRS, NASA's Horizon platform and the HartRAO space geodesy programme webpage to search for HartRAO datasets, aligns with the findings of Coetzer et al. (2015), that international/global data service providers such as the CDDIS, IVS, ILRS, provide HartRAO data via different access points. This increases the usage of HartRAO's geodesy datasets and makes for easy retrieval of required data from various databases. The design of a new data service system, such as the GRDMS for HartRAO, will increase the visibility of HartRAO's data holdings and will provide more access points to HartRAO datasets. Ultimately, the new GRDMS will also be added to the existing global network of geodetic data service provider data management systems, which will contribute to the increase of global geodetic data services.

With regards to the third question, the majority of participants (P1 and P3–P10) found the level of access experienced when accessing data to be easy. Only one participant (P2) indicated he/she found the level to be ‘moderate’ due to the terminology used on the ILRS website, while participant P7 specified that access is easy if the user has prior knowledge of the system. The new GRDMS will have to allow for easy access. The level of access provided by the new system should be designed with the novice person in mind. When designing the GRDMS it is, therefore, important to take cognizance of how users access data on current systems; where they go to find data; and how they go about accessing the data, to ensure that the new GRDMS provide the necessary search and retrieval service.

Speed is of the essence and users often become agitated and irritated if data is not presented in a timeframe acceptable to them. Therefore, the speed at which data is to be made available is important for designers, in order to ensure that the new GRDMS provides the requested data in an acceptable and reasonable period of time.

Participants were asked to explain how long it takes to retrieve data from the DMS they are using. Three participants (P2, P5 and P7) indicated that it takes them less than a few minutes to retrieve the data, with one participant (P2) explaining that this is due to the fact that he/she is familiar with the subject discipline and the access tools. Participant P9 indicated that it takes him/her “a minute or two” to retrieve data. Three participants (P1, P3 and P4) indicated that it takes them less than 5 minutes, with P1 stating that this depends on the type of station from which data is being retrieved. Three participants (P6, P8 and P10) indicated that it takes them less than 10 minutes to retrieve data. In other words, participants take 1–10 minutes to retrieve the data from the DMS they are using. The retrieval of data is dependent on the speed of the fibre cables used; the speed of the input and output query handler within a computer system; and the amount of data they download.

When designing the GRDMS, it is important to ensure that the system provides data within minutes or, better still, seconds. Linked to the literature of Purdon (2018), depending on the fibre cables and devices, binary messages travel at the speed of light.

The devices situated at either end of the fibre cables are configured to control how many megabits of data is transmitted per second. Tamara (2009) points out that downstream (the direction towards the user) bit rates can be as much as 400 MBps for business connections and 320 MBps for residential services in some countries, and that upstream traffic (originating at the user) bit rates range from 384 KBps to more than 20 MBps.

What is to be considered when designing the new GRDMS is the inclusion of new wireless Internet technologies such as Light Fidelity (LiFi). This high-speed bi-directional wireless network communication technology uses light to transmit wireless data embedded in its light beam at very high speeds (a hundred times faster than standard WiFi (Mercer 2018)). This technology is ideal for radio astronomy observatories as it does not produce high levels of radio-frequency interference (Mercer 2018), which is highly detrimental for radio astronomy research.

5.3 Theme 2: Scientific data management systems: Explanation, analysis and interpretation

Scientific data management systems are designed for a specific scientific purpose, such as the collection, archiving or processing of scientific data (Martin & Ballard 2010; Valle 2013). One main category related to scientific data management namely, 'Research data management systems (RDMS)', was identified via the literature and feedback from participants and is included under Theme 2.

This category deals with research data management systems used in scientific disciplines, which link to scientific data management systems. Key data relevant to this category are grouped and presented in Table 5.9. Linked to the main category, three sub-categories pertaining to components, characteristics and criteria could be identified (Table 5.9). This relates to the literature presented in Tables 3.1 and 3.2 (Chapter 3), where components, criteria and characteristics suggested by the theorists were presented in a matrix format.

Table 5.9: Research data management system (RDMS) breakdown

Main category	Subcategory
Research data management systems (RDMS) are computer systems and tools used to support unified research data management and data curation (Tufts University 2013).	**Components (Ca)** are parts of a computer, a computer system or a network of computers, for example, hardware components such as mechanical and electrical parts, and software components such as programs used for operating the computer(s) (Kaif 2013).
	Characteristics (Cb) are attributes of a computer and computer systems, such as: speed, accuracy, diligence, versatility and storage capacity (ScienceHQ 2013).
	Criteria (Cc) are standards by which something (such as computers and computer systems) can be judged (*The Merriam-Webster dictionary* 2018).

The focus of this theme is on obtaining feedback from participants regarding characteristics, criteria and components of data collection and retrieval that they wish to see form part of a new GRDMS. Detail on the findings in this section is presented per subcategory in the following sections.

The sub-category research data management systems (RDMS) consist of different components, e.g. hardware, software, processes, characteristics and criteria (Rob & Coronel 2007). Participants were asked to list the components they perceive as forming part of a modern data management system to ensure that the system can cater for large volumes of data. This information is important for designers, so as to identify components that would allow the new GRDMS to handle, manage and process large volumes of data in a short period of time. Components identified for inclusion by the participants are presented in Figure 5.3.

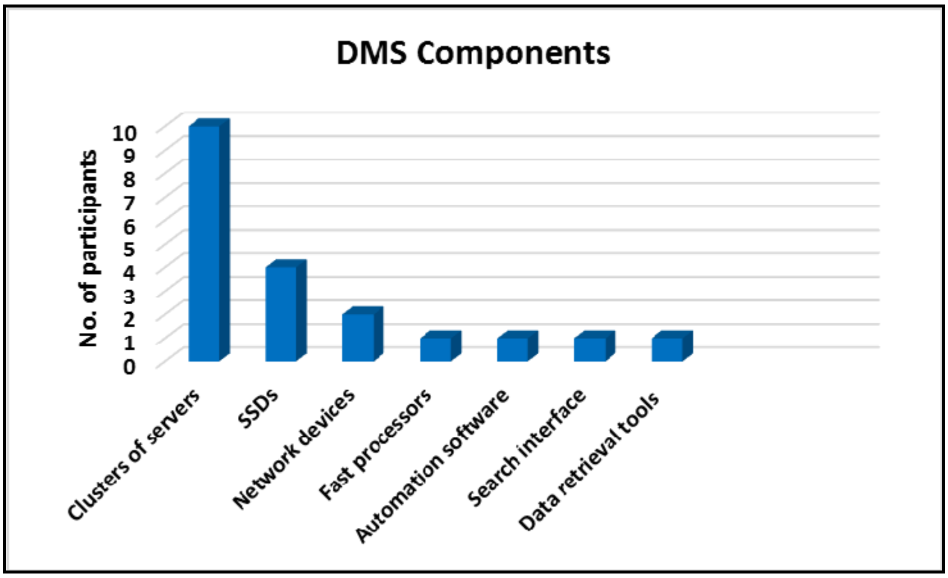

Figure 5.3: DMS components identified by participants

Apart from standard components, e.g. monitors, keyboards, motherboards, etc., most of the participants indicated that servers or clusters of servers should be included to cater for large data volumes. Four participants (P1, P2, P4 and P9) identified solid state drives (SSDs) as a requirement for a modern DMS. Two participants (P3 and P4) indicated that network devices should be included. One participant (P5) was of the opinion that fast processors are required, while another participant (P1) listed automation software, search interfaces and data retrieval tools, such as DOIs, as components for inclusion.

Linked to the literature of Austin (2016) and Wier, Boler and McWhirter (2012), data management systems should consist of sets of interacting smaller systems known as *subsystems* or *functional units*. Geodetic data management systems consist of both hardware and software components that can receive, structure and store data on server clusters and disseminate data to the scientific community (Noll 2013). According to Wier, Boler and McWhirter (2012) the UNAVCO GSAC system consists of various components contributing to the whole, such as an application server, subsystems within the application server, such as a query and output handler, repository with subsystems and servers for data transfer.

The GRDMS designers need to include all the necessary computer network information technology and computer communication components, such as (but not limited to) modern storage technology (e.g. solid state drives, server clusters, etc.); modern processors capable of processing data within a petaflop (a unit of computing speed equal to one thousand million million (10^{15}) floating-point operations per second) (Indiana University 2018); state-of-the-art automation software; and network devices. To be included are modern interacting functional units and subsystems (e.g. graphic user interface features and web applications, etc.) in the design of the new GRDMS. Also, to be included are data management software to manage the data collection activities; the data correlation and processing activities; storage and archiving activities; the data retrieval activities; and the data access and supply activities (as depicted in Chapters 2 and 3).

Participants were asked to answer three questions relevant to subcategory: characteristics (Cb):

- The first question asked of participants to list key characteristics of an RDMS, so that the designers can consider these characteristics when designing the new GRDMS.
- The second question asked of participants to identify the objectives of an RDMS to be considered by the designers as objectives for the new GRDMS, in order to ensure that the end product meets the objectives it was designed for.
- The third question asked the participants to identify the type of data they consider to be managed by an RDMS.

In terms of the first question, participants listed a number of characteristics an RDMS should possess:

- manage and include data types (e.g. metadata, ancillary data, RINEX and SINEX format);
- manage and structure data;
- apply data standards (similar to those used by the CDDIS);
- provide accurate data;

- contain a high-level of data integrity and simplicity;
- manages data (e.g. capture, store, archive, retrieve, display, backup, etc. data);
- provide access techniques, methods and tools such as remote access, secure access, multi-user access, multiple data access techniques, navigation, links to datasets, active persistent identifiers modern features, graphic user interface features, not too many pop-ups and graphics, help functions and glossary;
- can accommodate user growth, system expansion and versioning;
- allows for updates and scalability;
- compatible with different operating system (e.g. Linux and Microsoft Windows);
- store raw and processed data, as well as data products;
- display data holdings; and
- conduct fast processing and manages user accounts.

The views of the participants align to the views of Noll (2010 & 2016) and Thakur (2016b) that general characteristics of a data management system include the management of the flow of data within the system and between systems, and the integration and storage of large amounts of raw and processed data of homogeneous and heterogeneous nature and of different types and formats (e.g. SINEX and RINEX) It should also be able to provide multi-user access; manage the processing of user requests in a matter of seconds; allow the use of DOIs and other access tools; capture metadata according to metadata standards; support data sharing and interoperability; manage access; and provide an user-friendly graphic user-interface for easy navigation.

According to Noll (2010), the CDDIS system possesses characteristics that allow users to register on the system and provides an on-line interactive menu-driven interface. The CDDIS allows for multiple data access techniques, creates and manages different user account types, data independence, transaction processing and backups. The aforementioned characteristics provided by the participants and those found in the literature (discussed in Chapter 3) are important considerations in the design of the new GRDMS to ensure that data types from current and future geodetic instrumentation can be catered for and that the new GRDMS is fully functional.

With regards to the second question, participants listed a number of objectives for an RDMS, which include to: organise datasets; capture metadata; enable searching; provide easy quick access to data in a simple interactive way; provide data usage statistics; provide features to guide end-users; maintain data management functions (e.g. ingest, transfer data at high speeds, store data according to standards and provide access to data); and provide accurate high-quality reliable data.

Objectives identified for an RDMS by the participants are in alignment with the literature of Noll (2010) that RDMS objectives are to manage data management activities and to cater for the needs of the users in providing high quality accurate geodetic data. According to the views of Wier, Boler and McWhirter (2012) the objectives of a data management system should include the streaming of data received from various instruments to application servers and storing it in a database, which will serve as a repository. According to Thakur (2016a), the objectives of a data management system are to manage the processing of data requests and to manage analysed data. The University of Minnesota Libraries (2015) indicates that data management system objectives should be to cater for data governance and stewardship; architecture management; data development; database operations management; data security management; reference and master data management and curation; data warehousing; document and content management; metadata management; and data quality management. The system should cater for multiple-user access and share metadata between systems. Designers of the new GRDMS need to understand the objectives of a RDMS, to ensure that the end product (which is a fully functional GRDMS) will be met.

Regarding the third question, five participants (P2, P4, P5, P9 and P10) indicated that they believe a RDMS should manage raw/pre-processed and processed data. Three participants (P1, P3 and P6) were of the opinion that all geodetic data should be managed by the RDMS. Two participants (P7 and P8) thought that the RDMS should cater for data generated by HartRAO's geodetic instrumentation. Four participants (P1, P5, P9 and P10) also mentioned ancillary data, auxiliary data, data products, data usage information and metadata.

The view of participants that a RDMS should cater for raw, pre-processed and processed data is in alignment with the literature of CDDIS (2015), Noll & Michael (2016) and Thakur (2016). These sources indicate that an RDMS should be able to manage raw and processed data of different types and different formats. The GRDMS needs to be designed to cater for specific types of data (as indicated above and discussed in Chapters 2 and 3). The new GRDMS must be designed to cater for different data types, which are unique and scientific in nature and that require specialised treatment and management.

To obtain detail on the criteria that participants viewed as essential in the development of a scientific data management system – i.e. criteria that will cater for the provision of geodetic data in particular participants were asked to answer two questions relevant to this subcategory:

- The first question asked of participants whether a RDMS should have capabilities for interaction with other DMS. The question was asked to determine whether the participants have knowledge of capabilities for interaction between systems that could be leveraged by the designers of the new GRDMS.
- The second question asked participants to list criteria that should be considered for the design of the GRDMS. The aim was to establish a list of criteria – not to determine which criteria are more important than the others.

Regarding the first question, all participants agreed that an RDMS should have capabilities for interaction with other systems. Five participants (P2, P3, P4, P5 and P9) elaborated on their answers as follows:

- Participant P2 – "*yes, because it allows for improvement in system performance and access, e.g. The Centre for High Performance Computing (CHPC) system*".
- Participant P3 – "*yes, because for example VLBI networks are situated on different geographical sites. Systems need to be interoperable to allow for the sharing of data*".
- Participant P4 – "*data service providers such as the IERS obtain data from the CDDIS*".

- Participant P5 - thought that interoperability between systems is important.
- Participant P9 – "*yes, as long as it is setup correctly and have strong security measures to protect the data and the systems. Access should be limited*".

Participants were of the opinion that the RDMS should be interoperable to allow data sharing over geographically dispersed sites and to increase system efficiency. This aligns with the views of Mashaba et al. (2016) and Wier, Boler and McWhirter (2012) that a DMS should have capabilities for interaction with other systems (e.g. sharing different types of data, including metadata and storage). According to Behrend (2013), RDMS interact with other systems by sharing, hosting, distributing and transferring geodetic data generated by data providers (i.e. HartRAO) to global data service providers (i.e. the CDDIS, IVS and IERS), where the data is correlated, analysed and made accessible to the scientific user community. These systems and the units within the systems interact with one another to facilitate data distribution and transfer (Mashaba et al. 2016). These are important considerations in the development of the new GRDMS to ensure that the system has the necessary capabilities to manage data within the system itself as well and to interact with other systems, such as that of the CHPC, CDDIS and IVS. This is an essential capability, as it is HartRAO's mandate to collect and provide scientific data to the global scientific community.

With regards to the second question, participants identified thirty-two criteria that should be considered for the design of the new GRDMS as indicated in Table 5.10). The list of criteria is considered in its entirety: criteria frequently mentioned carry the same weight as those less frequently mentioned.

Table 5.10: Data management system criteria provided by the participants

Criteria	Participants codes	Criteria	Participants codes
Effective management of data processes/flow	P5, P10	Large storage capacity	P4, P5
Fast processing	P2, P9	Load balancing	P2, P9
Complete	P1, P2	Reliable	P2, P9
Accurate	P2, P9	Modern	P2, P3
Searchable	P1, P4	Query facilities	P3, P5

Criteria	Participants codes	Criteria	Participants codes
Access:		Data:	
– Secure	P2, P4, P9	– Centralised storage	P4
– Fast	P4, P7	– Standardised, structured	P1
– Easy	P4, P7	– From various instruments	P5
– User-friendly interface	P2, P3	– Raw, processed	P3, P5
– Remote	P6, P8	– Rapid analysis	P5
– Multi-user	P2, P5, P9, P10	– Archive processed data	P5
Analysis tools	P4	Products	P4
DOIs	P4	– Metadata	P5
Publications	P4	– Sharing of metadata between systems	P5
Maintained	P3	– Current, historical	P3
Communication - users and system manager/administrator	P3	– Up-to-date	P4

Criteria identified by participants align with the findings by Tenopir, Birch & Allard (2012), Behrend (2013), Fox and Harris (2013:WDS3), Wier, Boler and McWhirter (2012) and Noll (2017) that key criteria to be considered for an RDMS should include the data life-cycle; data structuring and organisation; up-to-date, fast, reliable, user-friendly, accessible and searchable data; secure access; data streaming and load balancing; large storage capacity; data preservation (archiving); and catering for the needs of the users. According to Core Informatics (2016), RDMS developed for managing large volumes of geodetic data should be able to provide automated data capture frameworks; drive data ingestion; integration; and management. According to Noll (2017), RDMS should be able to receive, structure, analyse, store and disseminate data and systems should be able to monitor folders and network files automatically; capture data files at remote locations/sites; and store raw data files in repositories. Noll & Michael (2016) indicate that RDMS should be able to receive different categories of data (i.e. raw and processed data) from the regional data centres and the systems should store data files in main/root directories in an archive/repository. Research data management systems should be able to stream various geodetic instruments' data to application servers, where the data can be stored in databases/repositories (Wier, Boler and McWhirter (2012). Computer storage devices, such as SSDs, should be used to store large data volumes indefinitely (ZDNet 2015). Metadata standards and DOIs can be used for organising geodetic datasets. According to the University of Minnesota Libraries (2015), a system should cater for multiple user access and users should be able search for data via the Internet.

Linked to the above suggestions and findings for Theme 3, key considerations relating to the design of the GRDMS, with specific reference to the management of research data, should be the consideration of key objectives of an RDMS, as indicated above and discussed in Chapter 3. Also to be considered when designing the GRDMS, are the inclusion of the above-mentioned system components (e.g. modern network information technology and computer communication components, fast modern processors, large storage devices, etc.); characteristics (e.g. able to manage, integrate and store large amounts of raw and processed data of homogeneous and heterogeneous nature and of different types and formats, able to manage the data flow within and between a system/systems, etc.); and criteria (e.g. up-to-data, fast, reliable, user-friendly, accessible and searchable system, etc.) and the rest of the components, characteristics and criteria listed by participants and discussed in the literature (Chapter 3).

5.4 Theme 3: Data management systems used for geodetic data

As can be concluded from information presented in the literature review (Chapter 3), data management systems enable researchers to document, deposit and manage their data, facilitating efficient search and retrieval (Tufts University 2013). One main category relevant to the uses of a data management system applicable to this research is how it can be applied to ensure the deposit, management and retrieval of geodetic data in line with the technique-specific data that will be streamed to HartRAO. Key data relevant to this category are presented in Table 5.11.

Table 5.11: Geodetic data management system (GDMS) breakdown

Main category
Geodetic data management systems (GDMS) are computer systems used for capturing, cataloguing, archiving and storing data generated by instruments and applications in a compliant, pre-defined manner best suited for its intended use, whether it is for structured, unstructured or semi-structured data (Hayward 2017).

Linked to this main category, participants were asked to answer three questions:

- The first question asked participants whether they use DMS to manage their data and to explain why they make use of the particular DMS. This question was asked to obtain an understanding of the reasons behind participants' choices of DMS, in order for their knowledge could be tapped into for the design of the new GRDMS.
- The second question asked of participants whether they are familiar with the CDDIS and UNAVCO GSAC databases. The designers need to know whether participants are familiar with these systems, as they are the international geodetic data service providers currently being used by the global scientific user community and for which HartRAO provides data.
- The third question asked of participants to what extent the design of the GRDMS for HartRAO complies with technological expansions and institutional needs, in order to gauge whether participants feel there is a need for, and would make use of, a new GRDMS.

In terms of the first question, Table 5.11 displays participants' views on the use and their reason/s for the use of a particular DMS. Four participants (P1, P2, P5 and P9) indicated that they do use a DMS to manage their data and provided reasons for using the particular DMS. Six participants (P3, P4, P6, P7, P8 and P10) indicated that they do not use a DMS. Those who indicated that they do use a DMS provided a variety of reasons for using the specified DMS, be it familiarity, ease of use, suitability, or for specific purposes such as data processing and analysis. According to the views of Dubois-Felsmann (2016:18), DMS are used by people to facilitate data management processes and activities. Noll (2017) states that centralised data management systems, such as the CDDIS, have been designed to cater for the data needs of the geodetic scientific community. Designers of the GDMS must ensure that the system caters for the needs of the users (as per reason for usage displayed in Table 5.12) while facilitating data management processes and supporting research activities of the organisation it serves.

Table 5.12: Research data management systems used by participants

Participant	Data management system	Reason for usage
P1	Yes, in-house system, not all functional	All that is available currently
P2	Yes, Explorer, for structuring files	Easy to use for type of data he/she works with
P3	No, not really, Microsoft Office Excel	Not applicable
P4	No, VieVS platform	Software used internationally for VLBI data processing and analysis, ideal for his/her research
P5	CDDIS	Easy to use, contains complete, accurate raw and processed geodesy data and products
P6	Not yet	Not applicable
P7	No, not currently, have used HartRAO system previously	HartRAO provides geodetic data to national and international scientists
P8	Not really	Well familiar with Microsoft Office Excel, good enough for their purposes
P9	New Computer Control System (NCCS) for HartRAO astronomical data	Suitable, used by other astronomical observatories
P10	Not at present (currently developing software)	Not applicable

With regards to the second question, all except one of the participants (P8) indicated that they are familiar with the systems and databases of the CDDIS and UNAVCO GSAC and are aware of the fact that HartRAO is a regional data centre that provides data to global data service providers such as the CDDIS and UNAVCO. Participant P2 indicated that, if he/she has to choose between the databases, he/she would use the CDDIS. Participant P4 indicated familiarity with the CDDIS but not with UNAVCO GSAC, and mostly uses the CDDIS to access data. UNAVCO GSAC hosts mostly GPS data, which the participant is not currently using. Participant P7 indicated familiarity with these databases but does not really make use of them.

According to Wier, Boler and McWhirter (2012), Appleby et al. (2016:18), MacMillan et al. (2016:74), Searle and Petrachenko (2016:78–79) and Noll (2017), HartRAO provides data to internationally recognised data service providers, such as the CDDIS and UNAVCO. Members of the scientific research community can access geodetic data from both of these systems (but not only from these two systems). The GRDMS should be able to host all of HartRAO's geodetic data at a central point and should be able provide access to HartRAO's data to the global scientific community.

The system needs to be designed in such a way that members of the scientific community and the public can access HartRAO's data without having to access one of the current data service providers' data management systems (e.g. the CDDIS and UNAVCO GSAC systems). However, to comply with the organisation's data provision mandate, the new system also needs to be designed in such a way that it will maintain and support HartRAO's capabilities for the supply of geodetic data to the international data service providers.

With regards to the third question on whether the development of a new GRDMS for HartRAO complies with technological expansions and institutional needs, two participants (P1 and P4) mentioned the lack of capabilities of the current system at HartRAO to provide all technique-specific geodetic data. Participant P1 also referred to the lack of the current system to store data in a central databank and of using standardised structures for the organisation of geodetic data. Participant P1 was of the opinion that automation is crucial and that statistics on the use and downloads are required. Two participants (P8 and P10) pointed out the value of a new DMS that can host all geodetic data in a central location. Participants P5 and P9 mentioned the current expansion of HartRAO's geodetic instrumentation and capabilities, requiring a system independent of the IVS and CDDIS to manage the additional technique-specific data types, products and analysis tools. Participants P3 and P4 considered a new DMS to be able to fill the void in data service and provision existing in the network of international geodetic data service providers. Participants P2, P3 and P10 were of the opinion that a new DMS would establish HartRAO as a provider of quality geodetic data in the global research community. Six participants (P3, P5, P6, P7, P9 and P10) pointed out the benefit of such a DMS to the local and international geodetic science community, society and to the organisation itself.

According to the participants the benefits of having the new GRDMS include:

- it will replace the existing system;
- it should be an automated system, that should handle internal and external structured data (e.g. store and retrieve data);
- the end-user will be able to either receive data subsets manually or, to retrieve the subsets by using scripts/software to retrieve the subsets, which has less human interaction due to being automated. Less human interaction with the system will reduces errors in consistency and repeatability of results;
- all of HARTRAO's geodetic data is available at a central location with multiple access points;
- it will automatically process raw data into data products without requiring human interaction;
- it will automatically generate data holdings summaries; and
- it will be independent of the CDDIS, IVS and other data service providers (providing data redundancy).

According to participants P2, P9 and P10, the system would expand HartRAO's international scientific collaborations. According to participant P2, *such a system could even become an income generator*. The majority of the participants also indicated that it is necessary that a new GRDMS to be put into place for HartRAO, because in doing so, HartRAO will contribute to South Africa's pursuit of national and international cutting-edge research; knowledge production and exchange; the provision of access to research infrastructure; and the support of human capacity development.

5.5 Theme 4: Data standardisation

Data standardisation is the process of changing data into a common format that allows for collaborative research, large-scale analytics and the sharing of research tools and methodologies (Observational Health Data Sciences and Informatics 2018). One main category relevant to Theme 4 was identified namely, ‘Data Structure and Organisation (DSO)’, which deals with the structuring and organisation of scientific data. Key data related to this category and its identified subcategory is presented in Table 5.13.

Table 5.13: Data structure and organisation (DSO) breakdown

Main category	Subcategory
Data structure and organisation (DSO) can be viewed as the process of standardising data by using predetermined standards, data structures and data organisation tools (Cormen, et al. 2009).	**Digital object identifier (DOI) and file naming convention (FNC)** are data structuring and organisation tools used to locate datasets on computer systems.

Data structuring and organisation are important considerations during the design of a new DMS. The goal of data organisation and structure is to create the technical environment that allows for the development of systems that are cost-effective to implement; flexible to construct; and easy to use (ENotes 2018). Without proper data structures, data retrieval and usage can be tedious and time consuming. Linked to this category and subcategory of data structuring and organisation, participants were asked to answer the following two questions:

- The first question focused on how data should be structured in the GRDMS. This question was asked to determine whether participants have knowledge regarding data structuring and organisation within a DMS and data structuring and organisation that can be leveraged by the designers of the new GRDMS.

- The second question focused on data structure guidelines and tools to be considered in the design of the GRDMS for HartRAO. This question was asked to solicit recommendations for possible data structure and organisation guidelines and tools, which could be employed in the design of the new GRDMS.

With regards to the first question, five participants (P2, P4, P5, P9 and P10) indicated that data should be structured and organised in standardised, scalable, hierarchical structures according to technique, station, frequency of observation scan, year, day-of-year, experiment name and data types (e.g. NGS card files, auxiliary files, metadata, etc.), filename and compression format. Participant P1 indicated that data should be structured and organised according to international standards, similar to those used by the CDDIS. Two participants (P7 and P8) were of the opinion that data should be organised in directories, folders and file structures. Participants P3 and P6 were not sure about the structuring and organisation of data, with P3 giving consideration to data in table format.

The fact that participants mentioned various ways of structuring and organising data is supported by the literature of Trauth & Sillmann (2013) and Noll (2010), indicating that there are many ways in which data can be structured and organised. For example, in databases, data can be structured and saved in the form of folders and directories, which are a form of hierarchical structure. Directories can be classified as main/root directories containing subdirectories. According to the literature of Trauth & Sillmann (2013), Noll (2010), Tenopir, Birch and Allard (2012) and Fox and Harris (2013:WDS3), data can be organised into different categories, such as observational, experimental, simulated, derived or compiled data and reference or canonical data. Noll & Michael (2016) suggest that data in an RDMS should be organised as files in main/root directories according to technique type, processing type, data frequency, station, year, day-of-year (the day on which the actual observation was conducted), filename and compression format. At the CDDIS, for example, geodetic VLBI data is organised and stored in subdirectories in format type SINEX and GNSS data is stored in top-level directories with various subdirectories consisting of folders and subfolders in format type

RINEX (CDDIS 2015). Directories are identified by a unique directory name and files of the same nature are usually stored under the same directory.

According to HartRAO (2012), geodetic data can be stored according to one or more directory name/s, device name (e.g. database, disk, etc.), format, instrument type, day-of-year, filename and compression format. The designers of the new GRDMS need to consider the various ways in which data can be structured and organised. The literature of Noll and Michael (2016) provides valuable information that may serve as a guideline.

In terms of the second question, four participants (P1, P5, P7 and P10) recommended guidelines similar to those of the CDDIS for structuring of data. Participant P5 also mentioned the NRF guidelines, which relate to the rating of scientist and their scientific output. Two participants (P5 and P10) suggested the use of data structuring tools such as DOIs and FNCs. Participant P10 also indicated metadata standards as a tool for structuring metadata. Participant P9 recommended using guidelines of hierarchical standardised structures, while participant P2 was of the opinion that data structuring will be determined by data requirements. Four participants (P3, P4, P6 and P8) indicated that they have no knowledge of any guidelines that could be considered.

Potential guidelines and tools for data structuring identified by the participants align with those found in the literature of the Federal Agency for Cartography and Geodesy (2013), the CDDIS (2015) and Riley (2017), indicating that there are various tools and guidelines that can be used for structuring of data, such as DOIs, FNCs, metadata and ORCIDs. DOIs are becoming increasingly important for the retrieval of large volumes of dispersed data (Simmhan, Plale & Gannon 2000). In recent years, the CDDIS has implemented DOIs as a means of organising and retrieving technique-specific geodetic data and data products (Noll 2016:136).

File-naming conventions (FNCs) are used to identify network stations, as well as names/titles of observations conducted at various network stations and file formats. FNCs provide sufficient information in the filename for data files to be transformed from one file format to another with no additional information required (VLBI.org 2009). Both the IERS and the CDDIS are using the ‘8.3.Z’ FNC to organise SINEX and RINEX data (Federal Agency for Cartography and Geodesy 2013 and CDDIS 2015).

The CDDIS and UNAVCO guidelines provide information about the types and standards of GNSS data that are accepted for their archives. – The CDDIS follows IGS standards outlined in the IGS Site Guideline documentation (CDDIS 2017), while UNAVCO follows their own data policy (UNAVCO 2012). According to Hogrefe & Stocks (2011), metadata schemas can be used to enhance the retrieval of data within DMS and can assist to ensure continued data retrieval and access. The authors indicate that data collected in accordance with guidelines set out in established metadata standards are not only more easily shared, but studies relying on such data is more easily funded.

Another tool that can be used to link datasets to its originator is ORCIDs. According to Clark (2012), ORCIDs can be used to attribute research reliably outputs to their true author or originator by assigning a machine-readable, 16-digit unique digital identifier to each author or originator of the data. The new GRDMS needs to accommodate data tools and structuring guidelines, such as DOIs, FNCs, metadata schemas and ORCIDs. These tools and guidelines are similar to those used by the CDDIS, IVS, ILRS and UNAVCO should be considered for the design of the GRDMS. These tools and guidelines are also being used by the global scientific community.

Different relationships between the responses to the question, “How should data be structured and organised in a geodetic research data management system?” and responses to the questions asked in main Category 3: RDMS. Sub-category 1: Characteristics and Sub-category 2: Criteria have been observed. The participants’ responses to the question on how data should be structured and organised in a GRDMS are a form of characteristic and criteria in its own way and can serve as such. The responses provided by the participants in connection with data structuring and organisation guidelines and tools are interlinked and homogeneous in nature. The designers of the new system could consider the findings of this study as characteristics and criteria for a data management system. There also seem to be an alignment between data structuring and organisation guidelines and data structuring and organising tools such as, DOIs and FNCs.

5.6 Theme 5: Data management system architecture

Data management system architecture/conceptual models are viewed as sets of rules and methods that describe the functionality, organisation and implementation of computer systems (Jaakkola & Thalheim 2011). These are tools that can guide the designers in structuring the design of a system(s). Detail on the data architecture/conceptual model was discussed in Chapter 2 to provide a framework of various DMS architectures that may be considered during the design of the GRDMS. One main category relevant to Theme 5 was identified namely, ‘Conceptual Models’. Key data related to this category are presented in Table 5.14.

Table 5.14: Conceptual models breakdown

Main category
Conceptual models are formal descriptions and representation of a system, organised to explain the structure and behavior of a system(s) (Jaakkola & Thalheim 2011).

Participants were asked to answer one question relevant to this category, namely whether they were aware of any conceptual models that will be suitable for and should be considered in the design of the new GRDMS for HartRAO. This question was asked to determine whether the participants were familiar with any suitable conceptual models that the designers may consider as a blueprint in the design of the new system.

Six participants (P1, P4, P5, P6, P9 and P10) indicated that they do possess knowledge of conceptual models that may be suitable for and should be considered as blueprints in the design of the HartRAO GRDMS. Three participants (P1, P5 and P6) suggested that the conceptual models of both the CDDIS and UNAVCO may be suitable, while another three participants (P4, P9 and P10) indicated that the structure of only the CDDIS should be considered. Four participants (P2, P3, P7 and P8) indicated that they do not have knowledge of such models. Examples of suitable conceptual models can be found in the literature by Wier, Boler and McWhirter (2012), Mashaba et al. (2016) and Noll

(2017). The conceptual models illustrated in Figures 2.3, 2.4 and 2.5 present the various system architectures, components and the interaction between units within these systems.

5.7 Summary and interpretation of findings related to the GRDMS

To demonstrate how the key themes (discussed in the previous sections) address the research objectives, a diagram is provided in Figure 5.4. Themes one to five is relevant and aligns to all the research objectives. The greater the weight with which the key theme maps onto the research objective, the wider the arrow pointing from key theme to the research objective.

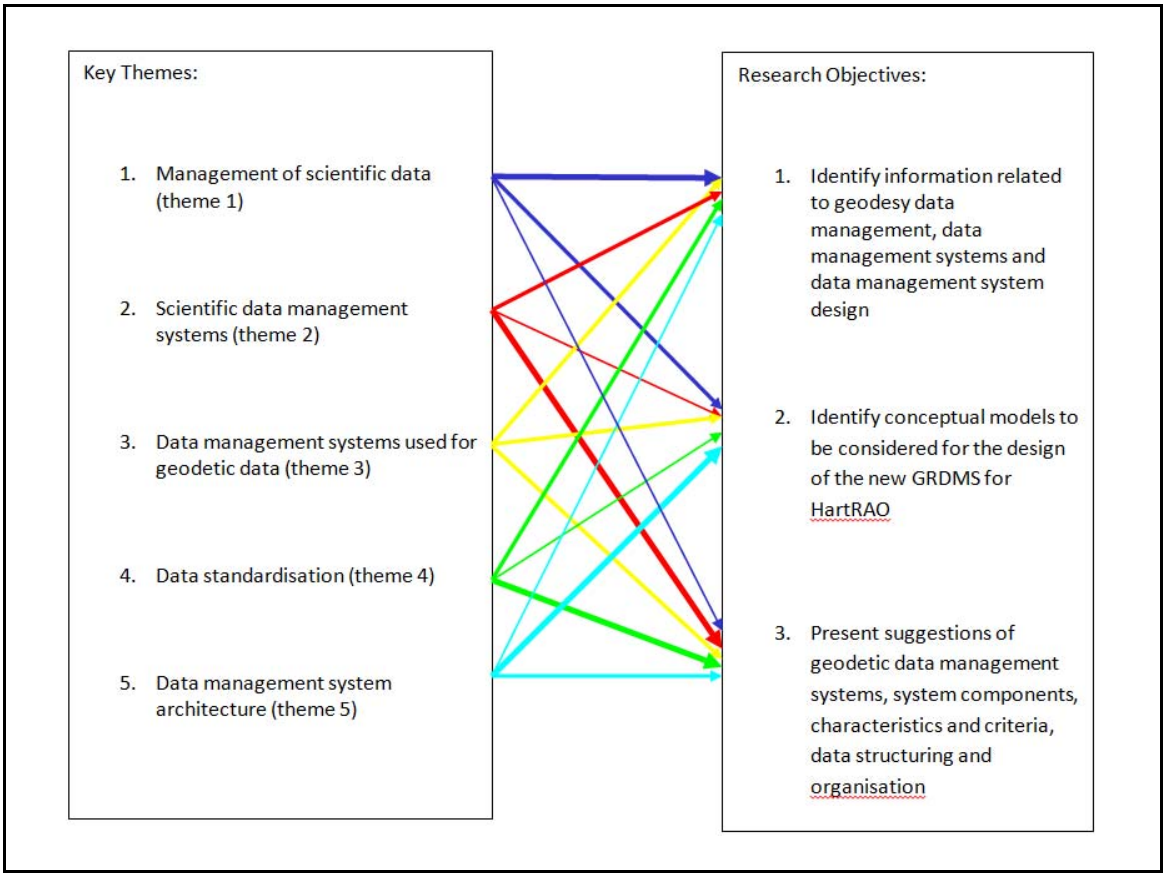

Figure 5.4: Mapping key themes to research objectives

The research participants (i.e. HartRAO researchers and students) perceive the need for implementing a new GRDMS at HartRAO as imperative, in order to ensure compliance with current and future technological expansion and institutional needs at

HartRAO. The current HartRAO DMS is unable to provide all technique-specific geodetic data. Data from new technological components are inaccessible via the current system. A new GRDMS, able to cater for extended capabilities and the increase in data volumes from extended geodetic instrumentation, is required for HartRAO.

Researchers, students and international data service providers, such as the CDDIS, are viewed as the main stakeholders in the design of a new GRDMS for HartRAO. Currently, participants make use of either an RDMS or a platform to manage their data. The choice of RDMS is mostly determined by availability, suitability, ease of use, data processing and analysis capabilities and/or deficiencies of the current DMS at HartRAO.

Participants search for space geodesy data on local HartRAO and remote repositories, servers and websites. They access HartRAO geodetic datasets by logging onto the HartRAO Geoid server or by accessing the HartRAO Space Geodesy programme webpage, as well as websites of international data service providers, such as the CDDIS, IVS, IGS, ILRS, IERS and UNAVCO and by using URLs, ftp and http. Members of the international scientific community also access data holdings and repositories of these international data service providers via their websites or http file construction to search for geodetic data specific to their field of study, including HartRAO datasets.

In general, participants found these databases to be easily accessible. Only a few of the participants found it to not be user-friendly or made difficult by terminology used by web designers. Most participants also found the way in which data is presented in these databases easy to understand, although some participants indicated that prior knowledge of the subject discipline may be required to ensure easy retrieval of data. Databases of global data service providers, such as those of the CDDIS, provide interactive menu-driven graphic user-interfaces that are more effective than that offered by the current HartRAO system. Depending on their experience, familiarity with and knowledge of a particular DMS, participants are able to retrieve the required data within minutes. The amount of time it takes participants to retrieve data also depends on factors such as familiarity with the scientific discipline, the type of data being retrieved, and the tools and systems being used.

Participants and members of the global scientific community make use of various technique-specific space geodesy data in their research. The researchers and students of HartRAO obtain and mostly use raw and processed GNSS (RINEX format); meteorological (compressed text file format); VLBI (NGS card file format); SLR (SINEX format); and also other geodetic data in various formats. Although participants listed various types of data that they wish to be included in the new GRDMS, they specifically mentioned geodetic data generated by HartRAO instruments. International data service providers, such as the CDDIS, supply various types and formats of geodetic technique-specific data to the international scientific community. Their databases are able to manage raw, processed and analysed data, homogeneous and heterogeneous data, ancillary and auxiliary data, data products, metadata, data usage information, as well as project management data, all of which should also be considered for and included as key components of the new GRDMS.

Participants make use of various search types/terms when searching for datasets with keywords, as well as station and experiment names featuring prominently. This aligns with practices applied by global data service providers, such as the CDDIS, which allow users to search by keywords, station, experiment and observation names, session codes and year and day-of-year.

Participants require data to be structured and organised according to standardised scalable hierarchical structures, techniques, stations, frequency of observation or scan, year, day-of-year, experiment name and data type, filename and compression format. This allows for quick and easy retrieval of various data types and formats and data products. International data service providers' structure and organise data according to categories of data; using specifying paths and device names; in directories; subdirectories, folders and files according to technique type; processing type; data frequency; station; year and day-of-year; filename; and compression format. Participants recommended the use of data structure guidelines as provided by the NRF and CDDIS, as well as data structuring tools, such as metadata, DOIs and FNCs. International data service providers, such as the CDDIS, make use of DOIs, FNCs, metadata and ORCIDs for organising and retrieving technique-specific geodetic data and data products. The CDDIS and UNAVCO provide guidelines about the types and standards of data that are accepted for their archives.

From their own experience in research and computer usage, participants are familiar with the components, characteristics, objectives and criteria an RDMS should possess. The participants recommend that the new modern GRDMS should be capable of handling large volumes of data and that it should, therefore, include components for large stable storage (clusters of servers, SSDs) and also fast networks (large bandwidth) and fast processors. It should be in line with the CDDIS system, consisting of hardware and software components for receiving, structuring, storing and disseminating data and also with that of UNAVCO GSAC, providing an application server and repository and data transfer servers.

Key criteria identified by the participants and indicated in the literature, which should be considered in the design of the new GRDMS, include an up-to-date system that is easily accessible, searchable and secure; has a large capacity for data storage; provides accurate reliable data; makes use of metadata standards and tools to structure and organise data subsets; and of which the architectural structure and processes are mirrored on the CDDIS and UNAVCO GSAC systems. Participants further require an RDMS with interaction/interoperability capabilities. Regional data providers, such as HartRAO, interact with international data service providers, such as the CDDIS and UNAVCO, by transferring data generated by its geodetic instruments to these data service providers.

Components, characteristics, objectives and criteria of a RDMS are not unique to one system but apply to most systems functioning in the realm of space geodesy. Some of the participants indicated that they have sufficient experience in DMS design to be able to make a significant contribution to the design of a fully functional new GRDMS. Conceptual models such as those of the CDDIS and UNAVCO GSAC systems appear to be the most suitable and relevant systems to be considered as models for the design of HartRAO’s new GRDMS.

5.8 Conclusion

This chapter focused on the reduction, categorisation, analysis, interpretation and presentation of empirical data. Qualitative raw data of the interviews conducted with the participants are presented as transcriptions in Appendix B. Interpretative qualitative analysis of the raw data was conducted. The raw data were divided into themes and categories relevant to the design of a new GRDMS for HartRAO. Subcategories, directly linked to the main themes and categories, were reviewed and discussed in the context of the interviews and the literature reviewed. Findings were presented as tables, figures and in discussions. Linkages and alignments to the literature reviewed (Chapter 3) were also discussed. The findings are used to guide the design of the new GRDMS.

Chapter 6 encompasses the conclusions and recommendations of this study. Achievement of the study objectives is indicated; conclusions and recommendations are made; and future research possibilities are presented.

CHAPTER 6: CONCLUSIONS AND RECOMMENDATIONS

6.1 Introduction

This study aimed at investigating what has been reported on the topic in the body of published knowledge and at soliciting information from people actively involved in geodetic research at HartRAO. To develop a contextual framework and to propose a conceptual model to be considered for the design of the new GRDMS for HartRAO, which will comply with technological expansions at HartRAO while meeting the requirements of local users, international collaborators, data service providers and the general public.

To achieve the aim of this study, a case study was conducted with data being collected through an extensive literature review and with one-on-one interviews. In Chapters 2 and 3, the components, characteristics, criteria, scientific data structuring and organisation and suitable conceptual models to be considered for the design of a new GRDMS for HartRAO were described, as identified in the real world and in the literature. The empirical exploratory research methodology followed in the study was described in Chapter 4. Participant responses to interview questions and the findings from the literature reviewed, with respect to the themes relevant to the design of the GRDMS, were presented in Chapter 5.

The purpose of this final chapter is to discuss the findings; present conclusions; and provide recommendations. The chapter concludes with a discussion of the alignment towards achievement of the study objectives, key recommendations (including a conceptual model for a GRDMS) and a proposal for future research.

6.2 Alignment towards the achievement of research objectives

To ensure that the research conducted is in line with the research objectives set for this study, this section aims to demonstrate how the information obtained from participant interviews and literature sources aligns with the objectives of the study and matching with the research questions. A summary of this alignment is provided in Table 6.1.

Table 6.1: Research objective and question alignment

Research objectives	Research questions	Alignment
Identify information related to geodesy data management, data management system and data management system design by means of data collection techniques.	What type of information in connection with geodesy data management, data management systems and data managements system design should be considered for designing the GRDMS?	Details of key information are discussed in Chapter 3. The findings of the literature reviewed and in-depth interviews are discussed in Chapter 5. Key information to be considered for the design of the GRDMS is provided in Sections 6.3.2 and 6.3.3.
Identify conceptual models to be considered for the design of the new GRDMS for HartRAO by means of data collection techniques.	Which conceptual models should be considered for design of a GRDMS for HartRAO?	Relevant conceptual models identified for achieving the objectives set for this study are discussed in Chapter 3. The findings (i.e. the three most suitable models) of the literature reviewed and in-depth interviews are discussed in Chapter 5. Various subsystems and data flow within the system are illustrated in Sections 6.3.2 and 6.3.3.
Present suggestions of geodetic data management systems, system components, characteristics and criteria, data structuring and organisation, relevant to the design of the new GRDMS.	Which geodetic data management systems, system characteristics, criteria, including data standardisation (structuring and organisation) should be considered for the design of the new GRDMS?	Details of the key information are discussed in Chapter 3. The findings of the literature reviewed and in-depth interviews are discussed in Chapter 5. Key information to be considered for the design of the GRDMS is provided in Sections 6.3.2 and 6.3.3.

In summary, the research objectives set for this study were matched to the research questions. Key information relevant to geodesy data management; data management systems (including system components, characteristics, criteria, data structuring and organisation); data management system design; and conceptual/architectural models was in alignment with the research objectives and addressed all the research questions.

6.3 The proposed GRDMS design model

Six key aspects, linked to findings from the literature review and participant responses with respect to the themes identified, are to be considered in the design of a new GRDMS for HartRAO. As indicated in Figure 6.1, these key aspects include: the GRDMS overall design, infrastructure architecture, data ingress, archiving, metadata retrievability and system front-end.

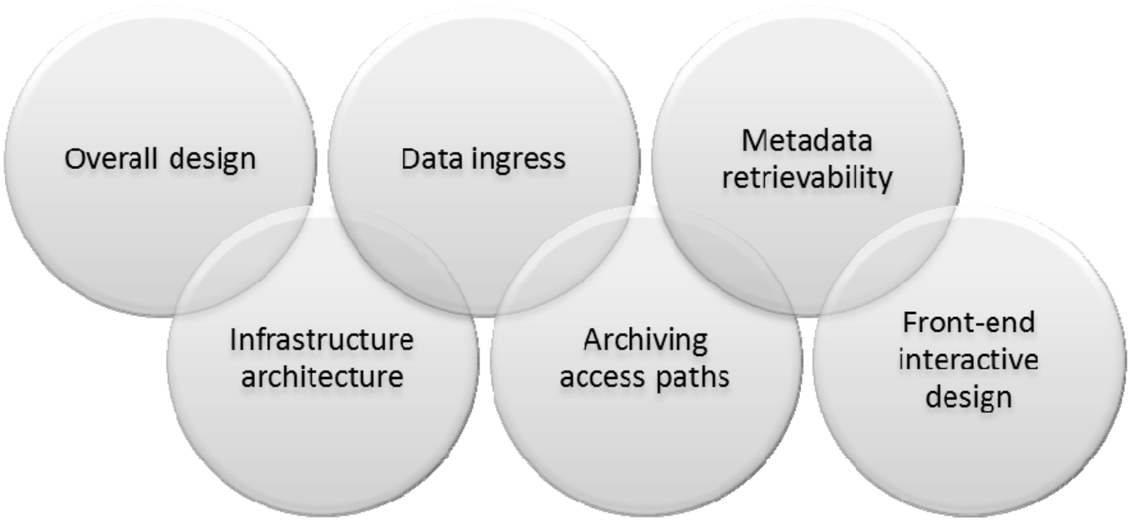

Figure 6.1: GRDMS design components

6.3.1 Overall design

It is recommended that a GRDMS is designed to cater for all of HartRAO's technique-specific geodesy data (Table 6.2) that is able to serve various types of users, but specifically researchers, students and global data service providers.

Table 6.2: Recommended HartRAO geodetic technique-specific data types

HartRAO technique / Instrument	Data type
NASA SLR (Moblas-6), LLR & Russian SLR (Roscosmos)	CPF files & compressed text files
VLBI products	NGS card and FITS files
GNSS	RINEX and SINEX files
Seismic	Seedlink records
Gravimetric	Compressed text files
Meteorological	Compressed text files
Tide gauge	Compressed text files
DORIS	RINEX files
Skycam	Images & videos

The new GRDMS should contribute towards establishing HartRAO as an important geodetic data service provider – the only one of its kind in Africa – and a vital link in a network of international data service providers. A new GRDMS that can cater for African geodetic datasets in a central data bank, using standardised data structures and organisation, is imperative to HartRAO's researchers and students, national and southern Africa's geodetic research community, regional and international geodetic data service providers and the global geodetic research community. The new GRDMS will have to comply with the strategic objectives of HartRAO and the mandate of the NRF, which is to:

> *"support and promote research through funding, human resource development and the provision of the necessary research facilities in order to facilitate the creation of knowledge, innovation and development in all fields of science and technology, including indigenous knowledge and thereby to contribute to the improvement of the quality of life of all the people of the Republic" (NRF, 2017:12). The strategic objectives of HartRAO are governed by the mandate of the NRF to "optimise South Africa's contribution to, and benefit from, the international Square Kilometer Array (SKA) project; establish and sustain globally competitive and transformed radio astronomy and space geodesy research and infrastructure in South Africa and abroad, where appropriate; optimize the associated national socio-economic benefit from radio astronomy and*

space geodesy activities; and promote radio astronomy and space geodesy in Africa" (National Research Foundation 2017:99).

The strategic objectives of HartRAO are governed by the mandate of the NRF to "optimise South Africa's contribution to, and benefit from, the international Square Kilometer Array (SKA) project; establish and sustain globally competitive and transformed radio astronomy and space geodesy research and infrastructure in South Africa and abroad, where appropriate; optimise the associated national socio-economic benefit from radio astronomy and space geodesy activities; and promote radio astronomy and space geodesy in Africa" (National Research Foundation 2017:99).

It is recommended for participants in this study to have experience in DMS design and to collaborate in the design of the new system. Their experience in system and software design can be leveraged to the benefit of a successful new GRDMS. Their input would be beneficial to ensure that the GRDMS complies with requirements and needs of end-users.

The main objectives of the new GRDMS is to receive, organise, structure, archive and store geodesy data and data products in a central data repository; disseminate data and data products to international data service providers; maintain metadata and provide matrixes for performance reporting; and to provide remote, secure, multi-user access and multiple data access techniques (e.g. via URL, http or ftp) via a user-friendly interactive graphical user interface to a growing research community (Coetzer, Botha & Jacobs 2018). The system should be divided into functional units (components) operating relatively independently of one another. This will assist in avoiding critical overall failures and will also enable faster repair of broken/damaged components. The proposed functional units are described in more detail in Section 6.3.2. Furthermore, to avoid and prevent data loss, system downtime and disaster recovery, it is recommended that a clone of the complete GRDMS be housed at a different location.

6.3.2 Infrastructure/Architecture

It is recommended that both the conceptual models of the CDDIS and UNAVCO GSAC be used as blueprints for the conceptual design of the new HartRAO GRDMS. (Findings in support of this recommendation are discussed in Chapter 5). In accordance with the conceptual models of the CDDIS and UNAVCO GSAC systems, which consist of various subsystems and units, the recommended architecture for the new HartRAO GRDMS is illustrated in Figure 6.2.

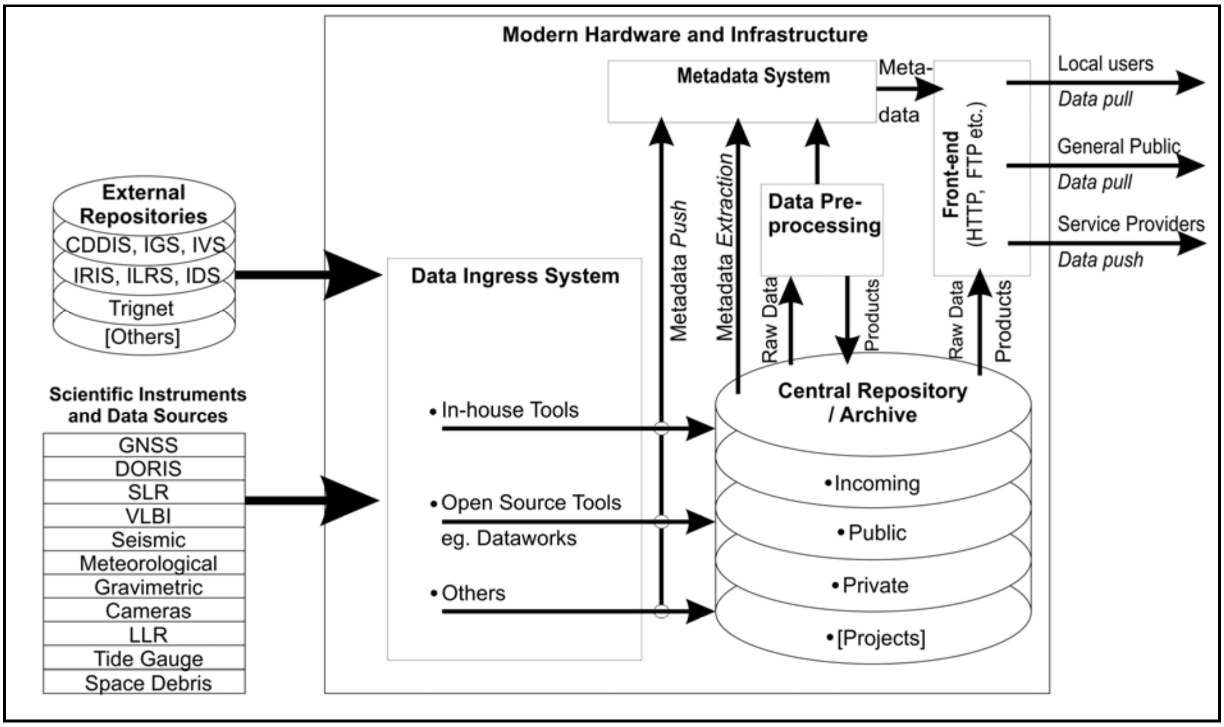

Figure 6.2: Conceptual model of HartRAO's GRDMS illustrating the various subsystems and data flow

With reference to Figure 6.2, the five functional units within the proposed *Modern Hardware and Infrastructure* system of the GRDMS that need to be considered, as well as the flow of data within these units, are outlined in the following sections.

6.3.2.1 Data ingress system/unit

The *Data Ingress System/Unit* will be a virtual machine with Linux operating system software. Its main functions will be the handling of data collection from the various stations, external repositories (e.g. CDDIS, IVS, etc.) and HartRAO's geodetic instruments (including data sources in Africa). This unit will conduct quality checks to verify the integrity and quality of the data and will be pushing the raw data to the *Central*

Repository/Archive Unit (Server), and the metadata to the *Metadata System/Unit.* It will achieve this by using in-house software tools, open source software tools, etc.

6.3.2.2 Central repository/archive Unit

The *Central Repository/Archive Unit* will be a server consisting of four subsystems. These subsystems will handle the incoming, public, private and project data. Within these subsystems, data will be structured and organised according to recommended directory structures and FNCs. A detailed discussion of these directory structures and FNCs is provided in Section 6.3.4. The *Central Repository/Archive Unit* will push raw data to the *Data Pre-processing Unit* for processing and will store raw and processed data. Metadata will be extracted and together with raw and processed data, as well as data products, will be available to local users, international collaborators, data service providers and the general public via the system *Front-end.*

6.3.2.3 Data pre-processing unit

The *Data Pre-Processing Unit* will consist of a Linux virtual machine with Space Geodesy PROCessing on Linux (SGProcL) software and a Microsoft Windows virtual machine with Space Geodesy PROCessing on Windows (SGProcW) software installed on it. According to Coetzer, Botha & Jacobs (2018), the SGProcL and SGProcW software is suitable for and compatible with the type of data processing required. Within this unit, some of the raw data (e.g. raw GNSS data) received from the *Central Repository/Archive Unit* will be pre-processed to form data products such as tectonic motion time series plots (HartRAO 2012).

Metadata will be extracted and stored in the *Metadata System Unit* and will be available to local users, international collaborators, data service providers and the general public via the system *Front-end.* Pre-processed data and data products will be pushed to the *Central Repository/Archive Unit* for storage. Local users, international collaborators, data service providers and the general public can access the data and data products via the system *Front-end.*

6.3.2.4 Metadata system/unit

The *Metadata System/Unit* will be virtual machine with Dataworks for GNSS software installed on it. This unit will extract, capture and store various types of metadata (e.g. metadata of raw and processed data, system operating metadata and user access metadata). Pre-requisite metadata of raw and pre-processed data and custom processed data products will be available to local users, international collaborators, data service providers and the general public via the system *Front-end.*

6.3.2.5 Front-end

The *Front-end* will consist of a number of virtual machines for different types of front-ends. Subsystems, such as query and output handlers, will be installed on these machines. The main functions of the different front-ends will be to provide access to raw, pre-processed data and custom processed data products, as well metadata to local user, international collaborators and the general public via a URL, http/website or ftp. The query and output handlers will be responsible for handling user queries and output formats according to user needs.

6.3.2.6 Data cycles

With reference to Figure 6.2 and relating to the data flow through the various subsystems of the proposed new GRDMS, logical processes within the data cycle to be considered for the new GRDMS include internal, external and custom data cycles:

- The **internal data cycles** consist of the receipt of raw data streamed to HartRAO from the various geodesy instrumentation and data sources. The *Data Ingress System/Unit* will be responsible for the receipt and ingress of the raw data. In this unit, raw data will be quality checked with in-house tools. The data will then be pre-processed into required file types, formats and filing structures. Quality-checked and preprocessed data will be pushed to the *Central Repository/Archive Unit* and stored within the different domains of the *Central Repository/Archive Unit.* Some quality-checked and pre-processed data will be sent to the *Data Pre-processing Unit* for processing into predetermined basic data products. Data that do not require processing remains in the *Central Repository/Archive Unit*. Predetermined basic data products will be sent to the *Central Repository / Archive Unit* for storage and archiving according to different domains. Throughout the entire process metadata will be extracted and stored in the *Metadata System/Unit.*

- The **external data cycles** consist of the process of retrieving data and metadata stored in the *Central Repository/Archive Unit* and the *Metadata System/Unit.* Local users, international collaborators, data service providers and the general public can request data stored in the *Central Repository/Archive Unit* via the system's interactive *Front-end* (http, ftp, etc.). Requested data will be pulled from the *Central Repository/Archive Unit,* packaged into a single compressed file and displayed on screen. For data service providers, e.g. the CDDIS and IVS, data will be pushed/transferred via the system *Front-end* to their receiving servers.
- The **custom data cycles** consist of the requests of data that are routine and do not require human intervention. These types of requests will be handled automatically by the *Modern Hardware and Infrastructure* system. Local users, international collaborators, data service providers and the general public can send their request to the GRDMS via the interactive *Front-end* of the system (i.e. http or ftp) or via emailing facilities. Requests will be translated into a script. Requests for pre-processed data and/or products will be pulled from the *Central Repository/Archive Unit.* Subsequently, the results will be sent to the requesting user via the *Frond-end* of the system or via email.

6.3.2.7 GRDMS components

Linked to Figure 6.2, the GRDMS will use modern technologies, such as centralised storage; large storage technology (e.g. clusters of servers, SSDs, redundant array of independent disk (RAID) arrays); CPUs; fast processors; fast networks with large bandwidth; firewalls; and servers with real-time casters (i.e. software that handles real-time data streaming (CDDIS 2015)). The GRDMS will consist of components that will support ease of access to various different geodetic data types (e.g GNSS tools to handle downloading of GNSS data to the local repository ('/archive/public/gnss')); a local repository with subsystems; and virtual machine technology (e.g. Microsoft Windows virtual machines and Linux virtual machines). In addition to the scientific software (e.g. SGProcL, SGProcW, Common Astronomy Software Applications (CASA), Vienna VLBI Software (VieVs), etc.), Dataworks for GNSS software will be installed on the virtual machines.

Dataworks for GNSS is open source software modules designed by UNAVCO and used by the UNAVCO GSAC (see Chapter 3). It can be used by small GNSS networks (e.g. tens of stations) to handle data streaming, ingress, formatting, renaming, archiving, storage, downloads and the submission/transfer of geodetic data to collaborators. In addition to fundamental data management, UNAVCO's Dataworks for GNSS software is recommended because of its capability to manage large volumes of GNSS data acquisition and sharing. Its GNSS modules provide: a database schema for metadata; tools for handling metadata (e.g. DOIs, FNCs, metadata standards, etc.); ingest software that can manage the metadata of incoming data file; scripts to manage actual data files; scripts for mirroring station data and metadata; and software to manage documentation (Boler et al. 2015).

Dataworks for GNSS's mirroring and federation among multiple networks modules are important because of the data service collaborations between HartRAO GRDMS and the other geodesy data service providers. The federation module allows each data and metadata collection to be kept distinct. Due to its repository data serving capabilities (typically ftp), Dataworks for GNSS software will be used to create Web services. Users will be able to use queries and the Web interface to discover and retrieve data without needing to know which data service provider holds any particular part of the overall collection.

Dataworks for GNSS software's query-handling Web services application programming interface (API) (i.e. a set of functions, protocols, tools and procedures (Webopedia 2018)) allows for the creation of applications that can access the data on the operating system (e.g. Linux or Windows operating systems). Application Programming Interfaces (APIs) are designed to accept formatted HTML requests and, in combination with the GRDMS Web graphic user interface, will allow users to search the GRDMS Repository/Archive and download files, etc. from a command line, making for a very user-friendly system.

In support of the retrievability, structuring and organisation of data and metadata within the GRDMS, tools such as DOIs, standardised vocabulary and FNCs will be applied. Each dataset will receive a DOI. Users will be able to use DOIs to retrieve datasets from the database. The systems interface will allow users to use standardised vocabulary to conduct keyword searches and to support the compatibility and transfer of data-to-data service providers. FNCs similar to those used by the major geodesy service providers such as the CDDIS, IVS, ILRS, etc. will be used. The tools are described in more detail in Sections 6.3.3–6.3.5.

6.3.3 Data ingress

The GRDMS should be designed to provide complete, up-to-date datasets for all geodetic instruments/techniques with the use of virtual machine technology (discussed in Section 6.3.2). The new GRDMS should cater for all HartRAO's geodetic technique-specific data types and formats be it raw, pre-processed or processed data, data products, ancillary data and metadata. In line with Noll (2017), it is further recommended that the system automatically peruses incoming data; verifies content quality; extracts pertinent metadata; moves files to appropriate archive directory locations/repositories; monitors and manages files; manages and executes utilities software; and retrieves deposited files. This ensures that the system is fully automated and reliable (less system failure), with less human interaction that may cause errors in consistency and repeatability of results. Special scripts and program routines will check incoming data; verify format and readability; and upload the files to the GRDMS. Metadata will be extracted from the incoming data files and will automatically be stored in the metadata repository. Automated archiving routines will peruse directories and move new data to the appropriate disc area. These routines will migrate the data based on file names to the appropriate directory.

Raw and pre-processed data and data products have to be stored in a well-organised and structured manner to cater for versioning, retrieval and access. HartRAO is a regional geodetic data service provider that collaborates in providing geodetic datasets to global data service providers (e.g. the CDDIS and IVS). Therefore, it is important for the data be stored in a manner compatible with international data service provider standards. To ensure effective transfer of data between HartRAO and global data centres, data need to be in the same format as that of the other data service providers. This supports speedy uploading/transfer of datasets between the various data service providers.

The data need to be archived and stored according to predetermined directory structures. An intelligent data structure translator, which will cater for all standardised structures, similar to those used by the international geodetic data service providers (one copy – many access paths), is recommended for the new system. The aim of the data structure translator will be to transform a program written in a given programming language such as C++ into a functionally equivalent program (the target language) such as Java or Python (Pring-Mill 2018). This data structure translator will ensure that the functional or logical structure (the intrinsic nature of each program) is not lost (Pring-Mill 2018). Based on the data structuring and organisation method used by the CDDIS, technique-specific geodetic data is to be structured and organised in a scalable hierarchical structure, from top-level/root directories containing various subdirectories with folders, subfolders and files, to individual datasets. A hierarchical directory structure is recommended for structuring files by technique (e.g. VLBI, GNSS, S/LLR, DORIS, etc.), with the final level containing the actual data files.

The following is an example of the recommended structure for raw technique-specific formatted data: */data/technique/type/station/frequency/year/DoY/filename.compression*. A description of the recommended structure is provided in Figure 6.3.

Descriptor	Description	Example value(s)
technique	technique abbreviation	gnss, slr, vlbi, doris, gravity, seismic etc.
type	data type	RINEX, SINEX, NGS card files etc.
station	technique-specific station code	HRAO, MATJ, Hh, Ht etc.
frequency	file frequency	daily, hourly, high-rate
year	Gregorian year	e.g. 2018
DoY	day-of-year	e.g. 028
filename	technique-specific	*see* Figure 6.6 for example
compression	un-/compressed, compression type	.Z, .gz, .zip

Figure 6.3: Recommended basic directory structure for the GRDMS archive

(Coetzer, Botha & Jacobs 2018)

With reference to Figure 6.3, it is recommended that technique-specific data and data products are structured within the root directory according to: processing type (e.g. raw data or data product), followed by the technique used; data type; station used; frequency of the observation; year, day-of-year (DoY); and filename and compression format. For the naming of files, it is recommended that the *8.3.Z* file naming convention (FNC) is applied in accordance with its use for the naming of technique-specific data files by the CDDIS. When HartRAO needs to share data files with international data service providers, or vice versa, the format of the files will be similar, which will increase the convergence of files among systems.

The new GRDMS needs to accommodate data versioning in order to remain compatible and effective. The new GRDMS will mainly focus on RINEX Version 2 data, but will also incorporate RINEX Version 3 data, which are similar to RINEX Version 3 data and that are currently being used by the CDDIS (Noll & Michael 2016). The following is an example of the recommended structure of a GNSS RINEX Version 2 filename using the *8.3.Z* FNC: *SSSSDDD0#.YY.Z* and an example of a composite filename for HartRAO geodesy data using the *8.3.Z* FNC: *hart0010.96d.Z*.

Description of the facets of filename in the example above is provided in Figure 6.4. The keys (descriptors) denote different facets of a composite filename.

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 uear	17
Z	un-/compressed	.Z

Figure 6.4: Example of the 8.3.Z file naming convention (FNC) for GNSS RINEX version 2 data

6.3.4 Archiving

Coetzer, Botha & Jacobs (2018) recommend all catalogues of data being stored in an online central repository to ensure long-term archiving of data (see Figure 6.2) to facilitate easy access, retrieval and use. This recommendation is supported by findings from the literature (Chapter 3) and participant responses (refer to Chapter 5).

HartRAO is a regional geodetic data service provider supplying geodetic datasets to global data service providers such as the CDDIS, IVS, scientific community and the public. Therefore, HartRAO's geodetic data should be archived in such a way that it would be compatible and shareable with international data service providers. For the structuring of data, data structures similar to those of the CDDIS are recommended.

A high level of caution needs to be applied when data is shared. The security of the entire system is a priority due to files containing information of high value to the organisation and scientific community. Because of security concerns, it is recommended that the GRDMS central repository/archive to have certain file protections in place to prevent security breaching (i.e. unauthorised access), infringements (i.e. intellectual property theft) and threats to the system as a whole (i.e. physical damage).

In order to protect data, it is recommended that data protection connections, such as Virtual Private Network (VPN) security, are implemented on the GRDMS to safeguard and protect data. According to the literature (WhatIsMyIp.com 2018), a firewall protects data while on the computer and a VPN protects data on the Web. VPNs use advanced encryption protocols and secure tunneling techniques, i.e. a protocol that permits secure propagation of data between networks (Techopedia 2018) to enclose online data transfers. This technology is used by corporations, organisations and the public to protect data. Even WiFi hotspots make use of this technology (WhatIsMyIp.com 2018).

Access to the data should be permitted or denied depending on several factors, one of which is the type of access requested. Access control limits who can access files and how they can access the files (Kaur, Singh & Kaur 2016:579). The most common access permission is that which allows the user to read the file but does not allow the user to modify the file (e.g. read-only permission). Only limited read-only or no direct access to mission-critical systems and the main data archive, should be allowed. Only authorised persons will be allowed access to mission-critical systems. The use of an online registration process is recommended in order to manage the use of data; protect the integrity of data; and to ensure the inclusion of security measures, particularly where private data storage is concerned. Username and password authentication should be implemented as a measure to safeguard the system and the data against misappropriation. It is recommended that the GRDMS provide a website with a landing page – i.e. a Webpage at which users arrive after clicking a hyperlink (Techopedia 2018) – where users can register to use the system and its content.

In light of the foregoing information, it is recommended that the GRDMS central repository/archive consists of the following structure and access permissions (Botha 2018):

- *Incoming domain*: for raw data obtained from our stations or from data services, read-only for local users and not visible publicly – */archive/incoming*;
- *Public domain*: for quality checked data that may be published publicly, read-only by public and local users – */archive/public*;
- *Private domain*: for data that must be kept private for a proprietary period, read-only locally and read-only user logins for authorised collaborators – */archive/private*; and

- *Projects/tools domain*: for data required in *GAMIT/GLOBK*[1] processing (automatic or ad-hoc), read-only for local users and read-and-write for authorised GAMIT/GLOBK users – */archive/[project/tool name].*

Another approach to protecting a file and its content from unauthorised access is to use a password with each file. If a user does not have a password or does not know the password associated to a file or directory, then he/she will not be allowed to access the system (Kaur, Singh & Kaur 2016:579). Therefore, it is recommended that the designers of the GRDMS consider password authentication, as it is a very effective way of protecting files to safeguard the system against foul play.

File naming is also a form of file protection (see Chapters 3 and 5) that can be applied in a system. A filename is attached to every file, so as to identify it uniquely and allow access to it through its unique name (Kaur, Singh & Kaur 2016:579). The use of the CDDIS *8.3.Z* FNC is recommended for the GRDMS to ensure compatibility, share-ability and interoperability between the GRDMS and international data service providers.

6.3.5 Metadata retrievability

With reference to Figure 6.2, metadata is collected from all the various units of the proposed GRDMS. UNAVCO Dataworks software modules (discussed in Section 6.3.2) are recommended for the management of metadata. According to Figure 6.2, metadata received by the *Data Ingress System/Unit*, are pushed to the *Metadata System/Unit* and extracted from the *Central Repository/Archive Unit (Server)* and the *Data Pre-processing Unit.* The *Metadata System/Unit* controls the collection of metadata.

It is therefore recommended that the *Metadata System/Unit* assigns and tracks all of the storage locations associated with each file, in order to direct file input/output (I/O) requests to the correct servers. It is also recommended that the *Metadata System/Unit* stores metadata such as filenames, directories, permissions and file layout. Storing the metadata provides an efficient division of work between computing and storage

---

[1] *GAMIT/GLOBK - collection of programs used for the analysis of GPS data and Kalman filter primary purpose of which is to combine various geodetic solutions such as GPS, VLBI, etc.*

resources (University of Tennessee Knoxville 2018). Files in the *Metadata System/Unit* must contain the layout of associated data files, including the metadata of the storage device and persistent interoperable identifiers.

For the creation of descriptive metadata stored by the *Metadata System/Unit,* the Dublin Core Resource Description and Access (RDA) standard, a standard for descriptive cataloging and access, designed for the digital era (British Library 2018), are recommended. Standards for resource description and access need to be incorporated in the new GRDMS to ensure that technique-specific geodetic data is described according to a set of guidelines and instructions for formulating data to support resource discovery. In support of resource/data discovery, it is recommended that datasets are described at all levels of aggregation and that persistent interoperable identifiers, such as DOIs, are assigned to each dataset to enhance searching and retrieval of datasets.

6.3.6 Front-end interactive design

It is recommended that the new GRDMS provides easy access to all HartRAO's geodetic technique-specific datasets (see Table 6.2), ancillary data, data products and metadata via a user-friendly graphic user-interface. An illustration of the current front-end interface of the HartRAO Geodesy Programme webpage is provided in Figure 6.5. Users can access limited technique-specific datasets by following the links on the webpage. This webpage was created to serve as precursor to the new GRDMS. Figure 6.5 displays the homepage of the HartRAO Space Geodesy Programme website with various search facilities. On the existing system, users can conduct keyword searches and/or click on hyperlinks to search for data and/or information. The webpage has several drawbacks, including lacking sufficient content; being outdated and incomplete; and lacking sufficient search/query facilities for users.

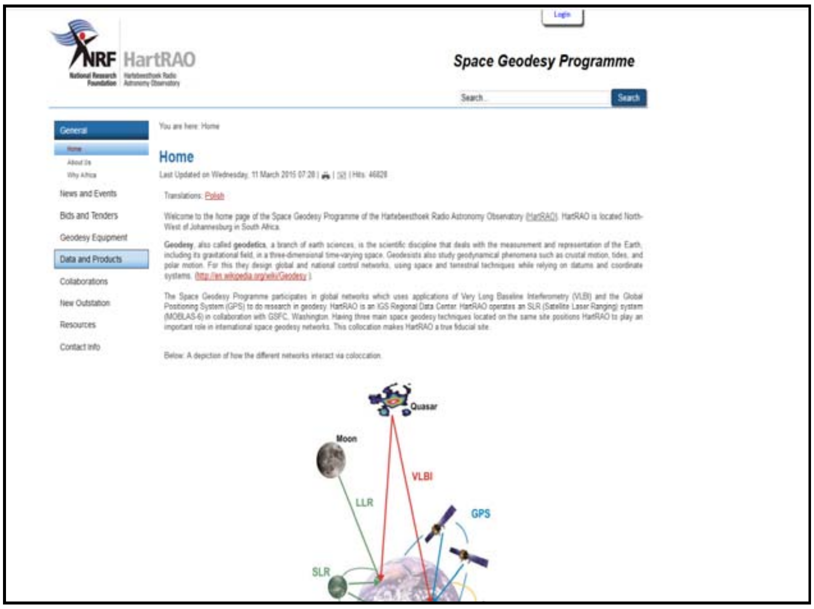

Figure 6.5: Front-end interface of the HartRAO Space Geodesy Programme webpage

(HartRAO 2015)

A new interactive graphic user interface is recommended for the website that will: allow for easy navigation of the system; provide effective searching/query facilities; allow for the use of various different search types/terms; display the data holdings in an easily understandable manner; provide hyperlinks to datasets; use active persistent identifiers; and provide help functionalities. It is also recommended that the designers of the new GRDMS incorporate a taxonomy consisting of familiar search terms to support keyword searching.

To cater for both technically skilled professionals and new users of the GRDMS, it is recommended that it should be possible to access the GRDMS repositories and archives via anonymous ftp and also via a browsing interface (URL or http). New users, both those familiar with space geodesy techniques as well as new research communities, prefer a browsing interface to access archive contents (Noll 2012). The new GRDMS interface should allow for browsing the archive, as users may require either current or historic datasets. A Web interface-based search tool that queries the GRDMS metadata is, therefore, required. In order for users to identify data and products of interest, the interface needs to allow them search criteria options based on temporal, spatial, target, site designation and/or observation parameters. The subsequent results of these searches should include a listing of sites or data holdings satisfying user input specifications.

Having an interactive front-end would aid users in the discovery of HartRAO geodesy data, products and information. The interface tool mentioned above can also assist staff in system archive management and could promote the HartRAO data holdings to a larger community (e.g. through metadata standards). Another feature recommended for supporting the discovery of HartRAO geodesy data, products and information is the inclusion of a map selection tool. The advantage of having this feature is that users can select a site/destination as the search type and the query would then yield a map and a list of valid sites that can be identified for retrieval. Figure 6.6 shows an example of this feature used as part of the CDDIS interface tools.

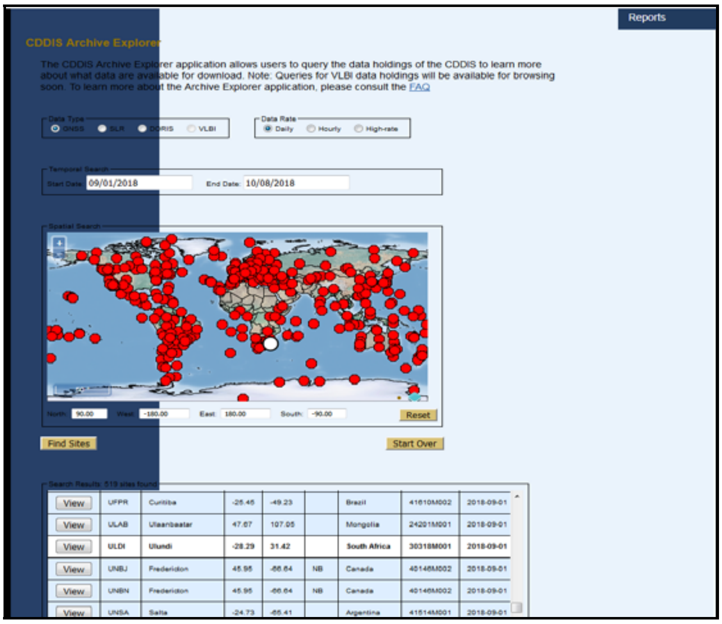

Figure 6.6: CDDIS Archive Explorer

(CDDIS 2016)

With reference to Figure 6.6, the CDDIS Explorer application allows users to query the data holdings of the CDDIS by selecting search parameters, such as data type, data rate and temporal target, on the interactive GUI. A map indicating the site/station and technique, as well as a table with more details regarding the site/station and technique are displayed.

6.4 Key recommendations

In order to achieve the execution of the design of the new GRDMS, various processes, hardware and software need to be implemented. Although the implementation and testing of the GRDMS falls outside the scope of this study, to ground the study it was necessary to provide a brief discussion on the future implementation of systems, processes, hardware and software as part of key recommendations for the design of the GRDMS. The following sections provide a brief discussion of the progress with the system design up until this point.

6.4.1 Hardware

HartRAO has recently procured a Dell Westmere processing cluster, containing 14.8 Teraflops and 3456 GB random access memory. Implementation and commissioning of the cluster is expected to occur during 2018/9 (Coetzer, Botha & Jacobs 2018). An ftp server needs to be implemented for archiving data in the new system. This server would allow users access to the raw and custom processed data and products and will also be responsible for streaming the raw data to the international partners and service providers.

6.4.2 Software

Within the *Data Ingress System/Unit,* open Virtual Private Network (VPN) connections need to be established to add security and protection of data stored on the proposed GRDMS. The VPN connections will connect all of the geodesy stations and instruments to the *Data Ingress System/Unit* virtual machine, which will handle and monitor the data collection process and also conduct quality checks. According to the designers, the following necessary software needs to be installed after the design of the GRDMS has been completed:

- *Joomla*® content management system software needs to be installed to manage all forms of content in the new GRDMS. Joomla® is free open source content management system software;
- *Astronomical image processing system (AIPS)* and the *Common Astronomy Software Aapplications (CASA)* software, data reduction software, will be installed

on the GRDMS *Data Pre-processing Unit.* These are necessary programs that will support the reduction and analysis of astronomy data taken by radio telescopes (National Radio Astronomy Observatory 2016);

- *MATLAB*®, a programming language for algorithm development, data analysis, visualisation and numeric computation (Tshwane University of Technology 2018);
- *Vienna VLBI Software (VieVS)*, VLBI analysis software for the analysis and numeric computation of geodesy VLBI data (Hellerschmied & Mayer 2016);
- *GAMIT/GLOBK*, a collection of programs used for the analysis of GPS and meteorology data, as well as *SeisComP3*®, seismological software for data acquisition, processing, distribution and interactive analysis (Deutsches GeoForschungs-Zentrum 2018);
- *Leica*® software packages for the acquisition, processing and analysis of site tie and GNSS data (Leica Geosystems 2018) will be installed on the GRDMS *Data Pre-processing Unit* on a Linux permanent machine and a Windows virtual machine. The software was recently tested and found to be working satisfactorily. Still to be implemented on the GRDMS *Central Repository/Archive* are scripts to collect metadata, assign DOIs and provide daily status reports and logs, as well as the installation of GNSS data search tools to redirect data from the Dataworks port (Coetzer, Botha & Jacobs 2018); and
- *UNACVO's Dataworks for GNSS* system software will be installed on Linux virtual machines to serve as middleware. The middleware will run scripts that will perform quality checks on the incoming data and convert the data to the required versions (e.g. convert RINEX Version 3 data to RINEX Version 2 data). These are necessary steps for the pre-processing, archiving and storage of RINEX format data.

In conclusion: the new GRDMS needs to be fully interoperable and compatible with similar data management systems (i.e. the CDDIS, IVS, ILRS, etc.), as well as fully operational as soon as the design of the system has been completed and software has been installed and tested. Full implementation is to follow once it is performing to the designers' satisfaction.

6.5 Summary and conclusions

This study will make a significant contribution in filling the gap that exists in availability of contextual information on global geodetic data management, systems, system design and geodetic data management service. The study will also contribute towards filling the library and information science knowledge gap. Lessons learnt and knowledge and expertise generated by this unique study will form part of a theoretical framework for designing a data management system for HartRAO. This will ultimately lead to the design of a GRDMS that will comply with the technological expansion of HartRAO, while meeting the needs of users at HartRAO and the global scientific geodetic data user community, as well as requirements set by international geodetic data service providers and funding agencies. This is a fundamental step in enabling the open science paradigm for which the space geodesy research community is striving.

The findings of this study, both attained from the results of the interviews conducted with participants and the literature reviewed, support and align with the achievement of the research questions and objectives related to components, characteristics, criteria, data structuring and organisation, and the proposed model for the development of a new GRDMS. The GRDMS, which will manage HartRAO's geodesy data repository with a Web service enabled application program interface, is intended to provide simple, consistent Web services to geodetic-focused international geodesy data service providers, users at HartRAO, members of the international geodesy research community and members of the public, in order to facilitate the discovery of, sharing of and access to geodesy data and products.

The proposed GRDMS model assumes that geodesy data is collected at instrument sites/stations, such as a VLBI, GNSS, SLR stations etc. The GRDMS will be designed to provide access to instrumental data files and log files and to query for metadata about geodesy stations and instruments. Ancillary site-based data, information, such as meteorological observations, seismic events, etc., can also be provided through the new GRDMS. The GRDMS software will include a web graphic user interface that will leverage the GRDMS web services for web-based search and access.

The new GRDMS will be designed to be fully interoperable and compatible with international data service providers, such as the CDDIS, IVS, IERS, IGS, etc. and will have several benefits. It is envisaged by the designers of the new GRDMS that it will contribute towards establishing HartRAO as an important geodetic data service provide – the only one of its kind in Africa – and a vital link in a network of international geodesy data service providers. The GRDMS will contribute to promoting and increasing the visibility of HartRAO's data holdings and services and will also contribute to establishing, promoting and increasing international geodesy collaborations between HartRAO's funding agency (National Research Foundation) and international funding bodies and governments. It will also establish and support collaboration between HartRAO and international geodetic research facilities.

Concerning the international geodetic community, the new GRDMS will fill the gap that exists in the current network of regional and global data service providers (with emphasis on southern hemisphere geodesy dataset provision). African geodesy data managed by the GRDMS could be used for studying earth systems, global climate change and natural disasters such as flooding, earthquakes and tsunami for the preservation of life on earth. The GRDMS could contribute geodesy data towards the study of the ecology, biodiversity, infectious diseases, etc.

For the South African government, the benefit of having the new HartRAO GRDMS is that it will contribute to the upliftment of the socio-economic and socio-political environment of South Africa. Geodetic data provided by the GRDMS can be used for housing, urban development, transportation, communication, natural resource management and sustainability for the citizens of the Republic of South Africa.
Geodesy data generated globally, including that generated by HartRAO local instrumentation and by geodesy instrumentation in African partner countries, are being used for the navigation of unmanned space probes and manned space crafts. Geodesy data is used to track and facilitate autonomous landing of space probes and space crafts on the Moon, asteroids and planets (Jacobs 2018).

The GRDMS will be to the benefit of space exploration and asteroid mining (exploration and excavation of carbon, silicon and metals contained in asteroids). Global companies are at present locked in a race to claim the trillions of dollars' worth of precious metals thought to exist in asteroids. Asteroid farming (i.e. use of soil on asteroids to grow plants and crops in space) is becoming increasingly important to scientists, governments and private industry, which can benefit from the HartRAO GRDMS. Geodesy data is of great importance to people such as Elon Musk, Chief Executive Officer of SpaceX (a private American aerospace manufacturer and space transportation service company), as it is used for orbit determination, geostationary orbit transfers and interplanetary spaceflight (Amos 2017). Elon Musk is currently constructing a constellation of 4 425 satellites capable of beaming the Internet to the entire globe, including remote regions that currently do not have Internet access (Kang & Davenport 2015). In partnership with other geodesy data service providers, HartRAO's new GRDMS can provide the much needed data for Musk's enterprise(s).

More importantly, data generated by HartRAO's geodetic instrumentation and managed by the new GRDMS, can and will be used in conjunction with other geodesy service provider's data, to overcome global information and data poverty and to improve the quality of life of all citizens on planet earth.

6.6 Proposed future research

As a recognised regional radio astronomy and space geodesy research data service provider, HartRAO is a vital link in a global network of high-quality scientific data service providers (HartRAO 2001; Gaylard & Nicolson 2007:51–52; De Witt et al. 2013:45; Mashaba et al. 2016). The various co-located radio astronomy and space geodesy instrumentation hosted by HartRAO makes this research facility the only fundamental Geodesy station in Africa (Nickola 2012:5) and one of seven fiducial sites worldwide (Combrinck 2014). As mentioned in Chapter 1, as HartRAO is currently the only space geodesy research facility in South Africa and in Africa, the design of a geodetic research data management system is a unique endeavour where South Africa and Africa are concerned.

Based on the evidence of this research, it is recommended that the proposed model leads to further research involving the construction, implementation, testing, debugging, fine-tuning, revising, integration and enhancement of the GRDMS.

Possible future research also includes:

- structuring and organising of data types within the GRDMS by using standards, schemes, conventions and tools (such as FNCS and ORCIDS);
- use of DOIs to enhance the visibility and retrievability of HartRAO's space geodesy data; and
- study of bibliometric tools, such as Scopus, Incite and Web of Science, in order to perform advanced bibliometric analyses of data usage, impact factor and return on scientific and financial investment.

Since HartRAO also focuses on training students, students can gain experience in data librarianship, data stewardship, digital data curatorship, data science and data science management by involving them in projects such as:

- open data /open science in the African context;
- transformations in research and librarianship as a result of the data revolution;
- big data in the scientific and commercial sectors;
- data management which can be international, interdisciplinary and trans-disciplinary in scope; and
- data management within the context of disaster risk research.

In South Africa, the study area of space geodesy data science and management has limited expertise. Hence, these projects could attract young scientists, information specialist, data librarians and curators to specialise in the field of research data management and research data management system design.

6.7 Conclusion

This chapter provided a brief reiteration of the study objectives to be achieved. Recommendations to consider for the design of the new HartRAO GRDMS model were formulated, based on the conclusions drawn from the literature review and interviews with participants.

This chapter concluded with suggestions for future research in South Africa and Africa. The final offering of this chapter was the proposal of a GRDMS design model that can be used as the foundation for the further development, implementation and testing of a new GRDMS for HartRAO.

REFERENCES

Acroyd, S. 2010. Critical realism, organization theory, methodology and the emerging science of reconfiguration. *Elements of a philosophy of management and organization.* Berlin: Springer.

Altintas, I., Berkley, C., Jaeger, E. & Mock, S. 2004. *Kepler: an extensive system for design and execution of scientific workflows.* Proceedings of the 16th International Conference on Scientific and Statistical Database Management, Santorini, Greece, 2004.

Amos, J. 2017. *Elon Musk: rockets will fly people from city to city in minutes*. https://www.bbc.com/news/science-environment-41441877 [Accessed 22 September 2018].

Amuno, A. 2018. *The five types of systems software.* https://turbofuture.com/computers/The-Five-Types-of-System-Software [Accessed 24 September 2018].

Antell, K., Bales Foote, J., Turner, J. & Shults, B. 2014. Dealing with data: science librarians' participation in data management at association of research libraries institutions. *College and Research Libraries*, 75(4):557–573.

Antin, K. 2016. *File naming conventions: why you want them and how to create them.* https://www.huridocs.org/2016/07/file-naming-conventions-why-you-want-them-and-how-to-create-them/ [Accessed 12 November 2018].

Appleby, GM., Bianco, G., Noll, C., Pavlis, E. & Pearlman, MR. 2016. *Current trends and challenges in satellite laser ranging*. Proceedings of the International VLBI Service for Geodesy and Astrometry 2016 General Meeting. Greenbelt, MD: NASA.

Austin, SN. 2016. *Types of network software*. http://www.techwalla.com/articles/types-of-network-software [Accessed 4 August 2016].

Babbie, E. & Mouton, J. 2011. *The practice of social research*. Cape Town: Oxford University Press.

Beal, V. 2018. *Interface*. https://www.webopedia.com/TERM/I/interface.html [Accessed 7 March 2018].

Behrend, D. 2013. Data handling within the international VLBI service. *Data Science Journal,* 12:WDS81–WDS84. http://doi.org/10.2481/dsj.WDS-011 [Accessed 4 March 2016].

Bezuidenhout, R-M. 2014. Theory in research, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M Bezuidenhout. Cape Town: Juta, pp 36–59.

Bezuidenhout, R-M & Cronje, F. 2014. Qualitative data analysis, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M Bezuidenhout. Cape Town: Juta, pp. 228–250.

Bezuidenhout, R-M. & Davis, C. 2014. From topic to research question, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M Bezuidenhout. Cape Town: Juta, pp. 60–71.

Black, P. 2014. Data structures, in *Dictionary of algorithms and data structures,* edited by Vreda Pieterse and Paul E. Black. http://www.nist.gov/dads/HTML/datastructur.html [Accessed 9 February 2017].

Boler, F., Meertens, C., Miller, M.M., Wier, S., Rost, M., Matykiewicz, J. 2015. *Dataworks for GNSS: software for supporting data sharing and federation of geodetic networks*. http://www.unavco.org/software/data-management/dataworks/lib/docs/AGU-2015-Boler.pdf [Accessed September 24 2018].

Boston University Libraries. 2015a. *Research data management: what is research data?* http://www.bu.edu/datamanagement/background/whatisdata [Accessed 30 April 2015].

Boston University Libraries. 2015b. *Research data management: organizing your data*. http://www.bu.edu/datamanagement/outline/elements/organize/data [Accessed 7 July 2016].

Botha, R. 2018. *GeoStation Manual v0.1. Hartebeesthoek Radio Astronomy Observatory Space Geodesy Programme.* http://geodesy.hartrao.ac.za/site/en/ [Accessed 7 August 2018].

British Library. 2018. *Collection metadata standards: resource description and access (RDA).* http://www.bl.uk/bibliographic/catstandards.html [Accessed 23 September 2018].

Bryant, R. 2013. *Connecting research datasets and researchers: ORCID use cases and integrations*. https://orcid.org/blog/2013/06/17/connecting-research-datasets-and-researchers [Accessed 17 May 2016].

Bryman, A. 2012. *Social research methods*. 5th ed. Oxford: Oxford University Press.

Butler, D. 2012. Scientists: your number is up. *Nature*, 485(4700), 30 May 2012:564. https://www.nature.com/news/scientists-your-number-is-up-1.10740 [Accessed 11 August 2017].

CA Technologies. 2018. *Understanding access paths*. https://docops.ca.com/ca-2e/87/en/building/building-access-paths/understanding-access-paths [Accessed 7 February 2018].

CDDIS (NASA's Archive of Space Geodesy Data). 2015. *VLBI Data holding*. https://cddis.nasa.gov/Data_and_Derived_Products/VLBI/VLBI_data_holdings.html [Accessed 8 February 2017].

CDDIS (NASA's Archive of Space Geodesy Data). 2016. *CDDIS Archive explorer*. https://cddis.nasa.gov/Data_and_Derived_Products/CddisArchiveExplorer.html [Accessed 24 September 2018].

CDDIS (NASA's Archive of Space Geodesy Data). 2017. *Standards for GNSS data submitted to CDDIS*. https://cddis.nasa.gov/docs/2017/GNSSDataStandards_v2.pdf [Accessed 17 November 2017].

Chisholm, M. 2016. *Seven phases of a data life cycle*. https://www.bloomberg.com/blog/7-phases-of-a-data-life-cycle [Accessed 7 July 2016].

Clark, R. 2012. *What is the relationship between ISNI and ORCID?* https://orcid.org/content/what-relationship-between-isni-and-orcid [Accessed 21 September 2017].

Coetzer, GL., Botha, RC., Combrinck, WL. & Fourie, CJS. 2015. A new geodetic research data management system at the Hartebeesthoek Radio Astronomy Observatory. Open Science at the Frontiers of Librarianship. Proceedings of a conference held at Astronomical Observatory of Capodimonte, Naples, Italy, 17–20 June 2014, edited by András Holl, Soizick Lesteven, Dianne Dietrich and Antonella Gasperini. *ASP Conference Series*, 492. San Francisco: Astronomical Society of the Pacific, p. 22.

Coetzer, GL. & Botha, RC. 2016. *The new geodetic research data management system at HartRAO*. Poster presented at the International VLBI Service for Geodesy and Astrometry 2016 Meeting: “New Horizons with VGOS”, Johannesburg, South Africa, 13–19 March 2016.

Coetzer, GL., Botha, RC. & Jacobs, L. 2018. *Progress with the new research data management system at HartRAO.* Library & Information Services (LISA) VIII Conference, “Astronomy Librarianship in the Era of Big Data and Open Science”, European Doctoral College, Strasbourg, France, 6–9 June 2017.

Combrinck, WL. 2014. *Hartebeesthoek Radio Astronomy Observatory: Space Geodesy status and progress*. Proceedings of the Second AfricaGEO Conference, CONSAS Conference. Cape Town, South Africa.

Combrinck, WL. & Combrink, AZA. 2004. *Proposed new international Space Geodesy facility for southern Africa*. 35th COSPAR Scientific Assembly, Paris, France, 18–25 July 2004.

Core Informatics. 2016. *The core SDMS (scientific data management system): capture, automate and integrate*. http://www.coreinformatics.com/products/core-sdms/ [Accessed 9 August 2016].

Cormen, TH., Leiserson, CE., Rivest, RL. & Stein, C. 2009. *Introduction to allgorithms*. 3rd ed. MIT Press.

Creswell, JW. 1994. *Research design: qualitative and quantitative approaches.* Thousand Oaks, CA: Sage.

Creswell, JW. 2012. *Educational research: planning, conducting and evaluating quantitative and qualitative research.* 4th ed. Upper Saddle River: Pearson Prentice Hall.

Datanet. [n.d.]. *Data capturing*. http://www.datanet.sa.co.za/index.htm [Accessed 5 March 2018].

Datt, S. 2017. *Importance of research approach in research.* https://www.projectguru.in/publications/selecting-research-approach-business-studies [Accessed 20 September 2017].

Davis, C. 2014a. Constructing arguments in research, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 121–130.

Davis, C. 2014b. What is research? in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 1–17.

Davis, W. 2016. *What is data structure? Structuring data and organization like Martha Stewart.* Trifacta. https://www.trifacta.com/blog/what-is-data-structure/ [Accessed 6 December 2017].

Davis, S. & Coetzer, GL. 2010. Astronomical libraries in South Africa. Library and Information Services in Astronomy VI: 21st century astronomy librarianship, from new ideas to action. Conference held at IUCAA & NCRA, Pune, India, 14–17 February 2010. *ASP Conference Series*, 433:69–72.

DBMS Intervals. 2018. *Three level database architecture*. http://www.dbmsinternals.com/database-fundamentals/basic-architecture/three-level-database-architecture/ [Accessed 25 September 2018].

De Vos, A., Strydom, H., Fouché, CB. & Delport, CSL. 2011. *Research at grass roots.* 4th ed. Pretoria: Van Schaik.

De Waard, A., Cousijn, H. & Aalbersberg, IJ. 2015. *Ten aspects of highly effective research data: good research data management makes data reusable*. https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data [Accessed 7 March 2018].

De Witt, A., Gaylard, MJ., Quick, JFH. & Combrinck, WL. 2013. *An overview of geodetic and astrometric VLBI at the Hartebeesthoek Radio Astronomy Observatory*. Proceedings of the 21st Meeting of the European VLBI Group for Geodesy and Astrometry. Finnish Geodetic Institute.

De Witt, A., Mayer, D., MacLeod, G., Combrinck, L., Petrov, L. & Nickola, M. 2016. *Optimizing the African VLBI Network for Astrometry and Geodesy*. International VLBI Service for Geodesy and Astrometry. 2016 Proceedings of General Meeting. Greenbelt, MD: NASA.

Deutsches GeoForschungs-Zentrum. 2018. *SeisComp3*. https://www.seiscomp3.org/ [Accessed 24 September 2018].

Driscoll, DL. & Brizee, A. 2012. *Ethical considerations in primary research*. https://owl.english.purdue.edu/owl/resource/559/02 [Accessed 5 May 2015].

Du Plooy-Cilliers, F. 2014. Research paradigms and traditions, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 18–35.

Du Plooy-Cilliers, F. & Cronje, J. 2014. Quantitative data collection, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp 147–172.

Dubois-Felsmann, G. 2016. *Large synoptic survey telescope (LSST): data management subsystem requirements*. https://docushare.lsstcorp.org/docushare/dsweb/Get/LSE-61 [Accessed 13 February 2017].

English Oxford living dictionary. 2018. sv ‘data’. https://en.oxforddictionaries.com/definition/us/data [Accessed 7 February 2018].

ENotes. 2018*. What is the need of data structures?* https://www.enotes.com/homework-help/what-need-data-structures-125063 [Accessed 25 September 2018].

Enslin, C. 2014. Limitations, delimitations and recommendations, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 274–281.

Eubanks, TM. 1993. Variations in the orientation of the earth, in *Contributions of space geodesy to geodynamics: earth dynamics. Geodynamics 24,* edited by DE. Smith and DL. Turcotte. Washington, DC: American Geophysical Union.

Federal Agency for Cartography and Geodesy. 2013. *File naming conventions and structure of the ftp archive*. https://www.iers.org/IERS/EN/Organization/About/History/WGCombination/filenamingconventionstructureftp.html [Accessed 7 September 2016].

Flick, U. 2011. *Introducing research methodology: a beginner's guide to doing a research project.* London: Sage.

Fox, P. & Harris, R. 2013. ICSU and the challenges of data and information management for international science. *Data Science Journal,* 12:WDS1–WDS12.

The Free Dictionary. Thesaurus. 2018. sv 'data management system'. https://www.thefreedictionary.com/Data+management+system [Accessed 9 November 2018.]

Gable, GG. 1994. Integrating case study and survey research methods: an example in information systems. *European Journal in Information Systems,* 3(4):112–116.

Gaylard, MJ. & Nicolson, GD. 2007. Forty years of radio astronomy. *African Skies = Cieux Africains,* 11:49.

Gaylard, MJ., Quick, JFH., Gotz, A. & Jonas, J. 1997. The new computer control system (NCCS) for the Hartebeesthoek 26m radio telescope. *Monthly Notices of the Royal Astronomical Society,* 56(3–4):32.

Georgia Technical Library. 2018. *Organize your data*. http://d7.library.gatech.edu/research-data/organization [Accessed 11 March 2018].

Goddard, W. & Melville, S. 2004. *Research methodology: an introduction*. 2nd ed. Oxford: Blackwell Publishing.

Gurav, S. 2017. *Research process*. https://www.slideshare.net/saileegurav/research-process-30441999 [Accessed 22 February 2017].

Harley, D., Acord, SK., Earl-Novell, S., Lawrence, S. & King, CJ. 2010. Assessing the future landscape of scholarly communication: an exploration of faculty values and needs in seven disciplines. *Open Access Publications from the University of California*, 138. http://www.escholarship.org/uc/item/15x7385g [Accessed 3 March 2015].

Hartebeesthoek Radio Astronomy Observatory. 2001a. *26m Telescope computer control system (NCCS).* http://www.hartrao.ac.za/nccs [Accessed 7 January 2017].

Hartebeesthoek Radio Astronomy Observatory. 2001b. *NCCS software*. http://www.hartrao.ac.za/nccsdoc/developer/developer-2.html [Accessed 7 January 2017].

Hartebeesthoek Radio Astronomy Observatory. 2001c. *The Hartebeesthoek Radio Astronomy Observatory*. http://www.hartrao.ac.za/summary/sumeng.html [Accessed 11 February 2016].

Hartebeesthoek Radio Astronomy Observatory. 2012. *Space geodesy programme: GNSS*. http://geodesy.hartrao.ac.za/site/en/data-and-products/gnss.html [Accessed 14 August 2016].

Hayward, S. 2017. *Experts explain: the rise of laboratory data lakes*. Advantage Business Media. https://www.laboratoryequipment.com/article/2017/05/experts-explain-rise-laboratory-data-lakes [Accessed 15 January 2018]

Hellerschmied, A. & Mayer, D. 2016. *Introduction to VieVs*. https://publik.tuwien.ac.at/files/PubDat_248832.pdf [Accessed 24 September 2018].

Higgins, S. 2007. *What are metadata standards?* Digital Curation Conference (DCC). http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards [Accessed 20 August 2018].

Hofstee, E. 2006. *The literature review: constructing a good dissertation: a practical guide to finishing a Master's, MBA and PhD on Schedule*. http://www.engl.polyu.edu.hk/BACapstoneProject/documents/Constructing_a_good_dissertation.pdf [Accessed 7 July 2017].

Hogrefe, K. & Stocks, K. 2011. The importance of metadata standards, in *The MMI guides: navigating the world of marine metadata,* edited by KI. Stocks, C. Neiswender, AW Isenor, J. Graybeal, N. Galbraith, ET. Montgomery, P. Alexander, S. Watson, L. Bermudez, A. Gale and K. Hogrefe. http://marinemetadata.org/guides/mdatastandards/stdimportance [Accessed 8 February 2017].

Holl, A. 2010. *Small data archives and libraries*. Library and Information Services in Astronomy VI: 21st Century Astronomy Librarianship, From New Ideas to Action, Pune, Maharashtra, India, 14–17 February 2010. *ASP Conference Series*, 433:201–206.

Howard, G-M. 2014. The literature review, in *Research matters,* edited by. F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 100–106.

IBM Knowledge Centre. 2010. *What is a database management system?* *https://www.ibm.com/support/knowledgecenter/en/zosbasics/com.ibm.zos.zmiddbmg/zmiddle_46.htm* [Accessed 9 November 2018].

Illinois University Library. 2018. *Introduction to data management for undergraduate students: file naming conventions*. http://guides.library.illinois.edu/introdata/filenames [Accessed 12 November 2018].

Indiana University. 2018. *Understanding measures of supercomputer performance and storage system capacity*. https://kb.iu.edu/d/apeq [Accessed 5 May 2018].

International Standard Name Identifier (ISNI). 2017. *ISNI and ORCID*. http://www.isni.org/isni_and_orcid [Accessed 24 August 2017].

Inspedium. 2018. *How many Mb's are there in a Gb, Tb, Kb or Byte? (Conversion Tool).* https://www.inspedium.com/website-hosting/mb-gb-converter/ [Accessed, 7 November 2018].

Jaakkola, H. & Thalheim, B. 2011. *Architecture-driven modelling methodologies*. Proceedings of the 2011 Conference on Information Modelling and Knowledge Bases XXII. IOS Press.

Jacobs, C. 2018. *Stellar GPS: navigating the solar system*. http://avntraining.hartrao.ac.za/index.php/news-and-events/news/82-invited-talk-friday-23-march-mr-christopher-jacobs [Accessed 25 September 2018].

Jacobs, L. 2014. The research report, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis & R-M and Bezuidenhout. Cape Town: Juta, pp. 294–308.

Jansen, J.D. 2007. The research question, in *First steps in research,* edited by Kobus Maree. Pretoria: Van Schaik, pp. 2–13.

Kaif, H. 2013. *Understanding what is inside your computer and how it works*. http://sites.jmu.edu/103oconnor-16/introduction-to-basic-computer-components-and-functions/ [Accessed 11 February 2018].

Kang, C. & Davenport, C. 2015. *SpaceX founder files with government to provide Internet service from space*. https://www.washingtonpost.com/business/economy/spacex-founder-files-with-government-to-provide-internet-service-from-space/2015/06/09/db8d8d02-0eb7-11e5-a0dc-2b6f404ff5cf_story.html?noredirect=on&utm_term=.b603d0c544b7 [Accessed 25 September 2018].

Kaur, M., Singh, S. & Kaur, R. 2016. Directory structure and file allocation methods. *International Journal of Computer Science and Information Technologies,* 7(2):577–582.

Keyton, J. 2011. *Communication research: asking questions, finding answers*. 3rd ed. New York: McGraw-Hill.

Konomi, E. & Marra, M. 2015. CRIS-INAF as a result of a fruitful collaboration among ITs, librarians and researchers. Library and Information Services in Astronomy VII: Open Science at the Frontiers of Librarianship. ASP Conference Series, 492, edited by A. Holl, S. Lesteven, D. Dietrich and A. Gasperini. San Francisco, California: Astronomical Society of the Pacific.

Open Science at the Frontiers of Librarianship. Proceedings of a conference held at the Astronomical Observatory of Capodimonte, Naples, Italy 17–20 June 2014, edited by András Holl, Soizick Lesteven, Dianne Dietrich, and Antonella Gasperini. San Francisco: Astronomical Society of the Pacific, p. 38. *ASP Conference Series*, 492.

Koonin, M. 2014. Validity and reliability, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 252–260.

Koopman, MM. & De Jager, K. 2016. Archiving South African digital research data: how ready are we? *South African Journal of Science*, 112(7/8):1–7.

Kraus, JD. 1989. *Radio astronomy.* 2nd ed. Ohio: Cygnus-Quasar Books.

Kumar, R. 1999. *Research methodology: a step-by-step guide for beginners*. Thousand Oaks, CA: Sage.

Lacy, A. & Luff, D. 2007. *Qualitative data analysis*. NIHR Research Design Service for Yorkshire & the Humber. https://www.rds-yh.nihr.ac.uk/wp-content/uploads/2013/05/9_Qualitative_Data_Analysis_Revision_2009.pdf [Accessed 19 September 2017].

Leica Geosystems. 2018. *Leica Geosystems Total Stations*. https://africa.leica-geosystems.com/en/index.htm [Accessed 24 September 2018].

Levy, Y. & Ellis, TJ. 2006. A systems approach to conducting an effective literature review in support of information systems research. *Informing Science Journal*, 9:54.

Lim, K-T, Bosch, J., Dubois-Felsmann, G., Jenness, T., Kantor, J., O'Mullane, W., Petravick, D., Comoretto, G and the DM Leadership Team. 2017. Data management system design, in *Large Synoptic Survey Telescope (LSST).* https://docushare.lsst.org/docushare/dsweb/Get/LDM-148 [Accessed 9 May 2017].

LIMSWiki. 2016. *Scientific data management system*. http://www.limswiki.org/index.php/Scientific_data_management_system [Accessed 4 August 2016].

Louw, M. 2014. Ethics in research, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 262–273.

MacMillan, D., Pavlis, E., Kuzmicz-Cieslak, M. & Koenig, D. 2016. *Generation of global geodetic networks for GGOS.* International VLBI Service for Geodesy and Astrometry, Proceedings of 2016 General Meeting. Greenbelt, MD: NASA.

Manzini, ST. 2012. The national system of innovation concept: an ontological review and critique. *South African Journal of Science*, 108(9/10). http://archive.sajs.co.za/index.php/SAJS/article/view/1038 [Accessed 7 February 2015].

Martin, E. & Ballard, G. 2010. *Data management best practices and standards for biodiversity data applicable to bird monitoring data*. U.S. North American Bird Conservation Initiative Monitoring Subcommittee. http://www.nabci-us.org/ [Accessed 10 June 2015].

Mashaba, Z. 2014. *Implementation and design of a web-based GNSS data management system at Hartebeesthoek Radio Astronomy Observatory (HartRAO).* 10th Annual Inkaba yeAfrica and !Khure Africa Conference, Matjiesfontein, Western Cape, 28 September–3 October 2014. http://events.saip.org.za/contribution [Accessed 7 February 2015].

Mashaba, Z., Combrinck, WL., Botai, JO., Munghemezulu, C. & Botha, RC. 2016. Design of a web-based GNSS data management system at HartRAO: preliminary results. *South African Journal of Geology,* 119(1). http://sajg.geoscienceworld.org/content/119/1/117 [Accessed 15 March 2016].

McLeod, R. 1983. *Management information systems. 2nd ed.* Chicago: Science Research Associates.

Mercer, C. 2018. *What is Li-Fi? Everything you need to know*. https://www.techworld.com/data/what-is-li-fi-everything-you-need-know-3632764/ [Accessed 20 September 2018].

Merriam-Webster. 2018. sv ‘searching’. Merriam-Webster Dictionary. https://www.merriam-webster.com/dictionary/search [Accessed 19 February 2018].

Merriam-Webster. 2018. sv ‘criteria’. Merriam-Webster Dictionary. https://www.merriam-webster.com/dictionary/criterion [Accessed 19 February 2018].

Mey, P. 2017. Putting South Africa on the VGOS map. *IVS Newsletter*, 47:6.

Miles, MB., Huberman, AM. & Saldaña, J. 2014. *Qualitative data analysis*. 3rd ed. Thousand Oaks, CA: Sage.

National Aeronautics and Space Administration. 2012. *Deep space navigation*. https://scienceandtechnology.jpl.nasa.gov/research/research-topics-list/communications-computing-software/deep-space-navigation [Accessed 16 May 2018].

National Information Standards Organisation (NISO). 2006. Building a metadata schema: where to start? *ISO 23081-1:2006 Information and documentation: records management processes. Metadata for records. Part 1, Principles.* http://www.niso.org/apps/group_public/download.php/5271/N800R1_Where_to_start_advice_on_creating_a_metadata_schema.pdf [Accessed 7 May 2016].

National Radio Astronomy Observatory. 2016. *What is AIPS?* http://www.aips.nrao.edu/aips_faq.html [Accessed 24 September 2018].

National Research Foundation. 2017*. Annual Performance Plan, 2017/18–2019/20*. Department of Science and Technology.

Nelson, B. 2009. Empty archives. *Nature*, 461:160-163.

Neuman, WL. 2011. *Social research methods: qualitative and quantitative approaches.* 3rd ed. Boston: Allyn & Bacon.

Nieuwenhuis, J. 2007. Introducing qualitative research, in *First steps in research,* edited by Kobus Maree. Pretoria: Van Schaik, pp. 47–66.

Nickola, M. 2012. *Astronomical seeing conditions as determined by turbulence modelling and optical measurement*. MSc dissertation, Faculty of Geography, Geoinformatics and Meteorology, University of Pretoria.

Noll, C. 1997. *The global positioning system for the geosciences.* Summary and proceedings of a workshop on improving the GPS reference station infrastructure for earth, oceanic and atmospheric science applications. National Academic Press.

Noll, C. 1999. *Data centres and data access*. GPS 99, Tsukuba, Japan, 21 October 1999. https://cddis.nasa.gov/docs/1999/igsdc_tutorial.pdf [Accessed 12 November 2018].

Noll, C. 2000. *International VLBI Service for Geodesy and Proceedings of Astrometry*. 2000 CDDS General Meeting, edited by NR. Van den Berg and KD. Baver. http://ivs.nict.go.jp/mirror/publications/gm2000/toc.html [Accessed 15 April 2017].

Noll, C. 2003. CDDIS Data Centre summary for the 2003 IVS Annual Report, in *IVS 2003 Annual Report*. http://cddis.gsfc.nasa.gov/docs/2003/cddis_ivs2003.pdf [Accessed 14 August 2016].

Noll, C. 2005. *IDS Data flow coordination*. https://cddis.nasa.gov/docs/2005/IDS_data_centers_2003_2005.pdf [Accessed 7 September 2017].

Noll, C. 2010. The CDDIS: a resource to support scientific analysis using space geodesy. *Advances in Space Research,* 45:1421–1440.

Noll, C. 2013. *An update on the CDDIS*. International GNSS Service (IGS) Workshop, Olsztyn, Poland, 23–27 July 2013. https://ntrs.nasa.gov/search.jsp?R=20120009701 [Accessed 15 September 2016].

Noll, C. 2015a. *Metadata standards for global geodesy*. Technical Meeting on Metadata Standards for Global Geodesy, Working Group on Data and Information, GGOS Bureau of Networks and Observations.

Noll, C. 2015b. CDDIS *Global Data Center Technical Report, 2014*. https://cddis.nasa.gov/docs/2014/IGS_cddistr_2014.pdf [Accessed 11 November 2018].

Noll, C. 2016. *CDDIS Global Data Centre technical report, 2015*. https://cddis.nasa.gov/docs/2015/CDDIS_IGStr_Noll_2015.pdf [Accessed 11 November 2018].

Noll, C. 2017. *CDDIS Data Center Summary for the IVS 2015–2016 Biennial Report*. https://cddis.nasa.gov/docs/2016/cddis_IVSreport_2015_2016.pdf [Accessed 15 April 2017].

Noll, C. & Michael, P. 2016. *Developments at CDDIS to support real-time and RINEX. Vol. 3.* IGS Workshop: “GNSS Futures”, Sydney, Australia, 8–12 February 2016.

Noll, C., Michael, P., Dube, M. & Pollack, N. 2012. *An update on the CDDIS*. International GNSS Service (IGS) Workshop, Olsztyn, Poland, 23–27 July 2012.

Noll, C., Michael, P. & Pollack, N. 2015. *Recent developments at the CDDIS in support of GGOS.* American Geophysical Union, Fall Meeting 2015. https://cddis.nasa.gov/docs/2015/aguposter_1512_v2b.pdf [Accessed 12 September 2016].

Observational Health Data Sciences and Informatics (OHDSI). 2018. *Data standardization*. 2018 OHDSI Symposium. https://www.ohdsi.org/data-standardization/ [Accessed 3 May 2018].

Onwuegbuzie, AJ., Frels, RK. & Hwang, E. 2016. Mapping Saldana’s coding methods onto the literature review process. *Journal of Educational Issues*, 2(1).

Open Research and Contributor ID (ORCID). 2017. *Josiah Carberry ORCID ID*. http://orcid.org/0000-0002-1825-0097 [Accessed 21 September 2017].

Orsborn, T. 2013. Five factors to consider when selecting a relational database. *SAP Insider.* http://sapinsider.wispubs.com [Accessed 3 August 2015].

Palmius, J. 2007. *Criteria for measuring and comparing information systems*. Proceedings of the 30th Information Systems Research Seminar, Scandinavia, IRIS 2007. http://www.palmius.com/joel/text/IRIS-30-final.pdf [Accessed 17 June 2016].

Pascoe, G. 2014. Sampling, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 131–145.

Pearce, JM. 2015. Return on investment for open source hardware development. *Science and Public Policy,* 43(2), April 2016. https://academic.oup.com/spp/article/43/2/192/2414129 [Accessed 13 June 2017].

Penn State University Libraries. 2015. What is data management? *Publishing and Curation Services.* https://www.libraries.psu.edu/psul/pubcur/what_is_dm.html#data-management [Accessed 28 March 2015].

Perret, E. et al. 2015. *Working together at CDS: the symbiosis between astronomers, documentalists and IT speci*alists. Open Science at the Frontiers of Librarianship. Proceedings of a conference held at the Astronomical Observatory of Capodimonte, Naples, Italy, 17–20 June 2014, edited by András Holl, Soizick Lesteven, Dianne Dietrich and Antonella Gasperini. *ASP Conference Series*, 492. San Francisco: Astronomical Society of the Pacific, p. 13.

Pring-Mill, D. 2018. *Why hasn't AI mastered language translations?* https://singularityhub.com/2018/03/04/why-hasnt-ai-mastered-language-translation/#sm.001xhz17fi5scsb10f01trgn05768 [Accessed 22 September 2018].

Purdon, A. 2018. *How to choose the correct fiber line speed and data package*. https://www.webafrica.co.za/blog/fibre/choose-correct-line-speed-data-package/ [Accessed 4 July 2018].

Purdue University Libraries. 2017. *Data management for undergraduate researchers: file naming conventions*. http://guides.lib.purdue.edu/c.php?g=353013&p=2378293 [Accessed 12 November 2018].

Quick, JFH. 1997. *New computer control system layout.* HartRAO Internal Communication Discussion Document, October 1997. http://www.hartrao.ac.za/nccsdoc/text/nccs.layout [Accessed 7 January 2017].

Quick, JFH. 2017. *New computer control system*. Notes of a formal discussion with Dr. JFH. Quick, HartRAO's VLBI Manager and Programme Leader: Instrumentation and Computing, 12 June 2017.

Rajasekar, S., Philominathan, P. & Chinnathambi, V. 2013. *Research methodology*. https://arxiv.org/pdf/physics/0601009.pdf [Accessed 7 September 2016].

Ray, JM. 2014. Introduction to research data management, in *Research data management: practical strategies for information professionals,* edited by Joyce M. Ray. Purdue: Perdue University Press, pp. 1–22.

Republic of South Africa. Presidency. National Planning Commission. 2011. *National Development Plan, 2030*. https://nationalplanningcommission.wordpress.com/the-national-development-plan/ [Accessed 27 July 2015].

Riley, J. 2017. *Understanding metadata: what is metadata and what is it for?* Baltimore, MD: NISO. https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf [Accessed 6 March 2018].

Rob, P. & Coronel, C. 2007. *Database systems: design, implementation and management.* 7th ed. Boston, MA: Thomson Course Technology.

Rouse, M. 2014. *Internet*. TechTarget. https://searchwindevelopment.techtarget.com/definition/Internet [Accessed 10 August 2016].

RWTH Aachen University. 2010. *What are research data?* http://www.rwth-aachen.de/cms/root/Forschung/Forschungsdatenmanagement/~lnaz/Was-sind-Forschungsdaten-/?lidx=1 [Accessed 10 August 2016].

Saffer, D. 2006. *Designing for integration: creating smart applications and clever devices*. Berkeley, CA: Peachpit Press.

Saunders, M., Lewis, P. & Thornhill, A. 2009. *Research methods for business students.* 5th ed. Prentice Hall.

Saunders, M., Lewis, P. & Thornhill, A. 2016. *Research methods for business students.* 6th ed. Prentice Hall.

ScienceDaily. 2018. *Astronomy*. https://www.sciencedaily.com/terms/astronomy.htm [Accessed 19 February 2018].

ScienceHQ. 2013. *Characteristics of computer*. http://www.sciencehq.com/computing-technology/characteristics-of-computer.html [Accessed 11 February 2018].

Sciencemag. 2017. *What is a DOI*? http://www.sciencemag.org/site/misc/doi.xhtml [Accessed 11 September 2017].

Searle, A. & Petrachenko, B. 2016. *Operational VGOS scheduling*. International Proceedings of the VLBI Service for Geodesy and Astrometry 2016 General Meeting. Greenbelt, MD: NASA.

Seeber, G. 2003. Satellite Geodesy. 2nd ed. Berlin: Walter de Gruyter.

Seliger, HW. & Shohamy, EG. 1989. Validity and reliability. *Second language research methods,* edited by Herbert W. Seliger and Elana Shohamy. Oxford: Oxford University Press.

Shaffer, CA. 2008. *A practical introduction to data structures and algorithm analysis*. 3rd ed. Blacksburg, VA: Prentice Hall.

Shuttleworth, M. 2018. *Research bias*. Explorable. https://explorable.com/research-bias [Accessed 24 September 2018].

Silverman, D. 2013. *Doing qualitative research: a practical handbook*. London: Sage.

Square Kilometre Array (SKA). 2018. *What is radio astronomy?* https://www.skatelescope.org/radio-astronomy/ [Accessed 17 May 2018].

Strydom, A. & Bezuidenhout, R-M. 2014. Qualitative data collection, in *Research matters,* edited by F. Du Plooy-Cilliers, C. Davis and R-M. Bezuidenhout. Cape Town: Juta, pp. 173–194.

Tamara, D. 2009. *Network+ Guide to Networks*. 5th ed. Boston, MA: Course Technology, Cengage Learning.

Tanase, A. 2014. *All about malware and information privacy*. https://techacute.com/malware-information-privacy/ [Accessed, 24 September 2018].

Taylor Redd, N. 2017. *What is astronomy? Definition & history*. https://www.space.com/16014-astronomy.html [Accessed 4 February 2018].

Techopedia. 2018. https://www.techopedia.com/ [Accessed 11 February 2018].

TechTarget. 2015. *Essential guide. Building an effective data governance framework*. http://searchdatamanagement.techtarget.com/definition/data-management [Accessed 28 March 2015].

Tenopir, C., Birch, B. & Allard, S. 2012. *Academic libraries and research data services: current practices and plans for the future*. Chicago, IL: Association of College and research Libraries.

TechTerms. 2018. *System requirements: definition*. https://techterms.com/definition/systemrequirements [Accessed 20 September 2018].

Terre Blance, M., Durreheim, K. & Painter, D. 2006. *Research in practice: applied methods for the social science.* 2^{nd} ed. UCT Press.

Thakur, D. 2016~~a~~. *Characteristics of computers*. http://ecomputernotes.com/fundamental/introduction-to-computer/what-are-characteristic-of-a-computer [Accessed 9 August 2016].

Thakur, S. 2016~~b~~. *Characteristics of database management system.* http://whatisdbms.com/characteristics-of-database-management-system/ [Accessed 9 August 2016].

Thanh, NC. & Thanh, TTL. 2015. The interconnection between interpretivist paradigm and qualitative methods in education. *American Journal of Educational Science*, 1(2):24–27.

Torge, W. & Muller, J. 2012. *Geodesy*. 4^{th} ed. Berlin: Walter de Gruyter.

Trauth, MH. & E. Sillmann (eds). 2013. Internet resources for earth science data, in *MATLAB and design recipes for earth sciences: how to collect, process and present geoscientific information,* edited Martin H. Trauth and Elisabeth Sillman. Berlin: Springer-Verlag, pp. 51.–74.

Tshwane University of Technology. 2018. *MATLAB & Simulink*. https://www.tut.ac.za/faculties/engineering/matlab/about [Accessed 24 September 2018].

Tufts University. 2013. *Research data management systems (RDMS).* https://sites.tufts.edu/rdms/ [Accessed 14 June 2017].

UNAVCO. 2012. *UNAVCO data policy*. https://www.unavco.org/community/policies [Accessed 9 September 2017].

UNC University Libraries. 2017. *Metadata for data management: a tutorial.* http://guides.lib.unc.edu/metadata [Accessed 14 November 2017].University of Cambridge. 2018. *Organising your data*. https://www.data.cam.ac.uk/data-management-guide/organising-your-data [Accessed 7 February 2018].

University of Colorado Libraries. 2008. *Guidelines on the file naming conventions for digital collections*. Digital Project Advisory Group. https://ucblibraries.colorado.edu/systems/digitalinitiatives/docs/filenameguidelines.pdf [Accessed 3 August 2017].

University of Durham. Department of Computer Science. 2007. *Guidelines for performing systematic literature reviews in software engineering*. https://pdfs.semanticscholar.org [Accessed 4 September 2017].

University of Leicester. 2015. *Naming files and folders*. http://www2.le.ac.uk/services/research-data/organise-data/naming-files [Accessed 12 August 2016].

University of Minnesota Libraries. 2015. *What is data? Glossary of data related terms*. https://www.lib.umn.edu/datamanagement/whatdatais [Accessed 12 May 2015].

University of Tennessee Knoxville. The National Institute of Computational Science. 2018. *I/O and Lustre usage*. https://www.nics.tennessee.edu/computing-resources/file-systems/io-lustre-tips [Accessed 25 September 2018].

University of South Africa (UNISA). 2013. *The UNISA Policy on Research Ethics*. https://www.unisa.ac.za/static/corporate_web/Content/Colleges/CGS/documents/Policy-on-Research-Ethics-rev-appr-Council-20.09.2013.pdf [Accessed 25 September 2018].

University of South Africa (UNISA). 2017. *Policy for Master's and doctoral degrees*. https://www.unisa.ac.za/static/corporate_web/Content/Colleges/CGS/documents/Policy%20on%20M%20and%20D%20degrees%20-%20rev%20appr%20Senate%20-%2029.03.2017.pdf [Accessed 25 September 2018].

University of Washington Libraries. 2014. *Data management guide: organization & format*. http://guides.lib.uw.edu/research/dmg/orgformat [Accessed 11 November 2018].

Valle, M. 2013. *Scientific data management*. Swiss National Supercomputing Center. http://web.archive.org/web/20120306015034/http://personal.cscs.ch/~mvalle/sdm/scientific-data-management.html [Accessed 4 August 2016].

VLBI.org. 2009. *E-VLBI File-naming convention*. http:// www.vlbi.org/filename/docs/049.1b-Filenaming%20conventions.pdf [Accessed 4 September 2016].

Wang, L., Shen, W., Xie, H., Neelamkavil, J. & Pardasani, A. 2002. Collaborative conceptual design: state of the art and future trends. *Computer-Aided Design,* 34:981–996. https://www.researchgate.net/profile/Weiming_Shen2/publication/ [Accessed 22 August 2016].

Watt, A. 2015. Characteristics and benefits of a database. *Database design.* 2nd ed. by Andrienne Watt and Nelson Eng. BCCampus Open Education. http://opentextbc.ca [Accessed 3 August 2016].

Webopedia. 2018. *API: application program interface*. https://www.webopedia.com/TERM/A/API.html [Accessed 24 September 2018].

Welman, JC. & Kruger, SJ. 2003. *Research methodology for the business and administrative sciences.* 2nd ed. Oxford: Oxford University Press.

WhatIsMyIP: The IP Address Experts. 2018. *What is a VPN?* https://www.whatismyip.com/what-is-a-vpn/ [Accessed 24 September 2018].

Wier, S., Rost, M. & Boler, F. 2016. *Dataworks for GNSS: manual for data repositories*. Colorado: UNAVCO.

Wier, S. Boler, F. & McWhirter, J. 2012. Geodesy data repository architecture with GSAC. UNAVCO GSAC WE: Services for geodesy data repositories. Colorado, USA: UNAVCO.

Willis, JW. 2007a. *Foundation of qualitative research: interpretive and critical approaches*. London: Sage.

Willis, JW. 2007b. World views, paradigms and the practice of social science research, in *Foundations of qualitative research: interpretive and critical approaches,* by Jerry W. Willis*.* Thousand Oaks, CA: Sage. http://dx.doi.org/10.4135/9781452230108.n1 [Accessed 13 November 2015].

Williams, J. 2015. *What is data structuring?* https://www.promptcloud.com/blog/What-is-Data-Structuring [Accessed 23 May 2017].

Williamson, K. & Johanson, G. (eds). 2013. *Research methods: information, systems and contexts*. Tilde University Press.

Wynn, D. & Williams, C. 2008. *Critical realism-based explanatory case study research in information systems.* Proceedings of the 29th International Conference on Information Systems, Paris, France, 14–17 December 2008.

Yin, R. 2013. *Case study research, design and methods*. Newbury Park, CA: Sage.

Zhang, Y. & Wildemuth, BM. 2009. Qualitative analysis of content, in *Applications of social research methods to questions in information and library science*, edited by BM. Wildemuth. Westport, CT: Libraries Unlimited, pp. 308–319.

ZDNet. 2015. *To flash or not to flash in the data centre*. http://www.zdnet.com/article/to-flash-or-not-to-flash-in-the-datacentre/html [Accessed 1 July 2015].

APPENDIX A: INTERVIEW SCHEDULE

Dear Participant

In my email requesting this interview I have indicated to you that I am busy with a research project on the design of a new Geodetic Research Data Management System for HartRAO. The aim of this interview is to obtain your ideas, thoughts and opinions regarding geodesy data, data management systems and data management system design. The information obtained will be used only for research purposes and no names of participants will be made known in the report. Are you satisfied with that and do you have any questions before we start the interview?

May I please take notes and audio-record the interview, as it would assist me to make a transcript of the interview for data analysis purposes?

Please answer the applicable questions as open and honestly as possible.

Section A: Space Geodesy Data

1. Do you use space geodesy data for your research?
2. Where do you search for space geodesy data?
3. How do you access geodetic datasets?
4. When searching for data, which database do you use?
5. Are you familiar with the CDDIS and UNAVCO GSAC systems?
6. What type of data do you normally obtain?
7. When searching for a specific dataset, which search type do you use – for example: author/PI searches, keywords, experiment names, observation name, station name, etc.?
8. Do you know where to find HartRAO datasets? If YES, how did you access the data?
9. How did you find the level of access? Very easy, easy, difficult, very difficult?
10. How long did it take you to retrieve the data - $\leq$ 5 min, $\geq$ 10 min?

11. Were the data displayed / presented in an easily understood manner?

Section B: Data Management Systems

1. Do you use a data management system to manage your data?
2. Which data management system do you use?

Section C: Data Management System Design

1. What do you see as the objectives of a data management system?
2. Who do you see as being the stakeholders that will use a data management system? In other words, who has a vested interest in its success?
3. What data/information should a research data management system manage?
4. Do you have experience in data management system design?
5. Should a research data management system have capabilities for interaction with other data management systems?
6. Which components should a modern data management system include to cater for large data volumes?
7. What criteria should be considered for designing the GRDMS?
8. What are the key characteristics of a data management system?
9. How should the data be structured in a geodetic research data management system?
10. What data structures guidelines should be considered during the designing of the GRDMS for HartRAO?
11. Do you know of any conceptual models that will be suitable and should be considered during the design of a GRDMS for HartRAO?
12. To what extent does the design of the GRDMS for HartRAO comply with technological expansions and institutional needs?

APPENDIX B: TRANSCRIPTIONS OF INTERVIEWS

Main category and code	**Subcategory and code**	**Sub-subcategory and code**	**Responses**
Data Management (DM)	Space Geodesy Data (SGD)	Data Usage (U): a. Do you use space geodesy data for your research?	*P1:* Yes, I use mostly SLR and meteorological data. *P2:* Yes, I use SLR data, looking at how to improve accuracy of station positions as well as instrument positions. *P3:* Yes, I use LLR data *P4*: Yes, I use mostly VLBI data, NGS card files. I use the card files for VLBI analysis using VieVS software. *P5*: Yes, I do, but not as much as some of my colleagues. I am more interested in new equipment and the provision of quality data and data products. *P6*: Yes, I use space geodesy data for my research purpose. *P7*: Not much at the moment but I have used geodetic data in the past for my research. *P8*: Yes, but very limited. *P9:* No, I mainly maintain the astronomical data systems at HartRAO. But I am aware and familiar with some of the geodetic data types. *P10:* No, but I am currently writing software for managing HartRAO's geodetic data.
		b. Who do you see as being the stakeholders that will use a DMS, who has a vested interest in its success? (Also relevant to Main category *Data management systems (DMS)* following below).	*P1:* Various researchers and research institutions *P2:* The end-users, which can be students, lecturers, researchers, laymen and curious people, managers of systems, institutions and governing bodies *P3:* Specialized individuals, students, researchers, curious members of the public *P4:* Scientists/researchers, governing bodies and the scientific community *P5:* In space geodesy, the people will be students – graduate and post-graduates, established researchers and most importantly the data service providers to whom data is provided. *P6:* Researchers

Main category and code	**Subcategory and code**	**Sub-subcategory and code**	**Responses**
			P7: Geodesists and meteorologists *P8:* Meteorologists and geodesists, undergraduates etc. *P9*: For astronomy data, it is mostly astronomers and students, so I would think the same would apply for geodesy data. Researchers (national and international) and students will use the system. *P10:* Students (graduate and under-graduates), established researchers and data centres
Data Management (DM)	Space Geodesy Data (SGD)	Data Type (T): What type of data do you normally obtain? (Also relevant to subcategory 2: *Data usage* discussed above).	*P1:* Rinex versions 2 and 3 *P2:* Meteorological raw data including air pressure, humidity, wind direction and speed, station positions, atmospheric correction models and precipitable water vapour data *P3*: Mostly processed data *P4*: Mostly NGS card files *P5:* I am mostly interested in geodetic data (e.g. Rinex format), which are generated by the SLR instrumentation and most possibly the LLR instrumentation in future. *P6:* Geodetic data converted into FITS files *P7*: Meteorological data *P8:* Meteorological data, such as weather data and geographical data of the area surrounding HartRAO *P9:* NA. *P10:* The programs that I am currently developing are for Rinex data.
Data Management Systems (DMS) (also relevant to main category:	Design (D): Do you have experience in data management system		*P1*: Some *P2*: Yes, I do and I have experience in Oracle, MySQL and Access *P3*: Not really *P4*: No

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
Geodetic data management systems (GDMS), following below).	design?		*P5*: Yes, I have experience in designing systems and using systems *P6*: No, not really *P7*: No *P8*: Limited experience *P9*: Yes *P10*: Yes
Data Management Systems (DMS) (also relevant to main category: *Geodetic data management systems (GDMS)*, following below).	Internet (DMS-I)	Search (S): a. Where do you search for space geodesy data?	*P1*: CDDIS mostly, IGS, sonel.org, local repository *P2*: ILRS repository *P3*: On the platform 'Horizon' which is provided by the NASA/JPL *P4*: I find the data via the IVS website and HartRAO servers *P5*: I usually gain access to the data via the space geodesy program URL *P6*: On the IVS website *P7*: HartRAO server *P8*: IVS and HartRAO *P9*: NA *P10*: The existing system provides access via the space geodesy program URL, with links to directories, folders and files.
		b. When searching for data, which database do you use?	*P1*: CDDIS, NASA EarthDATA *P2*: ILRS repository *P3*: NASA/JPL database *P4*: IVS and the CDDIS *P5*: I use the current data management and storage system database 'Geoid'. *P6*: NA *P7*: I access data via proxy from the HartRAO servers. Not sure what database software HartRAO is using. *P8*: Meteorological database and the HartRAO system *P9*: NA *P10*: I use 'Geoid'. The system contains a relational database.
		c. When searching for a specific dataset, which	

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
		search type do you use, for example: author/PI searches, keywords, experiment names observation name, station name, etc.?	*P1*: Station name *P2*: When using the ILRS site, I use keyword searches and station code. *P3*: I normally type in keywords. *P4*: I use experiment names, session names / codes, year. *P5*: I use the URL provided by the geodesy program to access data. *P6*: Experiment name and observation name *P7*: Station name and keywords *P8*: Station name, keywords, author searches *P9*: If I were to use such a system, I would probably use experiment names, station names or PIs. *P10*: I use the geodesy programme website. I do not really use Google or any other search engine to recover data.
		Interface (In): Were the data displayed / presented in an easily understood manner?	*P1*: Medium *P2*: Yes, the system displays the data in a format that can be understood. It is user-friendly. *P3*: Yes, users can specify how they would like the results displayed. *P4*: Yes, if you know what the data is all about, but if you don't know, then it is not easily understood. *P5*: It was okay, I could understand the manner in which it was displayed. *P6*: Yes *P7*: Yes *P8*: Yes, but the format in which it is displayed is old and uninteresting. *P9*: Yes *P10*: Yes, but the filenames given for the datasets are difficult for the layman to understand. The system's display features outdated and boring.
		Access (A): a. How do you access geodetic datasets?	*P1*: FTP and HTTP *P2*: By accessing the HartRAO 'Geoid' server *P3*: Via the Internet *P4*: Via IVS website or by logging onto HartRAO servers

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			P5: Via URL/FTP and HTTP
			P6: By logging onto the HartRAO server
			P7: By logging into the HartRAO Geoid server
			P8: Via the Internet by logging into the IVS http and the HartRAO servers
			P9: NA
			P10: Via HartRAO Geoid servers
		b. Do you know where to find HartRAO datasets? If YES, how did you access the data?	*P1*: Yes, by FTP, HTTP and direct access (staff member).
			P2: Yes, via the ILRS repository. I access all SLR site data from that repository by following the “Data and Product” link.
			P3: Yes, via ‘Horizon’ from the NASA/JPL database. I type in the ‘Horizon’ URL, then click on the ‘Observer’ (station) tab. I then specify what data I need by typing in keywords.
			P4: Yes, I access the IVS website and follow the tabs and links provided by the webpage.
			P5: Datasets for example GNSS. Datasets are stored in a directory. I select the data format and follow the link by selecting year, experiment or scan number, relevant file. I then follow the instructions onscreen to open the file containing the datasets.
			P6: Yes. First I log into the HartRAO server using username and password. Then I look for a specific folder where my data is loaded/saved. From the specific folders, I download the data.
			P7: Yes, on HartRAO servers and sometimes I access the IVS website.
			P8: Yes. I use my username and password to log into the HartRAO server. Then I look for a specific directory on the server. After that I search for the relevant folder where data is stored. I ftp download the data to my computer.
		c. How did you find the level of access: very easy, easy, difficult or very difficult?	*P9*: Yes, if I have to.
			P10: Yes, GNSS data is stored in a directory that contains the datasets. I select the data format and follow the link by selecting the year, experiment/ scan number. I follow instructions provided onscreen to open the file containing the datasets.

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			P1: Relatively easy *P2*: Moderate, because the terminology used on the ILRS website is too specialised. If you are a new student, researcher or layman you will find searching on the site difficult. *P3*: Not too difficult. Can be used by anybody. *P4*: Easy, but not very user-friendly *P5*: Easy *P6*: Easy *P7*: Easy if you know the system *P8*: Easy *P9*: Easy *P10*: Easy
		Retrieval (R): How long did it take you to retrieve the data?	P1: Less than 5 minutes but depends on the type and station P2: Less than a few minutes because I know the discipline and the website well P3: ≤ 5 minutes P4: ≤ 5 minutes P5: It took less than a few minutes to access what I was looking for. P6: Less than 10 minutes P7: Less than a few minutes P8: Only a few minutes, definitely less than 10 minutes P9: A minute or two P10: Less than 10 minutes
Research Data Management Systems (RDMS)	Components (Ca): Which components should a modern data management system include to cater for large data volumes? (Also relevant		*P1*: Automation software, search interface, data retrieval tools i.e. DOIs etc., server clusters with SSDs *P2*: Excluding the obvious components, such as network components, hardware and software, it would be beneficial if a system can have components to cater for storing large volumes of data, e.g. SSDs. *P3*: Large storage space on servers, network devices and large band width *P4*: Big stable storage such as SSDs on servers, proper Ethernet cables and modems.

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
	to subcategory *Criteria (Cc)* following below).		*P5*: Excluding the normal components of a computer system, a modern data management system should have the large and fast processors and stable servers. *P6*: Clusters of servers *P7*: Lots of servers *P8*: Large servers *P9*: SSDs or servers *P10*: All the usual components and large server clusters
Research Data Management Systems (RDMS)	Characteristics (Cb): a. What are the key characteristics of a research data management system? b. What do you see as the objectives of a research data management		*P1*: Proper metadata and structures, links to datasets *P2*: Accuracy, user-friendly interface, modern features, secure access, integrity and simplicity *P3*: It should have active persistent links, toolbars, not too many pop-ups and graphics that can slow the system down, should allow for versioning of datasets, easy to navigate, a 'Help' function, updated regularly, compatible with OS, Mac, etc. use and provide a glossary as well as being accessible via Linux. *P4*: Capture, store, archive, retrieve, display and manage data. *P5*: Multi-user via the Internet. Feature that allows users to register on the system and provides an online interactive interface. Allows multiple data access techniques, creates and manages different user accounts, processing and backups. Must be designed to accommodate user growth and system expansion. Organising and cataloguing datasets. Use data standards to organise and store raw and processed SINEX and RINEX data, data products and ancillary data. Displays data holdings. *P6*: Easy access, easy to navigate and use, reliable *P7*: Reliability, easy to use and modern *P8*: User-friendly, fast, compatible, expandable *P9*: Accuracy, secure access, integrity and reliable *P10*: Remote access, multi-user access, provides an online interactive user-friendly interface, scalable. Allow methods for organising and cataloguing datasets use international geodetic standards to organise and store raw and processed RINEX data, data products and ancillary data (same standards as those followed by the CDDIS).

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
	system? (also relevant to Subcategory *Criteria (Cc)* following below)		*P1*: Organise datasets, metadata, search enabler, data usage statistics *P2*: The system must be able to provide info/data to the end-user in a very simple interactive way. Have features to guide end-users on the use of the system. *P3*: A system should provide efficient access to data, should be updated regularly, managed, maintained, user-friendly. *P4*: Easy and quick access to data. *P5*: To perform the main data management functions, i.e. manage the datasets stored in databases, in structured manner, manage the system itself as well as provide access to users. *P6*: Store and provide access to reliable accurate data. *P7*: The system must provide quality data to the user. Must be easy to use. *P8*: To store and provide data. *P9*: Ingest data and transfer data at high speeds to the correlators and analysis centres. *P10*: Handle main data management activities, such as managing data stored in database, use a standardised structure, provide multi-user access from anywhere in the world.
	c. What data / info should a research data management system manage? (Also relevant to Subcategory *Criteria (Cc).*		*P1*: All geodetic data: GNSS, SLR, VLBI, DORIS, meteorological, seismic as well as various products *P2*: Raw and processed data. By keeping raw and processed data, researcher can conduct further research and discovery *P3*: Cater for different types of data *P4*: Raw and processed data *P5*: Depending on the system, I would think, pre-processed (raw) data, processed data, ancillary data, data products such as plots, data usage, just to name a few. *P6*: All relevant data *P7*: All types of data generated by HartRAO's geodetic instruments

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			P8: Data generated by the instrumentation *P9*: Raw and processed data. Auxiliary files and metadata. *P10*: Raw data, processed data, ancillary data, data products, data usage information, etc.
Research Data Management Systems (RDMS)	Criteria (Cc): a. Should a research data management system have capabilities for interaction with other data management systems?		*P1*: Yes *P2*: Yes, because it allows for improvement in system performance and access, e.g. the CHPC system *P3*: Yes, because, for example, VLBI networks are situated on different geographical sites. Systems need to be interoperable to allow the sharing of data. *P4*: Yes, for example the IERS data management system gets data from CDDIS. *P5*: Yes, I think so, interoperability between systems are important. *P6*: Yes, I think so. *P7*: Yes *P8*: Yes *P9*: Yes, as long as it is set up correctly and has strong security measures to protect the data and the systems. Access should be limited. *P10*: Yes
	b. What criteria should be considered for designing the GRDMS?		*P1*: Standardised structured data, completeness and searchable *P2*: Users need state of the art platforms where they can simultaneously access data. Load balancing is important, reliable, speed, accuracy, user-friendly interface, modern features, secure access and integrity. *P3*: The system should be able to provide current and historical raw and processed data, user-friendly, modern features and maintained, links for querying, e.g. email facilities to handle queries of users and facilities to allow for communication between users and the system manager/ administrator. *P4*: Easy and quick access, log user access (access control and user registration), DOIs to assist with retrieval of datasets, analysis tools, large storage capacity, data stored in a centralised area, retrospective search capabilities, up-to-date data, immediate access, data products and publications

Main category and code	**Subcategory and code**	**Sub-subcategory and code**	**Responses**
			P5: Should be able to receive data from various instruments and store raw and processed data it in a database / repository. Users should be able to send requests for data via the Web. Database software should be able to receive and interpret such requests, retrieve the requested data from the data repository and transfer the data via the ftp/http server to users. Processed data should be stored in an archive and collect and store metadata. Should manage the analysis of data in a matter of seconds. Should have storage devices, such as SSD and servers to store large data volumes. The system should facilitate the effective management of data processes ingestion, transformation, storage, retrieval, presentation and dissemination, manage and execute utilities software, automatically peruse incoming data, retrieve deposited files, extract pertinent metadata, verify content quality (formats and readability) and move files to appropriate archive directory locations. Cater for multiple users' access and share metadata between systems. *P6*: Remote access to datasets in an important consideration. *P7*: Access to the data should be easy and quick. *P8*: People's needs, they must be able to access data from their homes and universities. *P9*: Simultaneous access capabilities. Load balancing, reliable, fast, accuracy and secure access. *P10*: Facilitate the effective management of the flow of data and support processes such as data intake, uploading, exporting, processing, transformation, storage of large volumes of data, retrieval, presentation and dissemination. Allow multiple users' access.
Geodetic Data Management Systems (GDMS): a. Do you use a data management system to manage your data? Why do you use it?			*P1*: Yes, an old in-house system, not 100% functional. It is all that is currently available *P2*: Yes, for my own data I use "Explorer" for structuring files. Easy to use for type of data they work with. *P3*: No, not really, but I do use Excel. This question does not apply to me. *P4*: No, I use the VieVS platform. Software used internationally for VLBI data processing and analysis, ideal for their research. *P5*: My data is managed by using the CDDIS database system. Easy to use, contains complete, accurate raw and processed geodesy data and

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			products. *P6*: Not yet. Question is not applicable to me. *P7*: No, not at this stage, but I have used the HartRAO system previously. HartRAO provides geodetic data to international and national scientists. *P8*: Not really. I am familiar with Microsoft Office Excel and it is good enough for me. *P9*: Not my own data, but HartRAO's astronomical data is managed by the New Computer Control System (NCCS). It is suitable and used by other astronomical observatories. *P10*: Not at present. I am currently developing software. The question does not really apply to me.
b. Are you familiar with the CDDIS and UNAVCO GSAC systems?			*P1*: Yes *P2*: Yes, if needed I use the CDDIS. *P3*: Yes, it is another platform that can be used, but it redirects users to the NASA/JPL webpage. *P4*: Yes, I am familiar with CDDIS but not GSAC. I use mostly CDDIS to access the data I require. GSAC hosts mostly GPS data, which I do not use at present. *P5*: Yes, I am very familiar with both. *P6*: Yes *P7*: Yes I am, but I do not really use it. *P8*: No *P9*: Yes *P10*: Yes, I am familiar with both.
c. To what extent does the design of the GRDMS for HartRAO comply with technological expansions and institutional needs?			*P1*: Automation is crucial, statistics on use and downloads are also required. There are no systems currently used at HartRAO that can provide all the technique-specific data in a central data bank using a standardised system with standardised data structures. *P2*: The provision and access to quality data will improve visibility to the international research community. Having such a system will lead to multilateral collaboration. It might even be an income generating system and establish HartRAO as a centre for excellence. *P3*: It will benefit society and science. It will complete the various data services and puts HartRAO and Africa on the map as a world class data service

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			provider. *P4*: A geodetic data management system for geodetic African datasets, are needed to fill the data service and data provision void that exists in the network of international geodetic data service providers. There are no systems at HartRAO that can provide all geodetic data. *P5*: The network of remote observation stations, which consists of GNSS similar to GPS reference systems, seismic monitors and meteorological equipment is currently being expanded. This will densify the SADC network to enhance the science which can be done as well as add new data product capabilities. We are also developing a SA laser ranger, which will eventually be able to range to the Moon and also participate in novel experiments like intercontinental time transfer by laser link and deep space communications. Integrated and automated software tools for data analysis, as well as a set of new data products are under development which will further expand HartRAO's scientific capabilities. Data of high quality and accuracy in long term datasets are always greatly valued by the global research community. *P6*: With the new changes and developments at HartRAO, a new data management system that can manage and make geodetic data accessible, fast and easy will be of great benefit to the organisation and the international scientists. *P7*: Having a modern easy to use system will be beneficial to local as well as international geodesists. *P8*: A new data management system that has the capabilities to host all the different geodetic data in a centralised area will be of great benefit to the researchers and students. *P9*: At present, geodetic data is generated by instrumentation onsite and offsite. Most of the data is transferred to the IVS, CDDIS, etc. With the new expansion of geodetic instrumentation and geodetic capabilities, HartRAO will need an independent system to manage the different technique-specific data types. A system dedicated to managing all HartRAO geodetic data will improve HartRAO's collaborations and services to the scientific community. *P10*: The system will centralize all geodetic specific data types and will allow for more usage and visibility of geodetic data generated by HartRAO's instruments and collaborations. This will be of benefit for geodetic

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			research nationally and internationally.
Data Structure and Organisation (DSO): How should data be structured in a geodetic research data management system?			*P1*: Relatable in some international standard, e.g. the CDDIS *P2*: A hierarchical structure with capabilities for conducting different queries for different data. A standardised structure that is scalable will be beneficial. By having a standardised structure, problems with querying for different data products and raw data can be avoided. *P3*: Not sure, but I would guess in table form. *P4*: In a standardised structure according to technique, year, experiment name and then according to data type (e.g. NGS card files, auxiliary files, metadata, etc.) *P5*: I would think the data should be structured according to technique first, followed by data type, station, frequency of observation or scan, by year, day of year and lastly by filename and compression format. It is like a hierarchical structure. *P6*: Not sure *P7*: In directories, folders and files *P8*: Directories, folders, files, etc. *P9*: A hierarchical standardised structure that is scalable *P10*: In a hierarchical structure according to e.g. technique / type / station / observation / filename and compression type.
	Digital Object Identifiers (DOI) and File Naming convention (FNC): What data structure guidelines and tools should be considered during the design of the GRDMS for HartRAO? (also relevant to Main category: *RDMS* – subcategory 1: *Characteristics (Cb)*)		*P1*: CDDIS structure compliance *P2*: The requirements for data will determine and guide the structure. *P3*: Not sure *P4*: Don't know *P5*: For the supplying of data, an online data archive should be considered and for easy access a catalogue of pre-processed data, analysed data, ancillary data and data products should be provided. I propose the use of guidelines similar to that used by the CDDIS and the NRF for managing the datasets. To support data organisation and structuring, I recommend the use of the file structure as previously mentioned and similar file naming. Tools such as persistent identifiers, e.g. DOIs and naming conventions should

Main category and code	Subcategory and code	Sub-subcategory and code	Responses
			also be considered. *P6*: Not sure *P7*: Those used by the CDDIS *P8*: Not sure *P9*: Guidelines of a hierarchical standardised structure *P10*: I propose the use of that used by the CDDIS. DOIs for dataset should be considered for organising data. Remote access to the data archives via the Internet should be considered when structuring and organizing the data (i.e. how will the data be accessed). Metadata standards and file-naming conventions can also be used.
Conceptual Models : Do you know of any conceptual models that will be suitable and should be considered during the design of a GRDMS for HartRAO?			*P1*: Yes, CDDIS and UNAVCO *P2*: No *P3*: Not really *P4*: Yes, I am aware and have seen the CDDIS model. The CDDIS model illustrates the manner in which technique-specific geodetic data is managed and the architecture of the CDDIS system. *P5*: Yes I would think that because we are providing data to the CDDIS, their model should be considered for the GRDMS. UNAVCO's architectural model can also be considered as they also receive and manage geodetic data. These proposed models can be used as guidelines but we would have to design a model that will be suitable for HartRAO and still serve the research community and data service providers such as the CDDIS, ILRS etc. *P6*: Yes, CDDIS and UNAVCO might be applicable seeing that both systems cater for geodetic data. *P7*: No not that I can think of. *P8*: Not sure *P9*: Yes, CDDIS *P10*: Yes, the CDDIS architectural model

APPENDIX C: CONSENT FORM

INFORMED CONSENT

Full Title of Project: Designing a Geodetic Research Data Management System for HartRAO

Name of Principle Investigator: **Glenda Coetzer**

Please initial the box

1. I confirm that I understand the purpose of the interview and have had the opportunity to ask questions. ☐
2. I understand that my participation is voluntary and I am free to withdraw at any time without giving any reason. ☐
3. I understand that the information I am voluntary providing may be looked and used by responsible individuals or regulatory authorities where it is relevant to my taking part in this research. ☐
4. I give permission to the PI or any other relevant parties to access and use the information supplied by me.
5. I agree to take part in the above mentioned study. ☐

______________________	______________________	______________________
Name of Participant	Signature	Date

______________________	______________________	______________________
Principle Investigator	Signature	Date

1 copy for participant, 1 copy for Principle Investigator