AVN Training, 5 – 30 May 2019, HartRAO, South Africa

SCIENTIFIC DATA MANAGEMENT SYSTEMS: AN OVERVIEW

Glenda Coetzer

[1]Hartebeesthoek Radio Astronomy Observatory / SARAO, South Africa

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

NRF National Research Foundation | SARAO South African Radio Astronomy Observatory

Presented on the 15th of May 2019

OUTLINE

- Data, information & knowledge
- Research data life cycle
- Data management: why should we manage our data?
- Data management systems (DMS)
- Data structuring: hierarchical data structures and data structuring tools & methods
- Astronomy and geodesy DMS architectural models
- HartRAO's new geodesy DMS: architectural model, data structuring, data search & retrieval tools and interactive GUI front-end

DATA, INFORMATION & KNOWLEDGE

- **Data:** Facts concerning people, objects, entities, etc.
- **Information:** Data presented in a form suitable for interpretation. *DMS programs and queries convert data into information.*
- **Knowledge:** Insights into appropriate actions based on interpreted data.

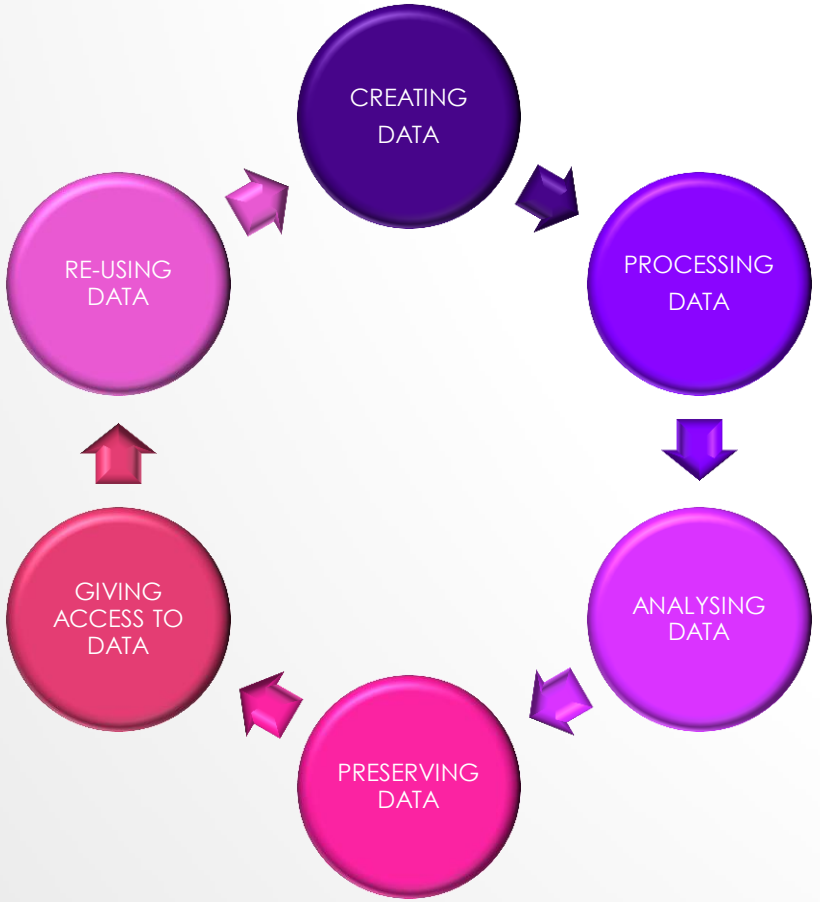
WHAT IS RESEARCH DATA?

"Research data, unlike other types of information, is collected, observed, or created, for purposes of analysis to produce original research results" (University of Edinburgh, 2017),

Types of research data:

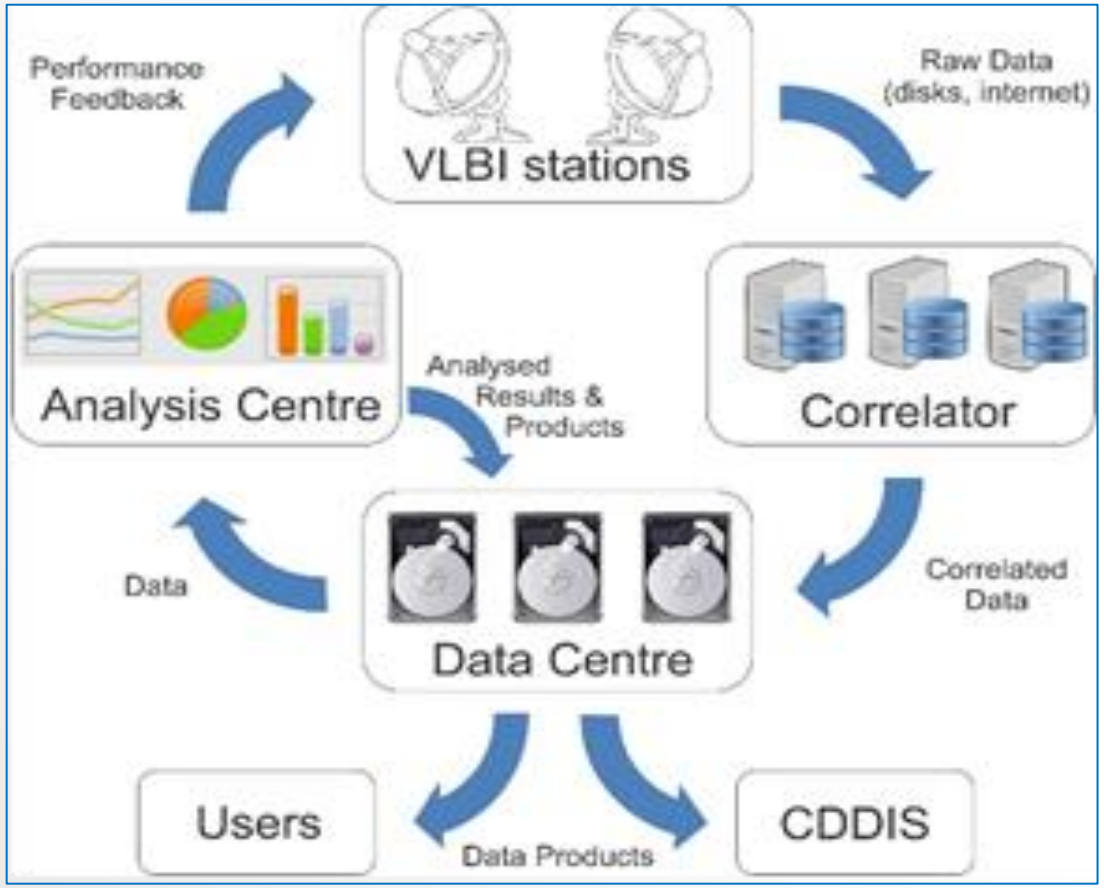
- **Observational data:** captured in real-time, usually irreplaceable.
- **Experimental data:** generated by lab equipment, often reproducible.
- **Simulation data:** generated from test models
- **Derived or compiled data:** 3D models, compiled database, etc.
- **Reference or canonical data:** a (static or organic) conglomeration or collection of smaller datasets, published and curated.

RESEARCH DATA LIFE CYCLE



ASTRONOMY & GEODESY DATA FLOW

Data flow between the various components of the VLBI scientific technique



Coetzer et al. 2015:24

DATA MANAGEMENT

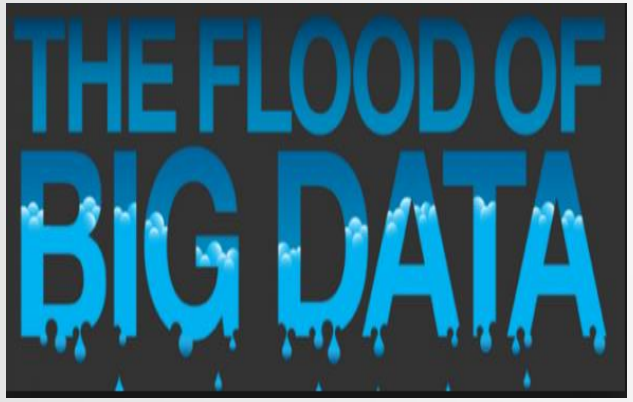
WHY SHOULD WE MANAGE OUR DATA?

Reason 1

DATA DELUGE

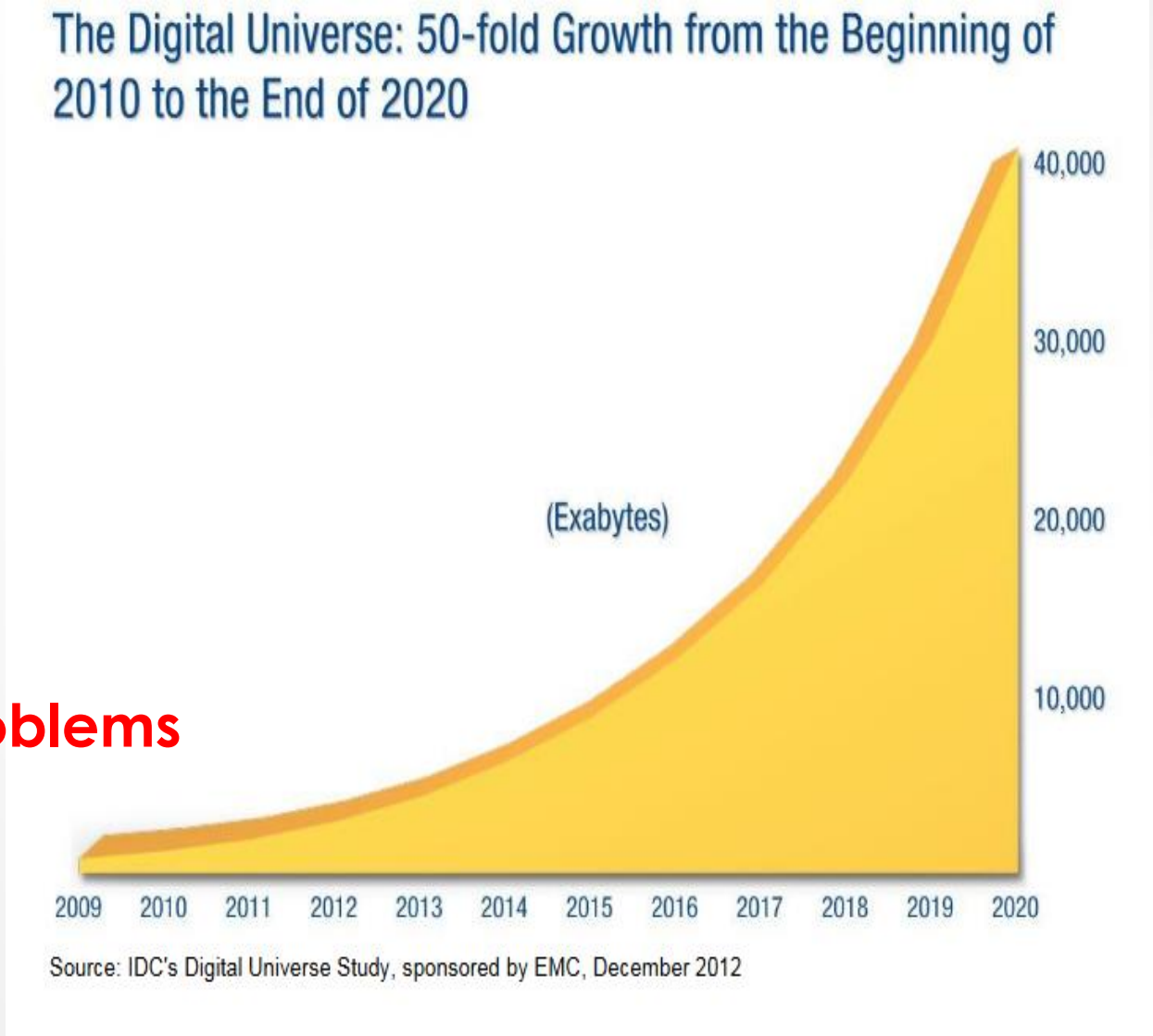
Data collected from scientific instruments, networks, observations, etc. calls for increased attention to *data management, data management systems (DMS)* and data stewardship.

Reason 2



SINK OR SWIM…

Storage problems



Reason 3

DATA LOSS

- Natural or disaster
- Facilities infrastructure failure
- Storage failure
- Server hardware/software failure
- Application software failure
- Format obsolescence
- Legal encumbrance
- Human error
- Malicious attack by human or automated agents
- Loss of staffing competencies
- Loss of financial commitment & stability
- Changes in user expectations & requirements

Reason 4 COMPLIANCE, TRANSPARENCY & RETURN ON INVESTMENT

- Funding agencies & governments open data mandates
- Data are a valuable commodity

NRF SARAO
National Research Foundation | South African Radio Astronomy Observatory

science & technology
Department:
Science and Technology
REPUBLIC OF SOUTH AFRICA

Reason 5

SCIENTIFIC PROGRESS

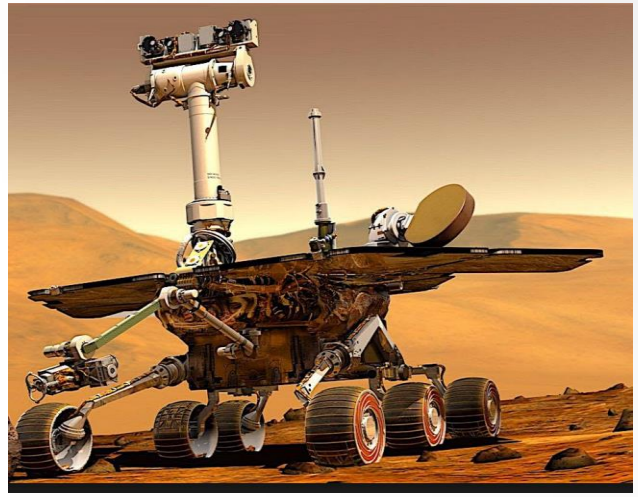
How we use astronomy data in our day-to-day lives -

- Historically: measuring direction and time, navigating the oceans, planting crops, etc.
- Today, scientific and technological development in astronomy, especially in optics and electronics, have become essential to our day-to-day life, with applications such as personal computers, communication satellites, mobile phones, Global Positioning Systems, solar panels and Magnetic Resonance Imaging (MRI) scanners.

"Astronomy data have been a cornerstone of technological progress throughout history, has much to contribute in the future, and offers all humans a fundamental sense of our place in an unimaginably vast and exciting universe" - Dave Finley, 2013.

Geodesy data spin-offs:

- Orientation and navigation.
- Space navigation, telecommunication and exploration.
- Earthquake and tsunami early detection and monitoring.
- Climate change monitoring.
- Housing, agriculture & socio-economic development.



Mars Rower (NASA 2019)

Crop growth (Farmers Business Network 2019)

Alaska earthquake (ABC 7 News 2019)

Polar bear on melting ice (IMF 2019)

Tsunami (NDEA 2011)

DATA MANAGEMENT SYSTEMS

- Computerised software system that facilitates the creation, maintenance and use of electronic databases, designed to perform data management activities (IBM Knowledge Centre 2010).
- Components: hardware, software, data, procedures and processes.
- Types: commercial, academic / research, social, personal, etc. -
  - Astronomy / radio astronomy DMS = international correlators. HartRAO Single-dish multi-wavelength radio astronomy data are managed by the HartRAO NCCS.
  - Geodesy DMS = *CDDIS, UNAVCO* (international data service systems) and *HartRAO's GRDMS* (new regional data service system, *under construction*).

DATA STRUCTURING

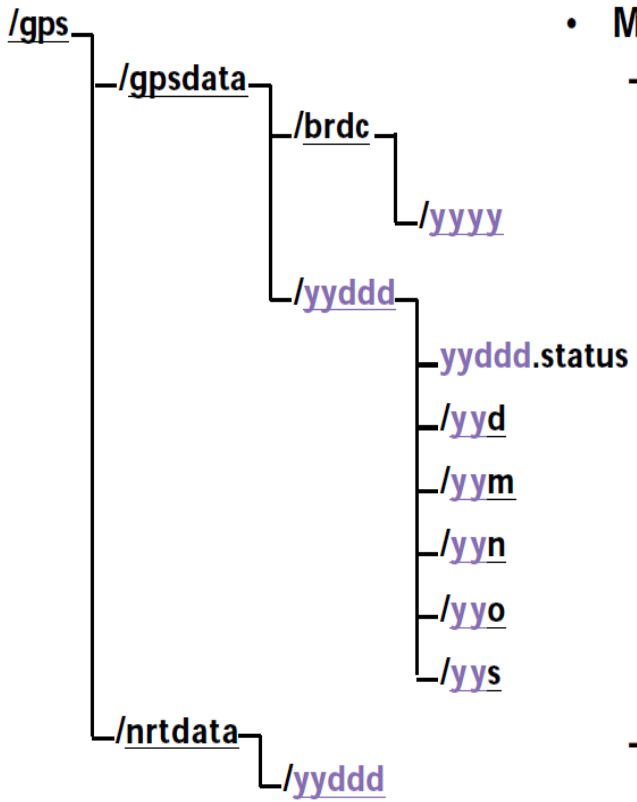
Within system databases -

- ***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.
- ***Trees:*** data structured in a hierarchical manner, consisting of one or more data nodes, e.g. *root/parent*; each node can consist of *zero* or more *sub/child* nodes.

HIERARCHICAL DATA STRUCTURES

- Research data is generally stored in *hierarchical structure* in directories.
- Directories are usually identified by a unique directory name and classified as *root directories*, which contain *subdirectories/parent directories*.
- Within these directories and subdirectories, data are stored in various *folders, subfolders* and *data files*.
- Files of the same nature are usually stored in the same directory.

EXAMPLE OF A HIERARCHICAL STRUCTURE



- **Main GPS filesystem**
  - **GPS daily data subdirectory**
    - Concatenated broadcast ephemerides
      - Yearly subdirectory (yyyy is year)
    - Daily GPS data subdirectories by year (yy) and day of year (ddd)
      - Daily GPS summary file
      - Compact RINEX observation files
      - RINEX meteorological data
      - RINEX broadcast navigation data
      - RINEX observation files
      - TEQC summary files
  - **GPS hourly data subdirectory**
    - Hourly GPS data subdirectories by year (yy) and day of year (ddd)

Noll 1999

DATA STRUCTURING TOOLS & METHODS

File Naming Conventions (FNCs), metadata schemas, Digital Object Identifiers (DOIs) and Open Research and Contributor IDs (ORCIDs)

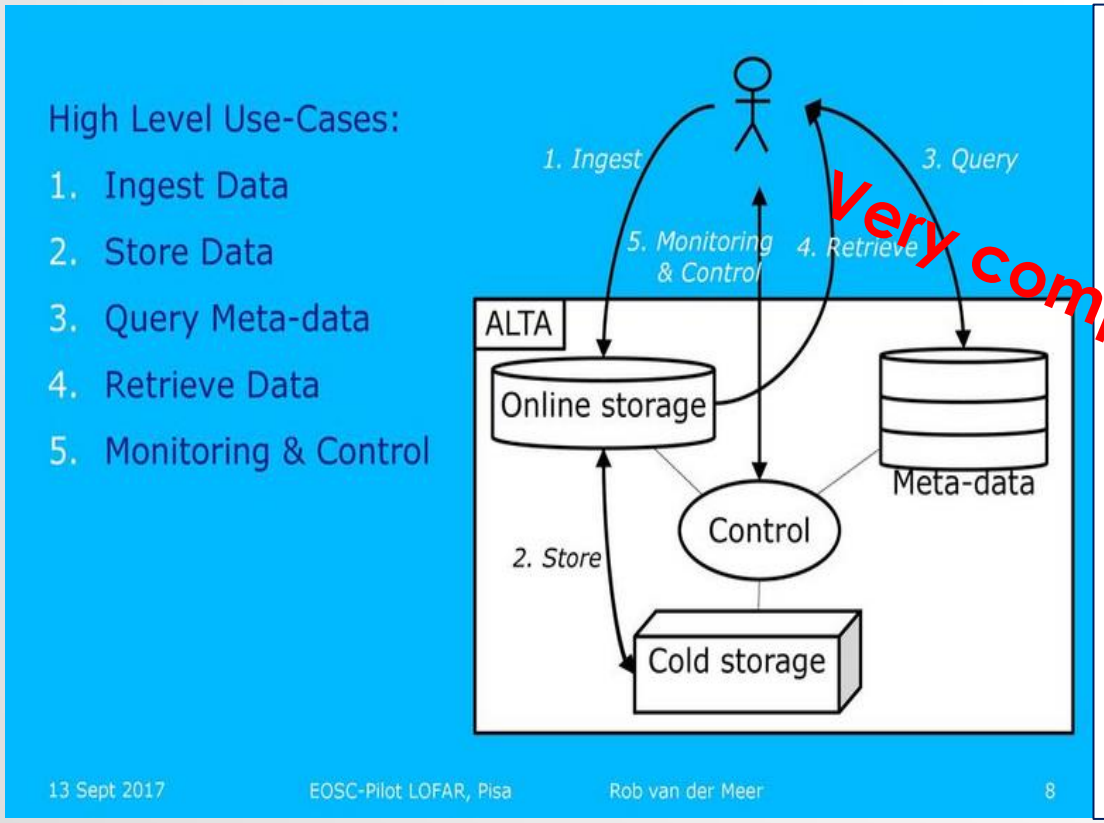
- ***File naming conventions (FNCs)*:** frameworks for naming files in a manner that delineate their content and the way in which they relate to other files, the *8.3.Z FNC* used by IERS and CDDIS for geodesy data.

- ***Metadata schemas*:** uniform sets of ground rules, for the use and management of metadata – specifically with regards to semantics, syntax and optionality of data values. Example of a metadata scheme used in different scientific disciplines: *Dublin Core* and the *IAU Astronomy Visualization Metadata (AVM) standard* (Noll 2015).

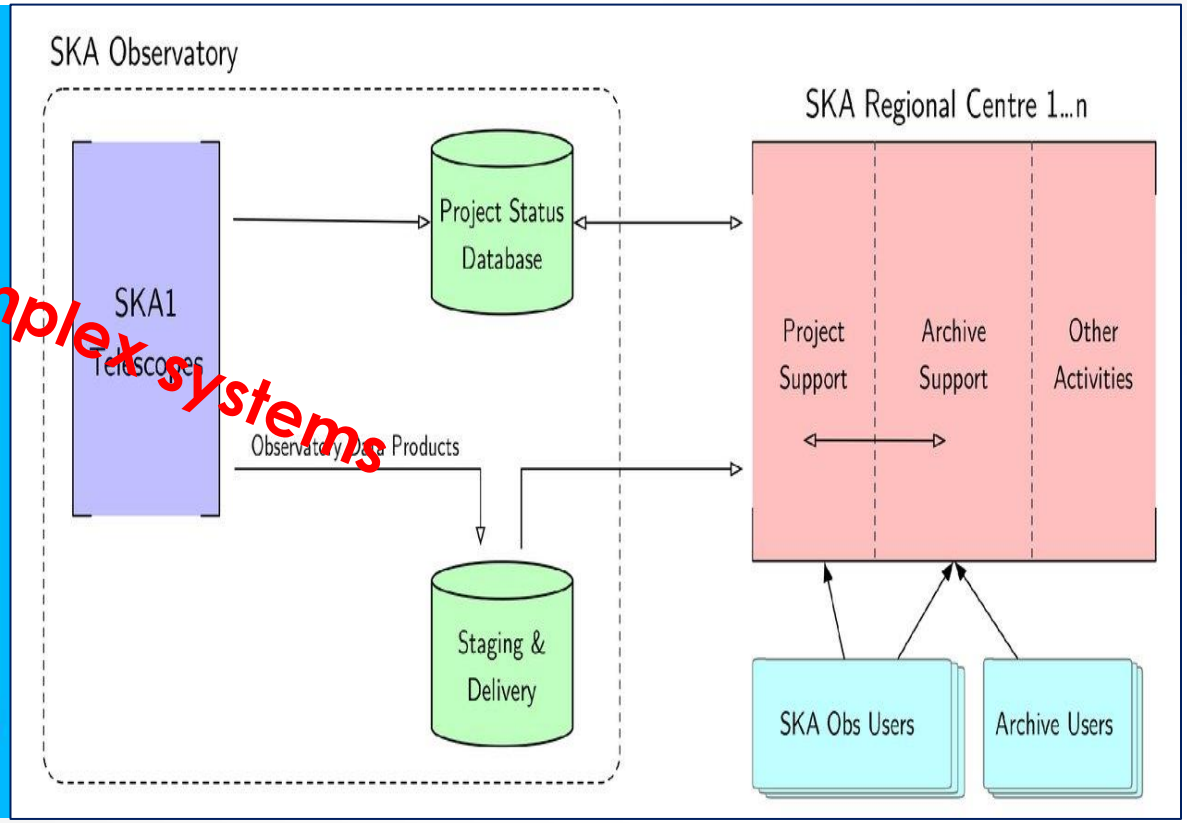
- ***Digital Object Identifiers (DOIs):*** unique persistent alphanumeric identifiers associated with a specific piece of intellectual property; specifies content of object rather than its location -
  - e.g DOI for accessing registered earthquake event research datasets is as follows: *doi:10.1594/GFZ.GEOFON.gfz2009kciu*
- ***Open Research and Contributor IDentifier (ORCID):*** a non-proprietary alphanumeric code used to link research outputs to its true author or originator, and consist 16-digits.
  - e.g. author's ORCID as it may appear on the ORCID platform: *0000-0002-1825-0097* and *https://orcid.org/0000-0002-1825-0097*

ARCHITECTURAL MODELS OF ASTRONOMY DMS

LOFAR & SKA



Rob van der Meer 2017

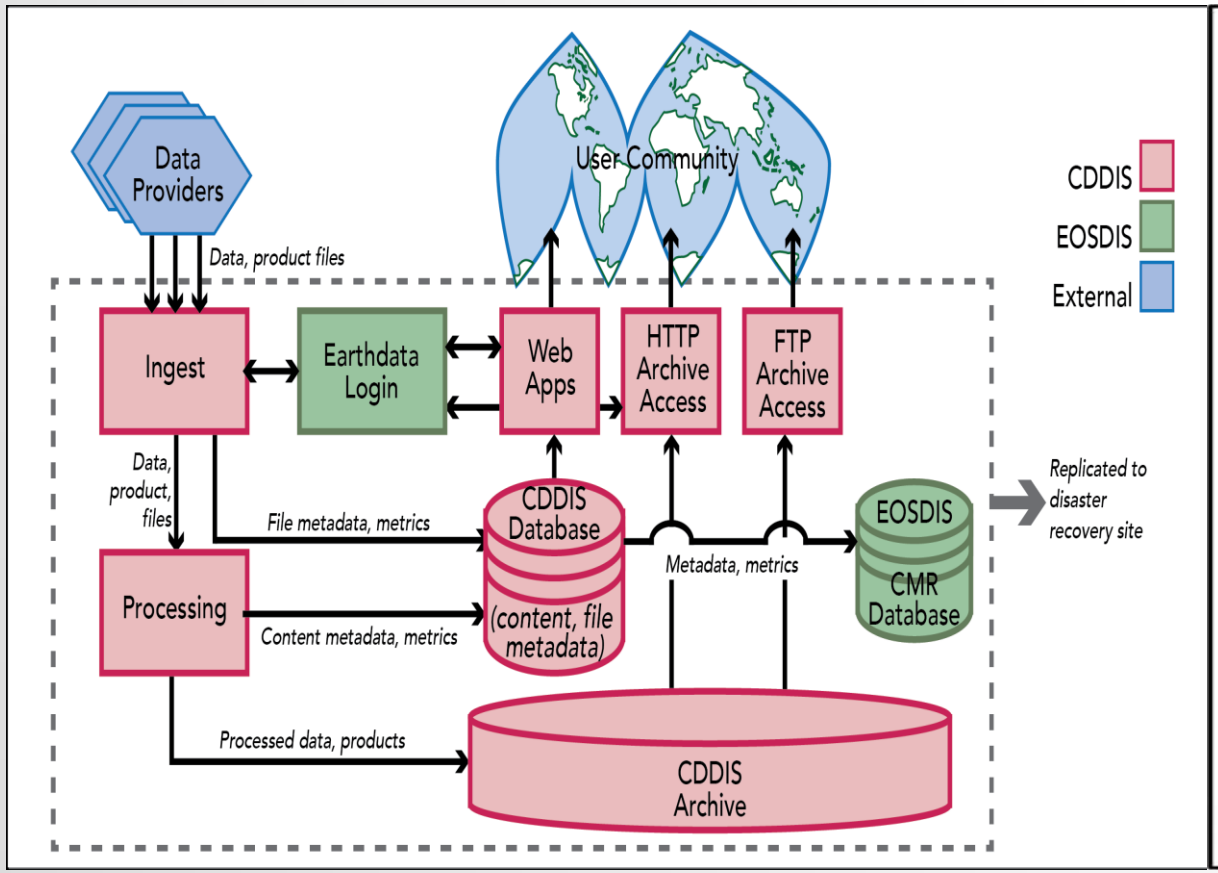


Chrysostomou, et al. 2018

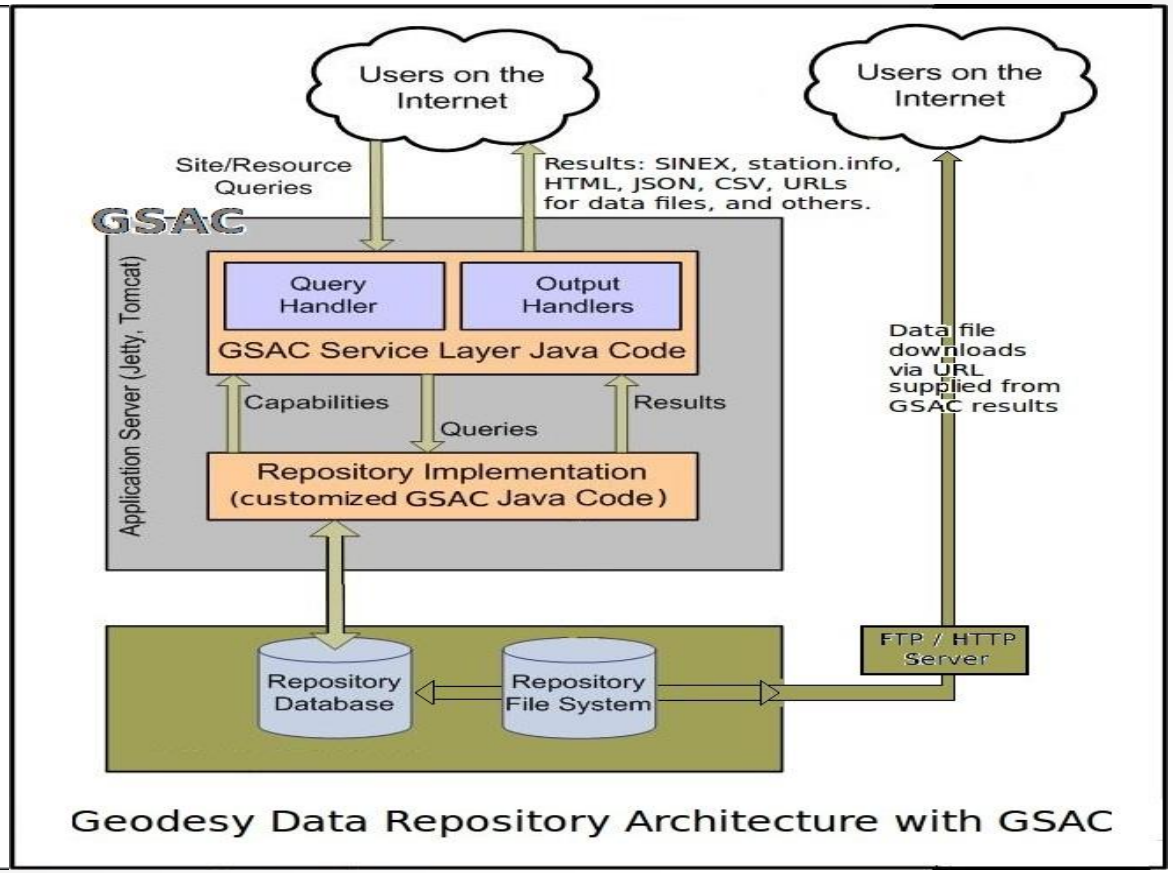
Very complex systems

ARCHITECTURAL MODELS OF GEODESY DMS

CDDIS & UNAVCO

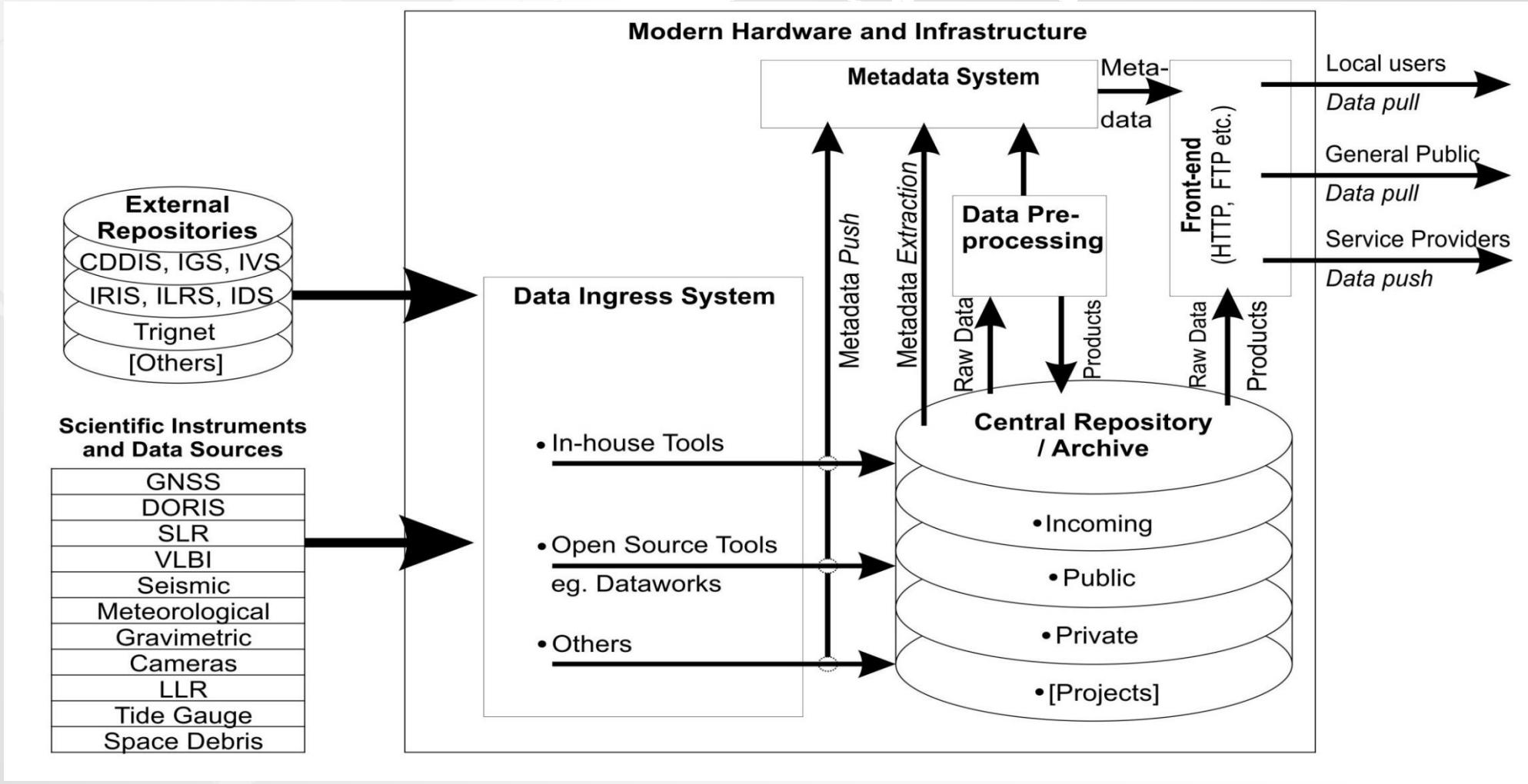


Noll 2017

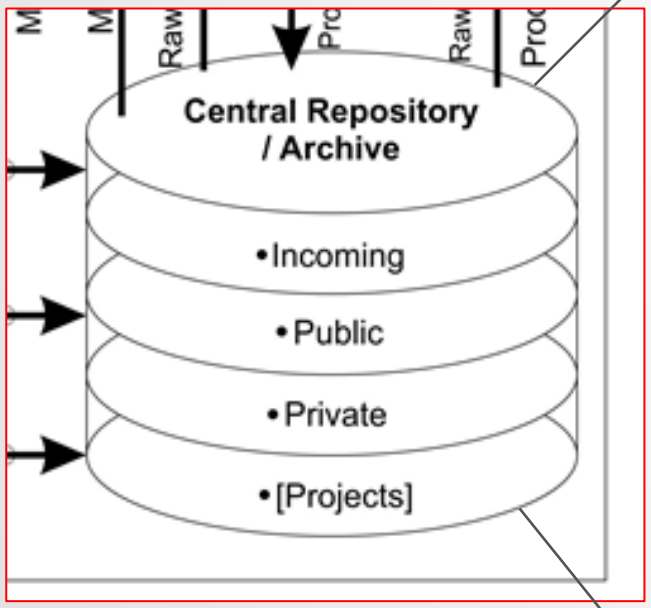


Wier, et, al. 2012

ARCHITECTURAL MODEL OF THE HARTRAO'S GEODESY DMS



DATA STRUCTURING WITHIN THE HARTRAO GRDMS



CDDIS structure for raw technique-specific Rinex data

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

8.3.Z FNC 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 year	17
Z	compressed?	.Z

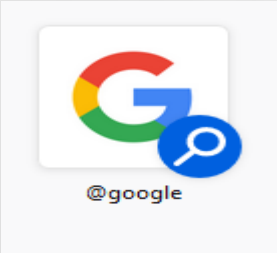
Example: **GNSS/rinex/HRAO/daily/2017/002/HRAO00ZAF_R_20170010000_01D_30S.rnx.zip**

DATA SEARCH & RETRIEVAL TOOLS

Enablers / tools e.g. *DOIs* and *ORCID* -

- DOIs
- Site (e.g. HRAO)
- Instruments (e.g. dish / antenna / receiver, etc.)
- Experiment group (e.g. CRDS)
- Datasets (e.g. raw Rinex files)

added to existing access paths = endless possibilities for data retrieval and discovery



ORCIDs

Site (e.g. HRAO)

Instruments (e.g. dish / antenna / receiver, etc.)

Experiment group (e.g. CRDS)

Datasets

ORCIDs:

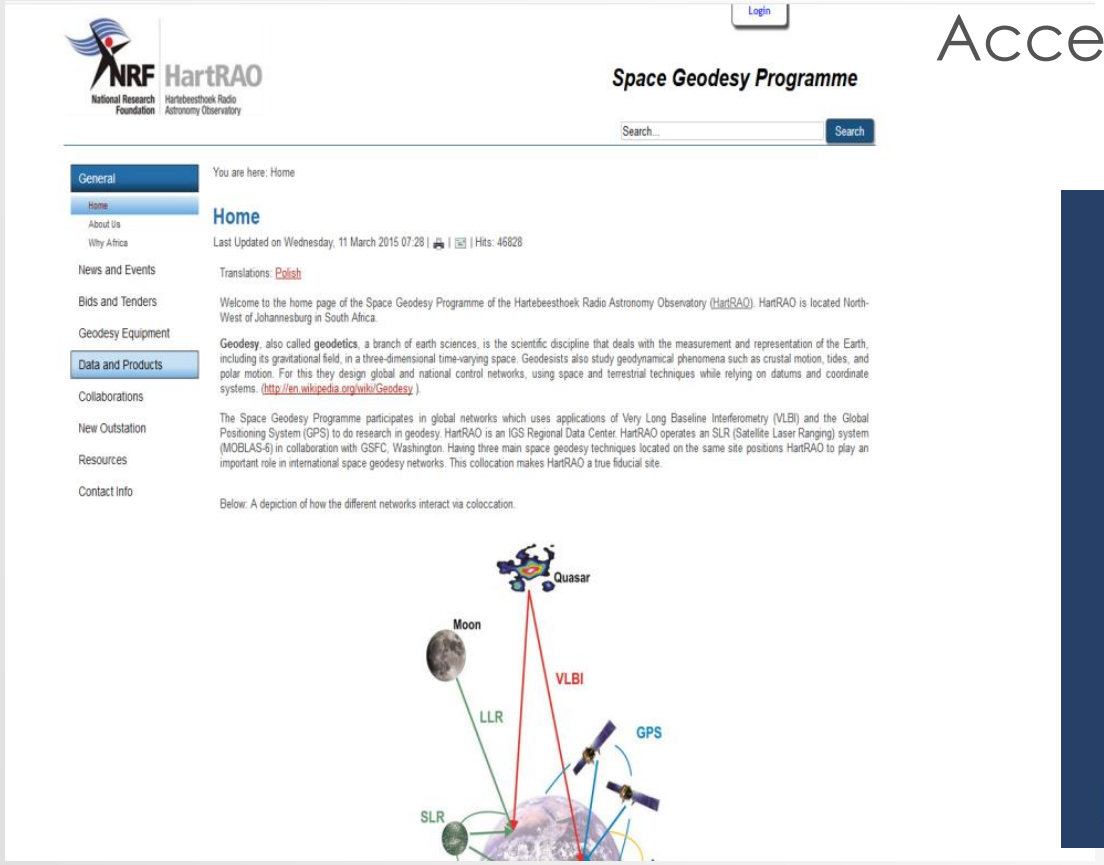
Authors (each having their own ORCID)

Scientific output (e.g. publications / articles, etc.) 1≥ ORCID / publication

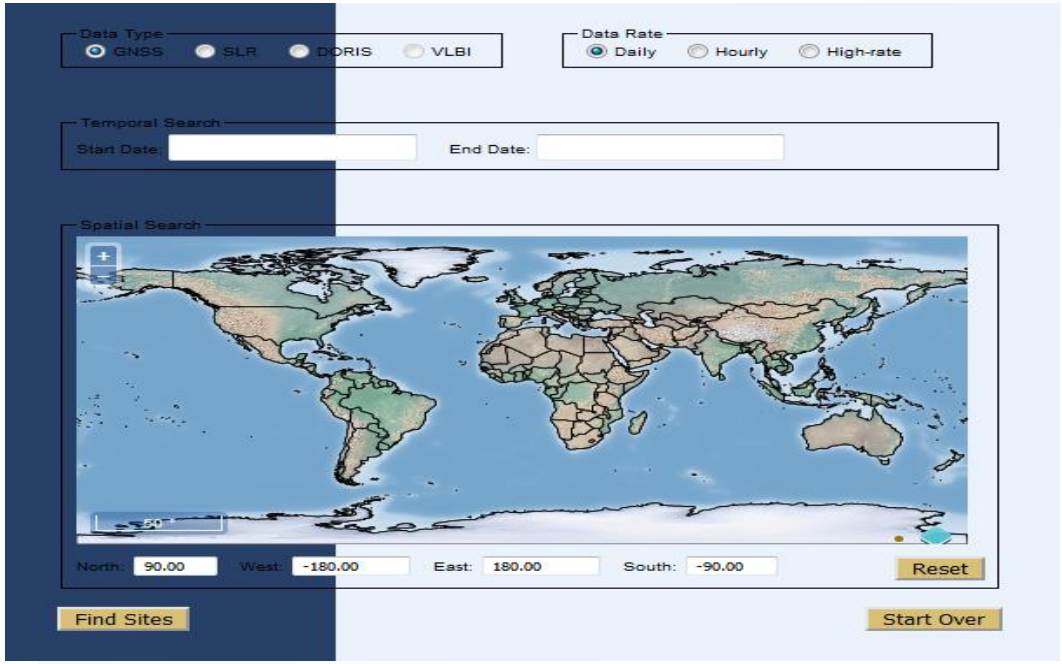
added to existing access paths = **EVEN MORE**
endless possibilities for data retrieval and discovery

INTERACTIVE GUI FRONT-END

***HartRAO Space Geodesy Programme* webpage with data discovery and search tools**



Access via anonymous ftp, http or URL



Work in progress!

THE END ….

THANK YOU !