South African Astronomical Observatory Data Archiving for Multi-wavelength Astronomy Research

Lucian Botha

October 2023

Thesis presented for the degree of Master of Science
in the Department of Computer Science
UNIVERSITY OF CAPE TOWN

Supervisor: Prof. R. Simmonds

Abstract

This research investigates the best practices for creating a data archive system that can store optical and radio astronomy data as well as interact with other astronomy archives. Archives are built and maintained by staff that specialise in big data storage, information security, database administration, and general information technology that provide platforms for access to research data to the broader astronomy community. The tools and services provided by the IVOA allow data to be located, compared, and retrieved, combining multi-spectral and multiple instrument data from different archives to show the evolution of data interoperability across wavelengths. The diverse methods applied and the organisation of resources and collaboration between different facilities to reach common goals will broaden the scope and availability of data throughout the research community. Using data from various archives makes multi-wavelength investigations a significant topic in astronomy. Astronomy data archives employ a variety of authentication mechanisms, including Open Authorization (OAuth), Security Assertion Markup Language (SAML), Central Authentication Service (CAS), and traditional web-based authentication with usernames and passwords. Challenges faced by archives include the increased complexity of data which can be addressed by choosing the appropriate level of standardisation. As astronomy and data science are rapidly evolving, it is a challenge to keep up with all of the new developments, so there needs to be a strategy of research and development so that new technologies and techniques can be adopted.

Acknowledgements

I want to express my gratitude to my wife Lisl, and my children Simoné, Leano, and Kyle for their support. To my supervisor, Robert Simmonds, I want to express my sincere gratitude for his insight and tenacity throughout the research. My research abilities have substantially improved thanks to your comments and suggestions.

Plagiarism Declaration

I, Lucian Botha, know the meaning of plagiarism and declare that all of the work in the document, save for that which is properly acknowledged, is my own.

Contents

List of Figures

Chapter 1

Introduction

1.1 Introduction

This research forms an investigation into the best practices used to build a data archive system to host optical and radio astronomy data and be interoperable with other astronomy archives. With the increase in the amount of astronomy data volumes, this investigation benefits the South African astronomy community and will enable astronomers to improve multi-wavelength data discovery. By following best practices performed by other Astronomy Data Centres, I intend to produce a comprehensive study of what is required to be adopted into the Virtual Observatory (VO). This introduction to the VO is a perfect opportunity to adopt the new vision of interoperability by the IVOA.

The South African Astronomical Observatory (SAAO) conducts fundamental research in astronomy and astrophysics by providing a world-class facility to promote astronomy and astrophysics in Southern Africa. The SAAO observing station is host to the Southern African Large Telescope (SALT) and other telescopes from collaborations and partners across the world. These telescopes can potentially produce 1 TB of data per night. Astronomy data growth will accelerate in the coming years with projects like the Square Kilometer Array (SKA) which will come into operation in the next 5 years with much higher data acquisition rates than what we are used to at existing facilities. The MeerLicht telescope, used to do simultaneous optical observations of objects pointed to by the MeerKAT radio telescope, would directly influence the ever increasing data volume. One of the biggest problems is that currently there is no central place for an observer to have a quick view of the existing observations. An archive will provide SAAO users (astronomers) the feature of searching metadata of observed data which could be used for future research projects and possible discoveries which were overlooked by others.

1.2 Problem Statement

The aim of this research is to identify the best practices for building astronomy data archives and show interoperability between observatories and multi-wavelength data, which will be the initial step for the South African astronomy community towards becoming VO compliant.

1.3 Research Questions

I, under the supervision of Prof. Rob Simmonds, am investigating the requirements of the SAAO to create a data archive prototype. The following research questions will be investigated and hopefully answered.

1. How to build a searchable database of the observations taken by the SAAO telescopes?
2. With the non-standard Flexible Image Transport System (FITS) header information produced by instruments used at the SAAO, how can the data be normalised for mapping into the VO?
3. Which best practices need to be followed to become interopable?
4. How can an archive be interoperable with other observatories and multiwavelength data?

1.4 Methodology

The research will investigate what techniques and methods of implementation, user experiences, support, and services that worked best for other astronomy data archives around the world. By conducting interviews with archive staff, developers and user of such system I intend to provide an in-depth understanding of what is needed to build an archive which contributes positively to the astronomy data archive community. The focus areas of the interviews include data ingestion, hardware, software tools, services, users experience, interoperability, storage, and querying.

1.5 Thesis Overview

This thesis begins with an introduction of the SAAO, SALT and the IVOA and the reason for doing this research. The problem statement, research questions, and methodology that details the procedures used to carry out this study follow the introduction. An understanding of astronomy, particularly the collection, organization, storage, and access of data throughout time. Data management, multi-wavelength astronomy, virtual observatories, and data archiving are introduced in Chapter 2. Interoperability is introduced in Chapter 3 along with its definition and astronomical applications. Chapter 3 also introduces the VO architecture

and the various data formats and describes the interoperable tools and services that are available in VO applications at various levels of the VO architecture. Chapter 4 examines current observatories and how they handle their data archives. I document the experiences of both archive users and the support personnel at some of the well known observatories that are housing data archives in Chapter 5. In Chapter 6, the thesis is summarized and a look at how this study may be expanded to cover other archives is given. It is also discussed how technology, procedures, data formats, and standards have advanced since this research's conclusion.

Chapter 2

Background

Astronomy is one of the oldest sciences with a rich history of data collections by amateurs who recorded their findings of observations of the night sky. The observational category of astronomy is concerned with the acquisition of the data from observations of objects by using telescopes and instruments with different capabilities to acquire the data. Due to the change in seeing, the astronomical atmospheric conditions, it's not always possible to make follow up observations of the same object with the expectation of obtaining the similar data quality. That coupled with the variation of the objects' physical characteristics are the key factors for building data archives, that can preserve each of these observations and make it searchable for the rest of the astronomy community through data archives.

This chapter gives an overview of the importance of astronomy data archives and how it forms a catalyst for other components such as the VO, interoperability across multi-wavelength astronomy and data management. The chapter concludes with a discussion on the relevance of astronomical data archives, how the VO evolved and how data management is impacting the existence of archives.

2.1 Astronomy Data Archives

Data is archived to keep a record of historical observations for further analysis or any variability over time. Archiving allows for research to be performed on the collection of information which could include the study of the evolution of objects observed over long periods. Data archives can serve as a useful educational resource that can be used as a teaching tool for students. Astronomy students can do projects like determining the temperature of a star from its spectral energy distribution or calculating the radius and surface brightness distribution of a galaxy which could be done by using archived data. This could bring about new research ideas that could unlock statistical revelations about the fields of time-domain analysis and multi-wavelength astronomy. Data collection at observatories has increased over the years as new telescopes and instruments were built and deployed for research programs.

Astronomers use particular data archives based on the types of data that they host. Astronomers are often searching for data to complement what they already have. One example of this is to obtain spectra if they already have an image or all the images of a particular field taken in different wavelength ranges or images taken with instruments at longer exposure times. Data archives have evolved from just data collection to a search space that provides tools and services that allow users to do data discovery and data access via platforms such as websites and APIs. A good balance between interactive services such as web services, database query services, and APIs to make data easily accessible is known to be some of the features of modern astronomy data archives. The ability to be able to do multiple object search, cone search, image previews, and name resolvers are qualities that users look for when making use of data archives. The overall approach of an archive can be to adhere to the FAIR principles, i.e. to make the data Findable, Accessible, Interoperable and Reusable. The following sections gives and overview of some of the science centers created to provide the data archive service to the wider astronomy community.

2.1.1 Astronomy Science Centres

The Strasbourg astronomical Data Centre (CDS) is host to Simbad [Simbad, 2020], a database containing identification records of astronomical objects, citation links, and bibliography references. The CDS hosts different services in use across global astronomy data centres. These services include desktop applications, online services, and tools that astronomers use for their daily tasks like Vizier and Aladin. Vizier is a catalog service containing spectra, time series, and images [Vizier, 2020]. Aladin is a visualization and image database application that accesses different sky surveys [Aladin, 2020].

The Infrared Processing and Analysis Centre (IPAC) is host to Infrared Science Archive (IRSA) and Keck Observatory Archive (KOA) which are archives from different NASA missions in its quest to explore the universe. KOA is a collaboration between the W. M. Keck Observatory and Caltech/IPAC-NExScI archiving for all WMKO data [KOA, 2020], [Keck, 2020]. IRSA is NASA's data collection of all possible infrared data from their missions to improve planetary science with a 10% footprint of all refereed journal articles in astrophysics [IRSA, 2020].

The International Centre for Radio Astronomy Research (ICRAR), which focuses on data-intensive astronomy research in Australia is responsible for the support and maintenance of the Murchison Widefield Array (MWA) archive. The MWA is an international collaboration and one of the SKA precursor telescopes. The MWA archive is hosted at the Pawsey Supercomputing Centre storing more than 32 petabytes of data accessible to astronomers. The archive makes use of the Next Generation Archive System (NGAS), an archive handling and management system that provides the data ingest service for the archive [NGAS, 2020].

The Mikulski Archive for Space Telescopes (MAST) which is supported by NASA is a service

that provides access to data archives that store data from different missions focusing mainly on optical, ultraviolet, and near-infrared parts of the electromagnetic spectrum. The archive was created to help with the Data Archiving and distribution of data from the Hubble Space Telescope (HST). Over time it evolved and is now part of NASA's distributed Space Science Data Services (SSDS) [MAST, 2015].

The Canadian Astronomy Data Centre (CADC) provides a similar service to the MAST project hosting data from different telescopes with partner collaborations that offer compute, storage, and cloud services to store, manage and publish data service to the astronomy community [CADC, 2020]. The archive offers its services to astronomers in many different countries providing access to over 2 Petabytes of data. The European Southern Observatory (ESO) Archive offers the service to different European telescopes. All data produced by these telescopes are stored at the archive which provides both raw and pipeline-processed data [Adam et al., 2018].

Most archives adopted different access methods, software standards, and data presentation which changes the complexity of how to users interact with it and thus predicts the likelihood of making use of them. This is one of the reasons for the existence of the VO, to provide standards that data archives adhere to that provides a platform for interacting with data from different archives without having to know too much how every one of them operates [Fabio, 2009]. MAST offers several options for access to data [MASTAccess, 2020] via the search portal which is a web interface with several parameters to choose from. The API options offer users the option of using Astroquery, Webservices, or classic API. Astroquery is a python module that offers python users different query types, data downloads, and cataloging options for available data. Access to proprietary data can be handled via a token or via a login function [MASTAstroQuery, 2020]. The Web service and classic API options perform requests via a URL which can be programmed to string together a bunch of requests in a single script. The CADC offers a similar search interface to get access to the data via its WebUI API for classic with classic form search option and a direct data service option [CADCDoc, 2019]. An enhancement of the services offered by modern archives is the Archive as a Service (AaaS) offered by the CADC provides the option of setting up and archive premise against one's own data collection [AaaS, 2019]. ESO provides multiple links to the data portals of the different observatories that provide access to different data types [ESOArchive, 2010].

2.2 Virtual Observatory

The VO is a compilation of international standards that helps with astronomy data discovery, data access, data sharing, and analysis. This set of standards is set up by the International Virtual Observatory Alliance (IVOA). It creates a framework for processing

new and old data that is enabled by the implementation of new algorithms and a suite of modular tools for working with distributed data. The VO consists of tools and services which astronomers can use to do data discovery, plotting, data analyses, and access data archives [Hanisch et al., 2016]. It is described by [Young, 2004] as a suite of software applications that work together that allows users to uniformly find, access, and use resources from distributed product and service providers. The idea of the VO is to achieve transparency for astronomical data. The VO is a data-intensive astronomy research environment that makes use of the platform created by the enhancement of the technology to provide access to astronomy data [Wynholds et al., 2011],[Pasion, 2013],[Chenzhou et al., 2012]. All archives using a common database query language can be accessed through a uniform interface, and diverse data can be analysed by the same tools. This kind of infrastructure enables collaborations for informal distributed research teams to share data, workflows, and analysis results in a transparent virtual storage system [Hanisch, 2007].

2.3 Multi-wavelength Astronomy

Multi-Wavelength astronomy is the term used to describe the study of the universe through the use of the full electromagnetic spectrum [NASA, 2013]. Objects in our universe get discovered because of the radiation they emit which are picked up in different wavelength ranges. The wavelength ranges include radio, infrared, optical, X-ray, and gamma-ray. Figure 2.1 below explains the multi-wavelength ranges and what type of objects fall in the categories.

By creating tools, services, and techniques that allow for multi-wavelength astronomy, there is an added challenge of hosting different kinds of data sets in the same space. More and more research involves follow up research in other wavelengths to access different information about the same objects. The reason being that instruments can only operate in limited ranges of the electromagnetic spectrum. This creates extra pressure to the already existing issues of how data is stored, accessed, and processed.

The most important capability from an astronomer's perspective is that of locating an archive that contains data from a particular region of the sky, in a particular waveband, with a particular instrument. It is feasible to do so now, but the process is time-consuming. Multi-wavelength studies will help understand the distribution of stars in our galaxy, how and why stars and galaxies change with time [National Research Council, 2007].

Type Of Radiation	Radiated by Objects at this Temperature	Typical Sources
Gamma-rays	> 10^8 Kelvin (K)	accretion disks around black holes
X-rays	10^6-10^8 K	gas in clusters of galaxies; supernova remnants; stellar corona
Ultraviolet	10^4-10^6 K	supernova remnants; very hot stars
Visible	10^3-10^4 K	planets, stars, some satellites
Infrared	10-10^3 K	cool clouds of dust and gas; planets
Microwave	1-10 K	cool clouds of gas, including those around newly formed stars; the cosmic microwave background
Radio	< 1 K	radio emission produced by electrons moving in magnetic fields

Figure 2.1: Typical sources in temperature and radiation ranges.

2.4 Data Management

Astronomy data management processes improved with the involvement of instrument scientists in the design and commissioning of new instruments to help predict the expected data volumes that will be produced by establishing the potential throughput of each instrument. Having these potential figures creates the opportunity to plan for the data avalanche and prepare the environment to manage the expectation of high speed data transfer and increased storage capacity. These include the network infrastructure for data transfer, communication, and the data storage which forms part of the planning when building a new instrument to make better decisions for data management. Data management is the processing of the data from the observation, to the reduced data sets until it is accessible by the principal investigator or the astronomy community through an online archive. Data volumes produced by telescopes have been increasing over the last few decades in orders of Petabyte scales in today's terms. This data avalanche created new challenges in terms of storage, access, transfer, and management of these large volumes of data. The development of networks and connectivity to provide access to control both space-and ground-based telescopes provided the capabilities of hosting large volumes of data and making it accessible to users from anywhere in the world.

2.5 Summary

This chapter has discussed the background and evolution of astronomy data archives as a useful tool to conduct research. Astronomy data archives are useful to keep copies of observations at different wavelengths. The accessibility to data archives has improved over time which allows more research to be conducted. The literature shows that tools and services provided by the VO improved the ease of access to data. Archives are built and maintained by staff that specialise in big data storage, information security, database administration, and general information technology that provide platforms for access to research data to the broader astronomy community. The archive users have different approaches to their expectations and preferred interactions with archived data. The flexibility of an archive to provide multiple ways of accessing the data is an indicator of how popular the archive becomes amongst the users.

Chapter 3

Interoperable Tools and Services

The chapter introduces the concept of interoperability, which is the capacity of systems and services to have a clear common expectation of the contents, context, and meaning of astronomy data across different wavelengths and from different data archives. Interoperability would not be possible without proper metadata descriptions that contain a name, for identifying what is in the data, a description, which is value-added expertise of the data, a characteristic of the data, and a measure that gives reference to reference systems or units. The IVOA provides interoperable tools, services, standards, and various data formats given by existing data archives, which enable seamless and transparent interactions across archives and multi-wavelength datasets.

3.1 IVOA

The IVOA was formed in June 2002 to combine the efforts of different astronomical archives under one umbrella which helped to stimulate the growth of development and deployment of tools, services, and archives to deliver interoperability between data archives. It currently constitutes 20 VO programs from different countries with membership open to any national programs following the guidelines set out in the IVOA participation guideline [Hanisch et al., 2010]. The IVOA is modeled on the same concept of the World-Wide Web Consortium (W3C), the group that is responsible for developing open standards for the Word Wide Web (WWW), [Berriman, 2022]. The initial mission of the IVOA was to facilitate the international coordination and collaboration necessary for the development and deployment of the tools and systems. This coordination and collaboration were necessary to enable the international utilization of astronomical archives as an integrated and interoperating VO. The VO is committed to open data and open access. In theory, the VO adopted the Findable, Accessible, Interoperable, Reusable (FAIR) principles even before it was formalised and implemented.

3.1.1 IVOA Architecture

The IVOA architecture,[Dowler et al., 2021], was designed in 2010 and remains the standard document that provides the layout of the components of that describes just how VO operates. The document builds up the architecture with the basic layout called level 0, which indicates the basic layers of the architecture which include the user layer, resource layer, finding, sharing, getting, VO Core, and using sections as below illustrated in Figure 3.1. This is the most simplistic view of the architecture meant for a general public view and understanding of how the VO works [Arviset and Gaudet, 2010]. The Resource Layer is made up of data archives that manage astronomy data collected from multiple telescopes and provide data and computational services to the scientific community. The User Layer of the architecture is formed by the user's access to the Resource layer via the research team and computer systems at the different data archives. The VO acts as the cohesive agent to connect the Resource Layer to the User Layer similar to what the World Wide Web (WWW) provides. With the WWW a search happens from a device connected to the Internet to search for information located on a different computer in a transparent way. Data and services are shared via the VO framework which allows the users to search for information using VO provided resources [UEDIN et al., 2022].

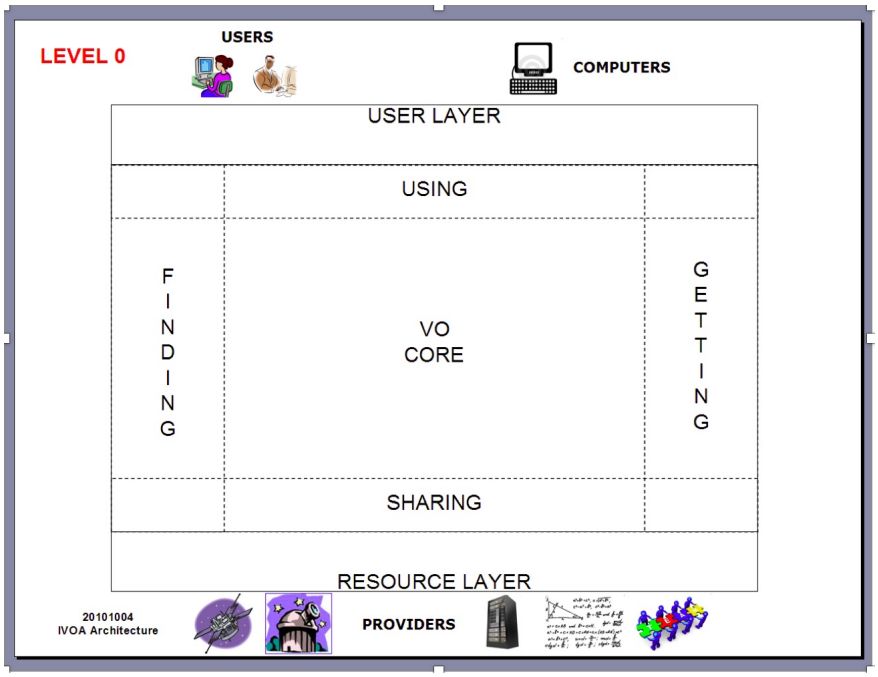

Figure 3.1: IVOA Architecture Level 0

3.1.2 IVOA Registry

The IVOA architecture is highly dependent on the registry. Registries facilitates the publication, resolution, and discovery of astronomical resources that are inherent in using it and designed to be accessed by software, often referred to more formally as a resource registry. A resource describes what data and computational facilities are available where, and how to use them. A registry is thus a repository of resources marketed by its owners to make users aware of them and know how to use them. In most cases, the registries, which are a very important part of the IVOA architecture are hidden from the normal user in the applications that they use [Nebot, 2022]. Data providers must register a service with a VO Registry and give the relevant information that characterizes the new service in order to promote its availability. Client programs can use the registration to find the best suited services for their application [Knowk and Tody, 2008].

Registries collect metadata and store it in a searchable database manner, similar to search engines. The registry, like other VO resources and services, is distributed throughout the network. To acquire access to data and metadata, Data Access Protocols are employed. These protocols provide a uniform way of accessing data and metadata from a variety of sources. A registry keeps track of resource metadata such as identifiers, titles, and descriptions. Additional metadata components can be presented in a catalog service resource to indicate how to use the service and its content [Greene and Plante, 2008a].

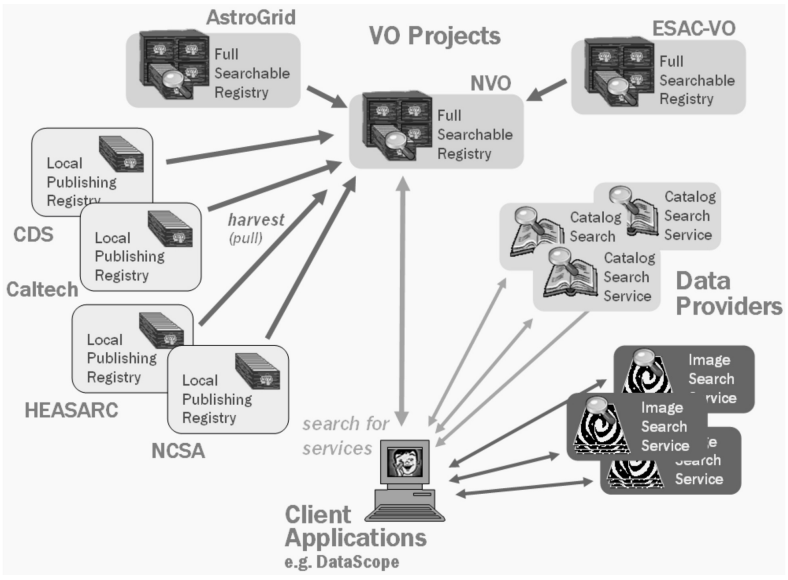

Figure 3.2: NVO Registry Model

The registry model shown in Figure 3.2 illustrates that publishers generate data and service descriptions and export them via publishing registries. Descriptions are frequently published through searchable registries. A fully searchable registry collects all known re-

source descriptions via a pull operation called as harvesting. A client program, such as DataScope, may now send a search for secondary search service descriptions, which it can then contact directly to locate images and catalog records [Greene and Plante, 2008b]. Registries are classified into published registries and searchable registries. Publishing registries are often run by data centers and are concerned with producing resource descriptions and disseminating them around the VO network. Searchable registries, on the other hand, allow users to look for resources that meet certain criteria [Arviset et al., 2006].

A registry resource description contains structured metadata that enumerates what the resource represents. These metadata are represented in XML with the help of a VOResource, which consists of a core metadata set and extensions. The core set of information represents principles that, in theory, apply to all resources. It is divided into three metadata categories: identification, curation, and content. The resource's identity metadata includes a title, a short name, and a globally unique identifier. The curation metadata describes who is in charge of the resource, whereas the content metadata describes what the resource includes. VOResource builds on a more general IVOA standard entitled "Resource Metadata for the Virtual Observatory", [Hanisch et al., 2007], which defines all of the core metadata concepts in the above categories plus core service metadata and gives them names. Applications use different formats such as XML, FITS, RDF, etc to exchange the metadata and VOResource explains how to render the metadata in XML.

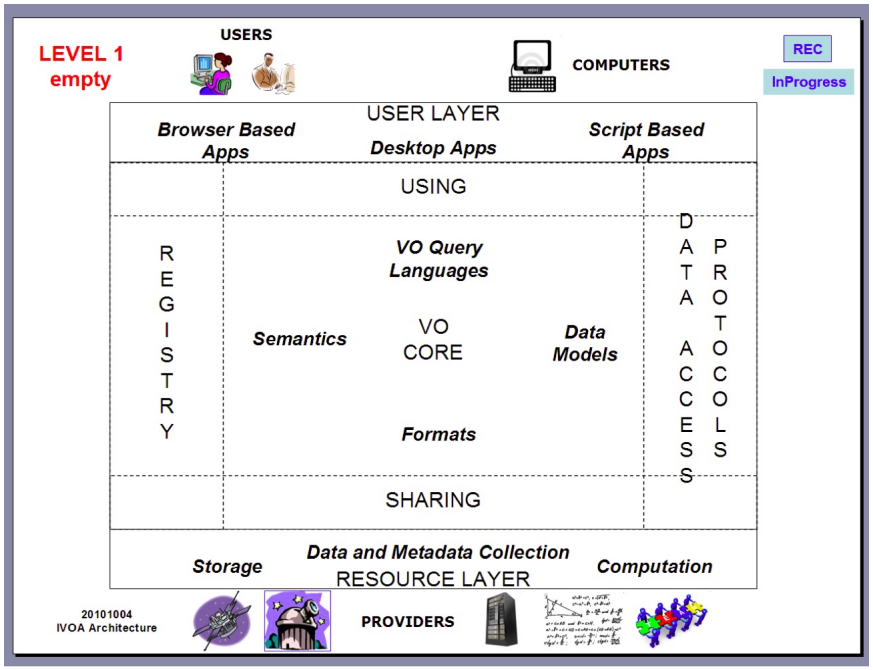

Figure 3.3: IVOA Architecture Level 1

3.1.3 IVOA User Layer

The User Layer illustrates how the user can interact with the IVOA by using the many applications and tools at their disposal. There are a number of applications that are IVOA compatible, these include web-based, and desktop applications. VOPlot and TOPCAT are plotting applications that take VOTables and provide a variety of display options which include statistics, filtering and cross-matching in table columns. Mirage, does plotting and image display and includes further tools for data classification and segmentation. Aladin is a freeware sky atlas that allows the user to see digitized photographs of any section of the sky and superimpose entries from astronomical catalogs [SAAO, 2022]. Data discovery tools like DataScope and SkyView make it easy for astronomers to find, compare, and retrieve data. DataScope is a web service that allows you to discover and explore data in the VO from archives and data centers all around the world [Iyer and DuttaDuwarah, 2018]. Skyview functions as a VO, generating photographs of any portion of the sky at wavelengths ranging from radio to gamma-ray. The VODesktop is a desktop program that allows you to interact with the VO. It can look through data resources, access external catalogs, and build workflows to automate processes.

VOPlatform is a tool that offers users with a location to store their regularly used VO tools and datasets, as well as other resources like as papers, web connections, and so on. The program is written in Java and operates on data in the VO Standard VOTable format. It can search through data resources, access external catalogs, and build workflows to automate processes. VOEventNet allows for quick observations of the changing night sky. VOEventNet collects streams of astronomical notifications and reports in a standard format so that both humans and robotic systems may respond with follow-up observations.

Open SkyQuery, is a robust interface for querying and cross-correlating objects from a variety of astronomical surveys. WESIX use the same questionnaires in conjunction with SExtractor to allow users to submit a picture, receive an item list, and have that list cross-matched with specified catalogs. Astronomers who come across photos with improperly calibrated astrometric solutions can correct them with WCSFixer. The National Virtual Observatory (NVO) Spectrum Services provides access and analysis tools for Sloan spectra as well as spectra from other surveys, as well as the ability for users to contribute their own spectra for comparison and analysis. The NED data and services available through the VO are also discussed here.

Montage is a mosaicking service that can stitch together very huge photos from the 2MASS and Sloan sky surveys. These mosaics have great photometric and astrometric precision, making them appropriate for scientific research. NESSSI is a general-purpose framework for handling massive computing operations on the TeraGrid. Users can combine repetitive operations via a script, or utilize VO services in custom applications, can make use of VOClient, VOAgent/VOLib, and IDL/VOLib. These provide access to the VO

data discovery and retrieval services for environments such as Java, Python, IDL, and IRAF.

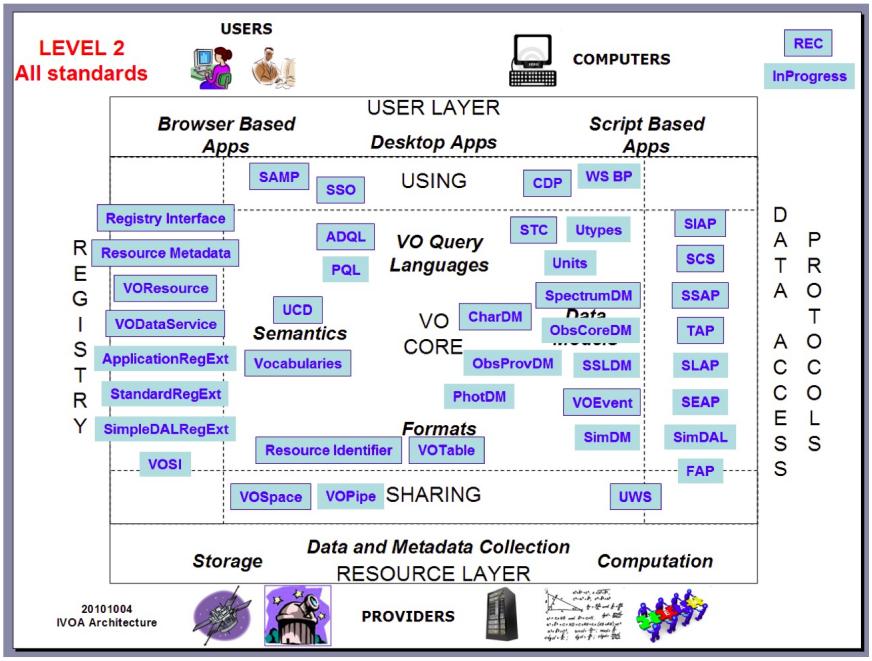

Figure 3.4: IVOA Architecture Level 2 - Standards

3.1.4 IVOA Data Access Layer Interface

The Data Access Layer Interface (DALI) defines how the Data Access Layer (DAL) service use IVOA standards as well as common Internet protocols and designs. The DAL standardizes the way applications can query data services, enable data discovery, describe resources, and define data access and retrieval methods via server-side processing [Galluzzi et al., 2019]. The individual data publishers can keep a local storage and access mechanism that matches their project-specific needs while using DAL protocols in the VO. The DAL manages several protocols that is used in a variety of services. Many of the issues encountered while building these protocols are common, such as query global syntax, validation and error query management, answer format, synchronous or asynchronous mode management, and availability or capability auto-description [Bonnarel et al., 2015].

TAP is a service protocol for acquiring access to common table data, such as astronomical catalogs and database tables. Information on the database tables, as well as real table data, is accessible. The service protocols have had to adapt to multidimensional data as new major observational studies provide open access and interoperability for cross-correlation of their data. DataLink offers several approaches and technologies for connecting online resources helpful for improved utilisation of existing or already identified datasets in this context.

Server-side processing for extracting information from datasets is driven by AccessData. Cutout, filtering, re-sampling or re-gridding, and combining several datasets are all possible uses. The DAL community has tackled multi-dimensional and multi-wavelength as well as keeping fundamental standards up to date. Time domain and radio astronomy data, are topics that remain a priority but attention was shifted to newer topics such observation location and object visibility information retrieval. According to [Molinaro and Dempsey, 2020] the DAL will work on the upgrades of existing standards based on the community feedback received about the new features that needs to be implemented. The standards that are under review for upgrades are, DataLink 1.1, DALI 1.2, ADQL 2.1 and ConeSearch 1.1. The TAP 1.1 standard have recently been updated and SIAP 2.0 and SODA 1.0 are waiting for community feedback before it can be reviewed for possible upgrades.

Large radio astronomy archives are generally used to store raw or calibrated visibility data. However, things are changing, and some facilities (for example, ALMA) now store and disseminate science-ready data. This will also be true for future radio telescopes, such as SKA, which will give calibrated and imaged data products. The primary purpose of these archives is, of course, to make their data discoverable and available to the astronomical community. Many archives now provide access to visibility data through project-specific online interfaces, allowing users to reprocess the data with fine-tuned reduction settings [Louys, 2020].

The IVOA has the mandate to provide interoperable data access to radio data. The IVOA founded a Radioastronomy Interest Group (RIG) for the specific task of addressing the shortcoming of interoperability of radio astronomy data archives. There is a need for the identification of metadata concepts for radio astronomy which provides different challenges because of the levels of diversity in its data types. More use cases are needed to improve data discovery access and visualization of radio data. The different data types include but is not limited to classic interferometry, interferometry with multiple beams and beamforming capabilities, very long baseline interferometry (VLBI), a single dish, time-domain, continuum, spectroscopy, and polarization properties. Apart from the diversity in the data types, the same can be expected from the different facilities and resources where observatories produce different data types [Lacy and Bonnarel, 2020]. There currently is a lack of proper schemas to include other radio data types like high frequency. The VO is striving to accommodate radio astronomy archives in a bit to become fully interoperable across multi-wavelengths.

Figure 3.5 illustrate the different components needed to be interoperable in the VO [Pasian and Molinaro, 2017]. For this research, I will focus on the VO Query Languages and Formats to describe how these components contribute to multi-wavelength astronomy in data archives.

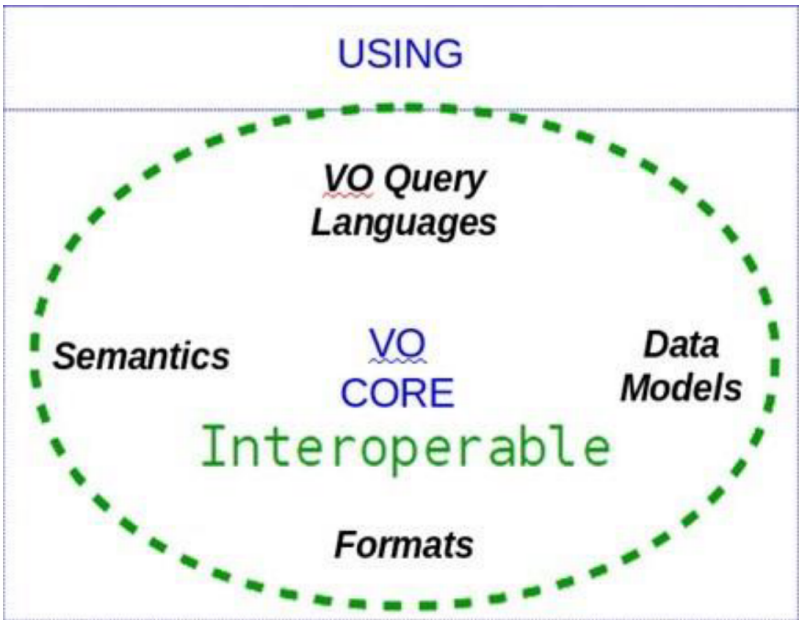

Figure 3.5: IVOA Architecture - Interoperability

Standards exist in the VO to facilitate the process of making data and service available across archives. The Astronomy Data Query Language (ADQL) and Table Access Protocol (TAP) service protocols provide access to general table data, catalogs, and general database tables. The Obscore presents the ability to pose a single scientific query to multiple archives simultaneously. The Cone Search protocol returns a list of astronomical sources from the catalog whose position lies within the cone, formatted as a Virtual Observatory Table (VOTable). The Simple Image Access Protocol (SIAP), provides capabilities for the discovery, description, access, and retrieval of multi-dimensional image datasets. High scientific value linked from the catalog will give access to the spectra, images, and light curves through Datalink's data discovery of metadata. Virtual Observatory Space (VOSpace) is the IVOA interface to distributed data storage. A celestial event and its measured attributes are represented by VOEvent. It incorporates existing IVOA standards and is intended to accommodate various data forms, including sections on provenance, scientific assumptions, and others. [Seaman et al., 2017]

3.2 Data Formats

The evolution of astronomy data formats is well documented and dates back thousands of years. Modern astronomical data formats are divided into four different categories namely processing, transfer, recording, and archiving. Early data formats for the digital computing focused on the transport and archiving of images.

3.2.1 Flexible Image Transport System

The Flexible Image Transport System (FITS) was defined as the standard data format for transporting images and archiving astronomy data sets [FITS Working Group, 2018]. FITS is a human and machine-readable data format that gives the user the power to self-document with key-value pairs of their choice. This format is extensible to include new features and future expansion on the original format which is key for the development of new instruments. FITS however is not adequate for the requirements of the types of science conducted with new instruments. For one the increased data volumes and different forms of storage media meant that the FITS logical and physical record length of 2880 bytes was not adequate to store data of that length, [Drakos, 1998]. More recently developed data formats build on the basis that FITS provides to improve software and instrument-specific processing and recording [Mink et al., 2014].

Further demands like the grouping of image data with its corresponding data quality and error arrays, the grouping of data from both detectors used, which could be solved by adding more keywords. However, the design limitation of the FITS standard meant that choices had to be made about changing the format to introduce these changes or make a software change to a create different outcome for the grouping permutations. Developed in the late 1970s, the FITS authors made many implementation choices that, while common at the time, are now seen to limit its utility for the needs of modern science. FITS is being pushed to its limit as astronomy evolves toward a more types of data products (data models) (data models) with richer and more complicated metadata. The issues with FITS are outlined in great detail by [Thomas et al., 2015]. One of the better qualities of FITS is its representation of metadata because of its readability and transparency. This and the fact that it has been a long-standing format and is well documented. Some of the drawbacks are the limitation of the key/value constraints with no room for organising of the content in the file. FITS does not have a hierarchical structure to do any grouping of data.

3.2.2 Hierarchical Data Format

The Hierarchical Data Format version 5 (HDF5), is a file format that can handle large data sets, complex data grouping and metadata association which describe the data set and specific details about the grouping of the data. It's an open-source file format which makes it widely applicable and supported by programming languages and software tools. HDF5 is heterogeneous and is capable of supporting different data types. In comparison to certain prior data formats (such as the FITS format), the HDF5 format offers far greater flexibility, including data chunking, external (i.e. distributed) object storage, and a data compression filter pipeline. Because it allows parallel I/O operations, the HDF5 format is especially well suited for quickly processing big data collections[Price et al., 2015]. HDF5 is a binary format that needs HDF5 toolset to inspect the content of each file. It is well documented

even though there is no documentation in the files which makes the reading of the files cumbersome because of the regular consultation of the lengthy documentation separate from the actual file. HDF5 is more suited for the role of a software API than a file format making it less attractive for an archival data format. The set of data types in HDF5 does not include a variable-length mapping datatype compared to the FITS data format that uses a Python dictionary style or Javascript object mapping. [Greenfield et al., 2015].

The size of files, as well as the size and number of objects in a file, are unrestricted in HDF5. The HDF5 format and library are expandable, and they're built to grow as the needs change. The functionality and data is portable across virtually all computing platforms. It has a simple but versatile data model that supports complex data relationships and dependencies through its grouping and linking mechanisms. It allows raw data to be stored outside, allowing raw data to be shared among HDF5 files and/or apps and frequently saving disk space. It supports a rich set of pre-defined data types as well as the creation of an unlimited variety of complex user-defined data types. Datatype definitions include information such as byte order, size, and floating-point representation, to fully describe how the data is stored, insuring portability to other platforms. HDF5 through its virtual file layer, offers extremely flexible storage and data transfer capabilities by chunking and compressing data into smaller sizes [Bonnarel et al., 2008]. By dividing the data into equally sized blocks or chunks, it is stored in separate smaller blocks which improves the storage efficiency and transmission speeds [Breitenfeld et al., 2020].

3.2.3 Advanced Science Data Format

The Advanced Science Data Format (ASDF), is an Ain't Markup Language (YAML) based text-based data format, which like FITS, is human-readable and could be viewed with simple text editors [Greenfield et al., 2015]. It contains a hierarchical metadata structure composed of fundamental dynamic data types such as texts, integers, lists, and maps. It features editable human-readable metadata that may be modified right in the file. The structure of the data can be automatically validated using a schema. ASDF is primarily intended as a standard for exchanging products from equipment to scientists or between scientists [Greenfield et al., 2015]. While it is relatively efficient to use and transport, it may not be ideal for direct usage on huge data sets in distributed and high-performance computing systems. ASDF keeps a readable header using a standard format and structural relationships explicit through key/value and list structures with key names longer than eight characters. Values can be of unlimited size and complex structures themselves. Support for schemas and corresponding validation tools allows much more flexible World Coordinate System (WCS) models and definitions of models in general.

ASDF is a generic scientific format; not restricted to astronomy. It offers better streaming support; is extensible and provides local tag definitions that can be constructed and

used with local libraries, and offers support both text and binary arrays and tables and [Mink et al., 2020]. ASDF's representation of the WCS is an improvement from the way that it's represented in FITS. It is more sensible and flexible with the organisation of the metadata and data. A YAML metadata header and optional binary data blocks comprise the file. ASDF is currently in use by astronomy projects at NASA's James Webb Space Telescope (JWST) and the Rubin Observatory mainly for its WCS support. The Gemini DRAGONS programme is using the ASDF format for data reduction software. It will also be used as the native data format for the Nancy Grace Roman Space Telescope [Greenfield et al., 2015].

3.2.4 Parquet

The Parquet data format, is a columnar format, originating from the Hadoop big data ecosystem in 2013. It is compressed along with the columns and handles variable-length columns like light curves. Parquet supports object storage and also works well with the popular Pandas package with good support in languages besides Python. A couple of applications are adapting the Parquet data format. Vera C. Rubin Observatory catalogs will be made available in Parquet format. The paper titled, Simulations from the Large Synoptic Survey Telescope(LSST) Dark Energy Science Parquet, suggest that LSST is considering Parquet as their preferred data format. The Infrared Science Archive (IRSA) is exploring Parquet for bulk distribution of large catalogs, moving from HDF5 to Parquet format [Shupe et al., 2020]. Parquet is an industry standard with an active developer community and well-maintained and supported. Its built-in compression provides the ability for efficient long-term data storage [Databricks, 2021].Because the data is kept in columns, it may be highly compressed and segregated (compression algorithms operate better with data with low entropy of information, which is typically included in columns).

According to the format's creators, this storage format is suited for handling Big Data challenges. A Parquet file's column metadata is placed at the end of the file, allowing for quick, single-pass writing. Parquet is optimized for the Write Once Read Many paradigm (WORM). It writes slowly but reads rapidly, especially when only a subset of columns is used. When reading large amounts of data, Parquet is an excellent choice. Only the necessary columns will be retrieved/read, reducing disk I/O. This is known as projection pushdown. Because the scheme goes with the data, the data self-describes. Because Parquet is merely files, it's simple to deal with, transfer, backup, and replicate them. Parquet provides very good compression up to 75 percent when using even compression formats like snappy. When compared to other file formats, this format is the quickest for read-heavy procedures. Parquet is ideally suited for data storage solutions that need aggregating on a specific column over a large collection of data. It also supports predicate pushdown, which reduces the additional cost of moving data from storage to the processing engine for filtering [Luminousmen, 2022].

3.3 Summary

This chapter gives an introduction into the tools, systems, services and standards that could be used to promote interoperability between data archives. The IVOA architecture is discussed and explained to show the evolution of data interoperability across wavelengths, which is an enhancement for astronomy research across wavelengths. It provides astronomers with access to all archived astronomical data from via a single interface. The tools and service provided by the IVOA allows data to be located, compared and retrieved, combining multi-spectral and multiple instrument data from different archives. Data archives are locally hosted by each institution, but data can be shared by adhering to extensible metadata standards and interchange protocols. Distributed virtual observatory operations, are supported by a core set of standards, interoperability, and management services will be required.

Chapter 4

Astronomy Data Management Facilities

This chapter provides an overview of how existing observatories and astronomy research facilities manage various astronomy data archives. The facilities I will concentrate on are CANFAR, ICRAR, NRAO, ASTRON, and NOIRLab, which handle survey data management from the point of observation to the production of raw data and finally the presentation of science output from the observations.

4.1 Canadian Advanced Network for Astronomical Research

Many astronomy research groups rely on the Canadian Advanced Network for Astronomical Research (CANFAR) for data-intensive storage, access, and processing [Berriman, 2011]. CANFAR is an environment that consists of the Canadian national research network (CANARIE), the cloud processing and storage resources provided by Compute Canada, currently known as Digital Research Alliance of Canada (the Alliance), [Bertocco, 2017] which provides an astronomy-specific cloud computing platform [Ball, 2010]. CANFAR provides a Virtual Cluster, over which the user has complete control, and which uses the Cloud Computing provided by Compute Canada. Hence, rather than build a new infrastructure for a project such as a sky survey, an individual or collaboration may utilize CANFAR. Figure 4.1, depicts the CANFAR architecture [Kavelaars, 2016]. The overall architecture is comprised of the CANFAR infrastructure, which hosts cloud processing and storage resources, an astronomy data center, and tools and services that adhere to IVOA standards. At CANFAR the implemented IVOA standards and recommendations are, VOSpace, Single-Sign-On(SSO), Credential Delegation Protocol (CDP), Universal Worker Service Pattern (UWS) and IVOA Support Interfaces (VOSI). VOSpace is the IVOA interface to distributed storage. SSO is the IVOA profile that describes approved client-server authentication mechanisms. CDP

allows a client program to delegate a user's credentials to a service such that that service may make requests of other services in the name of that user. UWS, defines how to manage asynchronous execution of jobs on a service. VOSI, describes the minimum interface that a web service requires to participate in the IVOA, i.e. a set of common basic functions that all these services should provide in the form of a standard support interface in order to support the effective management of the VO [Bertoccoa et al., 2018]. Users utilize CANFAR to create and configure virtual machines (VMs) that are saved in VOSpace. Instances of the VM images are launched in a batch processing environment. TAP's data access services are utilized to locate data utilizing ObsCore's data discovery services, which aid in the overall description of observations and archive data products. Datalink is a service that locates all files that are related to the selected science file and displays a table with all of those files' access points. VOSpace is used to store the output of the CANFAR processing system as well as to share files between members in a collaboration. VOSpace files are replicated in four physical locations to ensure that they are safe against disk failure [CANFAR, 2021].

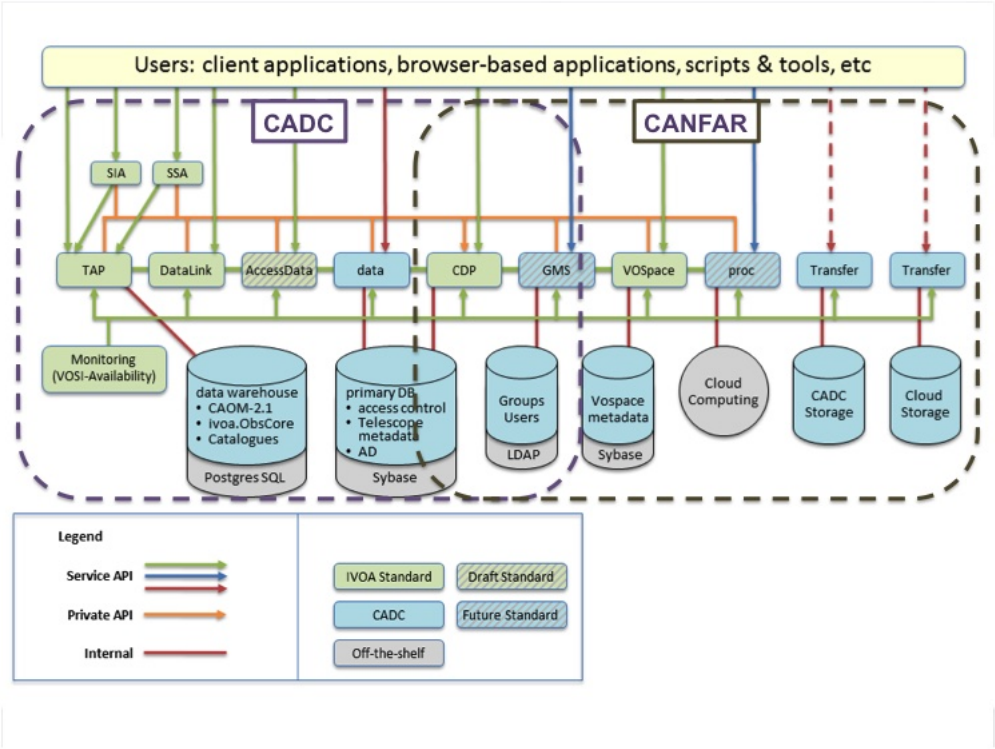

Figure 4.1: CANFAR/CADC - Architecture

CANFAR uses a cloud-based framework to provide its users with access to large resources for both storage and processing. A twin framework has been deployed at OATs-INAF, with

the same technical requirements and use cases and based on the same software libraries. A Java-written RESTful service manages the access control service for user access. It makes use of the Restful API and the Oracle JAAS (Java Authentication and Authorization Service). This service allows for user registration, authentication, and management of user groups. Multiple identities per user are supported (currently, X.509 certificate, user/password, and cookies are implemented, but they are easily extensible). An LDAP server stores information about users, groups, and group memberships. Users have access to a resource (for example, a service or proprietary data) if they are a member of the group(s) that protects that resource. The resource owner, who is also in charge of assigning group membership to other users, grants access rights to a specific resource [Bertocco et al., 2017].

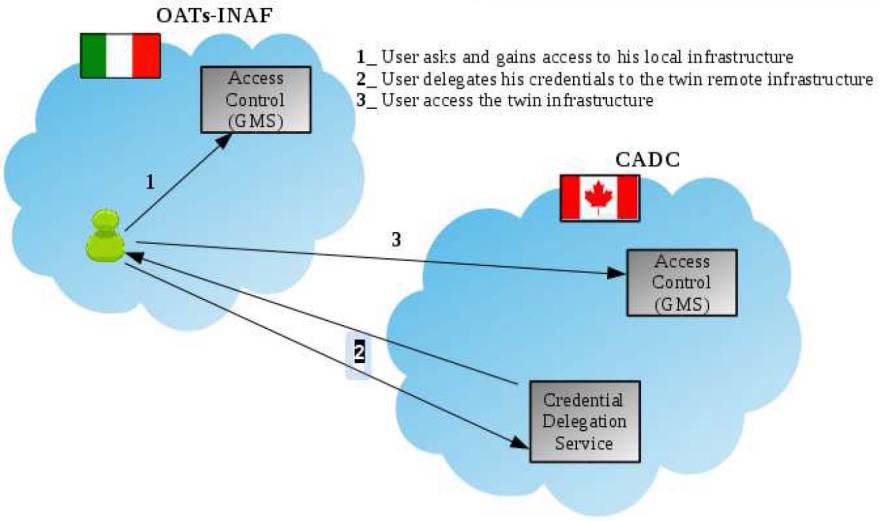

Figure 4.2: Access Control interoperability

CANFAR and the Italian National Institute for Astrophysics (INAF)-Osservatorio Astronomico di Trieste (OATs) created a federated international cloud facility with storage and computation capacity that supports astronomical-specific data sharing and calculation technologies as shown in Figure 4.2 above. This project was in part funded by the EGI-Engage H2020 European Project [EGI, 2017].

4.2 Australian Data Archives

The AllSky Virtual Observatory (ASVO) is an Australian organisation consisting of 5 nodes that provide researchers with to access data across multiple datasets, from all types of astronomical facilities in Australia. The Data Central and SkyMapper institutes, hosted by the Macquarie University and the Australian National University respectively are hosting the

optical astronomy data. The Murchison Wide Field Array (MWA), hosted at the Curtin University, and the Australian SKA Pathfinder (ASKAP) hosted by Commonwealth Scientific and Industrial Research Organisation (CSIRO), named CASDA, is hosting the radio astronomy data. The last node is the Theoretical Astrophysical Observatory (TAO), hosted by Swinburne University, which is responsible for all theoretical data [O'Toole and Sealy, 2019]. ASVO's purpose is to use a cloud-based data storage system to federate and distribute astronomical data to the general public. The ASVO is a significant infrastructural project that connects observational data with theoretical capabilities. It creates a platform for astronomers to access and use the exponential expansion in astronomical data volume expected over the next decade.

The MWA node sends out pre-processed data that hasn't been calibrated. Curtin University manages the telescope and ASVO node, and the data is stored in the Pawsey Supercomputing Centre. The MWA is a low-frequency radio telescope that operates between 80 and 300 MHz and is one of two SKA precursors in Australia. This node splits down data into smaller chunks so that astronomers who aren't directly participating in the study can access it. The MWA provides a TAP service that can be used to receive MWA observation metadata by VO / TAP compliant software such as TOPCAT. The IVOA "ObsCore" template is supported by the TAP service, as well as MWA-specific schemas providing richer observation metadata [MWA, 2021]. The SkyMapper node serves to process and calibrate multi-epoch, multi-band images and photometry from the SkyMapper Southern Sky Survey. The data are taken with a specially built 1.3m telescope located at Siding Spring Observatory and operated by the Australian National University, [ANU, 2021]. The Australian Astronomical Observatory created Data Central to provide raw Anglo-Australian Telescope (AAT) data as well as survey data products. It contains almost 45 years of AAT data, as well as survey data releases from the Galactic Archaeology with HERMES (GALAH), Sydney-AAO Multi-object Integral-field spectrograph (SAMI), and Galaxy And Mass Assembly (GAMA) surveys, among others. Data Central features a web interface and IVOA services, and a REST API is on the way.

The TAO stores queryable data from a variety of cosmological dark matter numerical simulations and galaxy formation models in a database designed for quick access. The astronomy community can access TAO via the "cloud" via a browser from anywhere in the world. It has a simple and user-friendly online interface. TAO eliminates the need for users to create SQL queries by offering a bespoke point-and-click interface for selecting galaxies and their attributes, which automatically generates query code in the background. Query results can be routed through various "modules" before being delivered to the Swinburne OzSTAR supercomputer for further processing. TAO provides web access to cloud-based fictitious extragalactic survey data created using powerful semi-analytic galaxy formation models coupled with massive N-body cosmological simulations. TAO is designed to be versatile, allowing numerous simulations and galaxy formation models to be saved and accessed

using the same data format from a single place. To reach as many astronomers as possible, the user interface design goal is to provide a simple portal that makes science modules easy to use. No programming skills (SQL or otherwise) are required to use any part of TAO, making it more accessible to astronomers, whether they are observers or theorists. TAO's capacity to post-process hosted data for various scientific applications is a key feature. This is accomplished using many science modules that can be coupled in user-defined configurations based on the astronomer's desired needs and module capability. TAO may also be easily expanded with new capabilities in the future because of its modular design [TAP, 2019].

The International Centre for Radio Astronomy Research (ICRAR) is a collaborative environment in which scientists and engineers can engage and collaborate with industry to produce studies, prototypes, and systems that contribute to the overall scientific success of the SKA, MWA, and ASKAP [ICRAR, 2021]. ICRAR develops software and systems, as well as expert systems (machine learning), and provides operating system support, including the in-house ICRAR compute lab. ICRAR plays a key role in the international Square Kilometre Array (SKA) project, the world's biggest ground-based telescope array. Attracting some of the world's leading researchers in radio astronomy, who will also contribute to national and international scientific and technical programs for the SKA, ASKAP, and the MWA. Creating a collaborative environment for scientists and engineers to engage and work with industry to produce studies, prototypes and systems linked to the overall scientific success of the SKA, the MWA, and ASKAP.

ICRAR enhances Australia's position in the international SKA program by contributing to the development process for the SKA in scientific, technological, and operational areas. Promoting scientific, technical, commercial, and educational opportunities through public outreach, educational material, training students, and collaborative developments with national and international educational organizations. Establishing and maintaining a pool of emerging and top-level scientists and technologists in the disciplines related to radio astronomy through appointments and training. Making world-class contributions to SKA capability, concerning developments in the areas of Data-Intensive Science and support for the Murchison Radio-astronomy Observatory.

ASKAP data will be delivered to Perth's Pawsey Supercomputing Centre at a rate of approximately 2.5 GB/s, equating to 75 Petabytes per year. The total archive data volume is expected to reach 5 PB per year [PAWSEY, 2021]. ICRAR's Data Intensive Astronomy program is to develop the data management and processing technologies required for the SKA. CASDA serves as the central repository for storing, managing, and sharing fully calibrated and science-ready data products [ICRARDIA, 2019]. It is a data management system that archives science data from the ASKAP processing pipeline and makes it discoverable and accessible.

The total volume of archive data is expected to reach 5 PB per year. As part of its multi-tiered system, CASDA makes use of ICRAR's Next Generation Archive System, commonly known as NGAS, to meet data storage and retrieval needs. NGAS is being used by many of the world's major astronomical observatories. Today, ICRAR's version of NGAS has been optimized and finetuned to deal with far higher data rates and volumes than initially possible. It has also been adapted to run in both supercomputing and cloud-based environments and to support the integrated usage of cloud resources for scientific analysis. The NGAS infrastructure is currently used to control many Petabytes of data stored in hundreds of millions of individual files, distributed and mirrored across four continents while providing transparent access for end-users. NGAS has an integration with the hierarchical storage layer and supports the large files produced by the ASKAP pipelines [ICRARDIA, 2019]. CASDA is designed to handle an ingest rate of 16 TB of data per day. The application supports long-term data tape and disk storage at Pawsey, the Australian Supercomputing Centre.

CASDA offers a range of data access services to search for data in different stages. CASDA's observation search allows astronomers to search for and access processed data from ASKAP via a simple web form. The Skymap service allows astronomers to interactively and visually search for an object in the sky with an Aladin-lite interface. CASDA VO services can be accessed via python scripts to automate access to ASKAP data. The TAP service can be used to search the metadata for specific images and cubes, or radio sources in the CASDA catalogs. To provide these, CASDA made use of the open-source Astronomical Data Query Language (ADQL) and Universal Worker Service (UWS) libraries obtained from Strasbourg astronomical Data Center (CDS). CASDA built a continuous delivery pipeline that includes automated testing of every change against the VO standards using the TAPLint and VOTLint tools from the STILTS package. The CASDA VO library is available for use by other data centers. These web services are provided through the CSIRO Data Access Portal (DAP). The DAP is an enterprise-wide system that archives and provides access to data across many areas of CSIRO research [Chapman et al., 2015].

4.3 National Radio Astronomy Observatory

The NRAO operates three radio telescopes for scientific research. The International Atacama Large Millimeter/Submillimeter Array (ALMA), Very Large Array (VLA), Green Bank Telescope (GBT), and Very Long Baseline Array (VLBA) [Kellermann et al., 2021]. The NRAO handles all aspects of telescope management, including the Proposal Submission and Management System, the NRAO Data Vault, data processing, and all interactive services [Radziwil, 2011]. ALMA, one of the world's largest radio telescopes, collaborates with the European Southern Observatory (ESO), the National Radio Astronomy Observatory (NRAO), and the National Astronomical Observatory of Japan (NAOJ), [ALMA, 2021].

These three observatories run data archives and serve as ALMA Research Centres (ARC), where data from the ALMA headquarters, including observation and calibration data, is copied. The ALMA could produce 80 Gigabytes per day at full capacity. The VLA, VLBA, and GBT data are still stored in the legacy data archive and can be accessed using the legacy Archive Access Tool (AAT). The NRAO's data policy enforces the proprietary period, during which data is only available to the observations' principal investigators[Chandler, 2014].

The NRAO operates the North American ALMA Science Center (NAASC) and the ALMA archive in Charlottesville, Virginia called the NAASC cluster (cvpost). It also operates the New Mexico Array Science Center (NMASC) and the VLA/VLBA archive in Socorro, New Mexico known as NMASC cluster (nmpost) [NRAO, 2021b]. Lustre, a parallel distributed filesystem used in large-scale computing facilities, is used by both clusters. Lustre enables NRAO desktops, public machines, and clusters at a specific site to share a large file space, eliminating the need for data to be copied between systems for processing. They are primarily intended to improve performance by aggregating individual disk throughput across a large number of disks. As a result, the resulting storage volume is typically larger than that of desktop storage [NRAO, 2021c]. The NAASC Lustre filesystem is comprised of four storage servers, each with four RAID arrays (16 total arrays) of 64TB capacity. The total storage capacity is 1.1PB. Individual nodes can read and write to the Lustre filesystem at rates in excess of 1GByte/sec, and the entire filesystem can handle aggregate I/O of around 10GB/s. The NMASC Lustre filesystem is made up of ten storage servers, each of which contains four 44TB RAID arrays (for a total of 40 arrays). There is a total storage capacity of 1.8PB. Individual cluster nodes can read and write to the Lustre filesystem at speeds exceeding 1GByte/sec. The filesystem can handle approximately 15GB/s aggregate I/O [NRAO, 2021a]. NRAO supports different data retrieval methods which include, SFTP, SCP, LFTP, RSYNC, and Browser Access which are all done with the NRAO user credentials provided by the access control functionality. The NRAO computing facility also now makes use of Globus, a service that facilitates large-scale data transfer between computing facilities [La Plante et al., 2021]. NRAO has plans to support the XSEDE's Globus Connect platform which is a virtual cyberinfrastructure that allows scientists to interactively share computing resources, data, and expertise.

The NRAO is working on a Science Ready Data Products (SRDP) project to help its users deal with massive amounts of data and the complex and intense data processing requirements that come with it. Users can picture for themselves using the SRDP's calibrated visibility data. The pipelines are automated to provide a set of standard picture products as well as a science platform where customers may specify processing specifics and fine-tune the outputs to meet their needs, according to [Lacy et al., 2020]. The NRAO will also replace the Archive Access Tool (AAT) with an easy-to-use, modern user interface that can search for and return data from all NRAO telescopes, in addition to the SRDP. Virtual Observatory (VO) methods will make data from both raw and SRDP generated products available,

allowing for greater support of multi-wavelength science and better integration of radio or sub-millimeter data into the global astronomy data ecosystem. The NRAO wants to link its archival data sets to the published literature so that researchers can quickly access the research publications that present and analyze the findings [Kern et al., 2019].

The new NRAO archive will have the capacity to detect whether metadata is inaccurate manually or automatically, as well as implement a reingestion recovery approach to help with "self-healing" [Dashofy et al., 2002]. Reingestion is the process of recovering data from an archive, extracting metadata (as in normal ingestion), computing the differences between the stored and newly-extracted metadata, and delivering those differences (the "diff") [Gosh et al., 2006]. All metadata starts out in the considered good form and can be flagged by a user or by an automated mechanism to become questionable. It becomes a higher priority candidate for reingestion if it has been flagged as questionable. There are no bug-proof means of ingesting data and keeping metadata in an archive that contains data from so many instruments over such a lengthy period of time. The self-healing archive is a design that will aid in the continual improvement of curation, not just of current data but also of previous data [Witz et al., 2019].

4.4 ASTRON

The Netherlands Institute for Radio Astronomy (ASTRON) is a research institute dedicated to radio astronomy. The LOw Frequency ARray (LOFAR) is an instrument that measures the earliest phases of the cosmos, transient flashes, whirling neutron stars, and colliding black holes. It was planned, developed, and is operated by ASTRON. This is accomplished by examining the signals emitted by the Universe at radio wavelengths. By offering access to tools, instrumental data, and science-ready data products, the ASTRON Science Data Centre will enable fundamental scientific discoveries. Astrophysicists can use data processing and analysis services to gain experience working with the extraordinarily vast and complicated data products created by present and future radio astronomy instrumentation. This data center serves as a gateway to data archives as well as international high-performance computer systems, allowing researchers to conduct data analysis that will lead to breakthrough discoveries. [Astron, 2021].

Open and FAIR data sharing is essential to achieve the greatest quality scientific discoveries by facilitating partnerships and peer-reviewing, as well as to optimize scientific output from research by lowering access to data and data-analysis abilities. These ideas have recently received a lot of attention in the scientific community. They have long been recognized and used in the field of astronomy due to the need to investigate astronomical phenomena by combining data from a wide range of astronomical devices all across the world over long periods of time. This has resulted in the development and deployment of open standards, as well as open data and software repositories. As a result, radio astronomy is

now being used to establish open data services in other academic domains. [Swinbank, 2021].

The scale at which data is generated and evaluated is a distinctive feature of radio astronomy study. New possibilities for cross-domain and global data and expertise exchange are emerging as network and computational technologies advance. Technological developments have resulted in an exponential growth in the amount of data produced by radio astronomy equipment, as well as the emergence of a new generation of globally distributed radio astronomical observatories such as LOFAR and SKA. By creating worldwide links with both public and private institutions, ASTRON seeks to solve these difficulties. ASTRON worked with partners in the ICT infrastructure to develop a distributed data archive with astronomical content and scope.

ASTRON, in conjunction with the European Open Science Cloud (EOSC), provides international researchers with access to a scalable computational infrastructure that supports data storage and processing services. ASTRON is working with EOSC to build and integrate services that allow scientists to create and use scientific data products. Data archive access services, scientific processing workflow services, and research data repositories are the building elements for a European-scale science data center. By providing data mining capabilities, VO functionality, and processing pipelines operating on integrated computing clusters, a user portal is being constructed to allow low-threshold access to the underlying capabilities [ASTRON Editorial Team, 2018].

Given the scale of data from instruments like LOFAR and SKA and the complexity of the processing required to generate science-ready data products, many researchers will benefit from having access to high-performing data infrastructure and high-throughput processing capabilities without having to organize resources or set up complex software installations. To make software deployment simple and portable, ASTRON is developing container-based technologies. Data analysis pipeline installations, as well as application images, are distributed over connected infrastructure. For the most data-intensive workflows, a user workspace is being constructed to temporarily store data, for example, to free up computing resources for the next steps in the processing workflow or to evaluate quality before ingesting data into an archive or a science data repository.

4.5 National Optical-Infrared Astronomy Research Laboratory

The National Optical-Infrared Astronomy Research Laboratory (NOIRLab), is part of the National Science Foundation, which operates all ground-based optical and infrared astronomy research[NOIRLab, 2019]. The NOIRLab currently is operating five research programs,

Cerro Tololo Inter-American Observatory (CTIO); the Community Science and Data Center (CSDC); Gemini Observatory; Kitt Peak National Observatory (KPNO); and the Vera C. Rubin Observatory. The Vera C. Rubin Observatory's current focus is to deliver the Large Synoptic Survey Telescope (LSST). LSST will generate nearly 20 TB of raw data per night in optical astronomy [Juric, 2015]. LSST implemented a Data Management (DM) system to handle raw data reductions to scientifically useful catalogs and images. The DM's primary functions are to process the incoming data stream within 60 seconds of observation, archive raw images, generate alerts to new sources or sources whose properties have changed significantly, and update the relevant catalogs [Juric et al., 2021]. It provides free online access to its science platform which provides a collection of Jupyter computational notebooks, web portals and application programming interfaces (APIs) for data analysis, browsing and retrieval [Guy, 2018], [Juric, 2015]. Using a web browser, the LSST's users will be able to write and run code in the Python programming language to analyse the entire LSST data set remotely on servers hosted at the National Center for Supercomputing Applications in Urbana, Illinois, rather than downloading the data to their own computer. All data releases will be archived for the duration of the LSST archive's operational life, with the two most recent releases available in a queryable database. By providing appropriate software, the DM system facilitates the creation of added-value data products, application programming interfaces (APIs), and computing infrastructure at the LSST data access centers [Juric et al., 2021].

All LSST data will be made available to the community through interfaces that utilize community-accepted standards to the maximum possible extent. The data management system is architected in three layers: an infrastructure layer consisting of the computing, storage, and networking hardware and system software; a middleware layer, which handles distributed processing, data access, the user interface, and system operations services; and an applications layer, which includes the data pipelines and products and the science data archives. The applications layer is organized around the data products being produced. The Science Platform of the LSST has three user-facing aspects. The portal is used for new data releases and to send out new alerts. JupyterLab provides the environments with notebooks, user databases, and access to user files. Web APIs together with IVOA standards provide software tools and end-user applications to perform user computing tasks [Dubois-Felsmann et al., 2019].

4.6 Summary

This chapter gives an overview of the respective astronomy data management facilities. It gives and indication of which surveys and research they are focused on. This illustrates the diverse methods applied and the organisation of resources and collaboration between different facilities to reach common goals. This will broaden the scope and availability of

data throughout the research community. Using data from various archives makes multi-wavelength investigations a significant topic in astronomy. The VO offers the tools and services to make that enable the multi-wavelength studies.

Chapter 5

User Experiences and Expectations

In this chapter, I discuss the experience of using various data archives from a user and archive support staff point of view. I conducted a survey by asking archive users and archive staff members to complete questionnaires that capture their experience with archives from two different perspectives. In some instances both questionnaires were completed by the same person from both perspectives. A data archive is a critical part of the research process because it enables reproducibility of research. If data is managed effectively, its veracity will increase, potentially leading to an improvement in data quality, increased trust in the data, and an increased number of citations [ANDS and NCRIS, 2017]. Since astronomy is an observational science, observations are recorded and archived. Many phenomena are time-dependent, making archiving even more critical. The user questionnaire focuses on the reasons for choosing the particular archive(s) and the tools and services these archives offer. The survey had 19 responses and provide insight into the state of astronomy data archives. The form of the survey, the representativeness of the respondents, and the specific questions asked will all have an impact on the significance and thoroughness of the findings reached from it. The following sections provide the questions and feedback from users and archive support staff.

5.1 User Feedback

I discuss what the archive provides to the users who do multi-wavelength research and how it is applied. The feedback addresses issues such as the background research required to use a given archive, the tools and services supplied by that archive, and how the user utilizes the features offered by different archives.

5.1.1 Data Archives

I received responses from data archive users in different locations. The first question asks which archives are used and why. This gives an indication of which ones are out there and what the reason is that people used them. With the second question, I am trying to find out what tools and services the particular archives provide to assist with their research. Some users give explanations of how they utilise these tools and services. Question 3 enquires about the technique of multi-wavelength research and if the archive offers such features. In the last question, I ask whether the tools and services offered by the archive are supporting IVOA standards to enable interoperability. From the opinions given by some of the astronomers it was clear that they are looking for specific data in a specific archive to assist them. Others stated that there is not one particular archive that satisfies their needs so they list a combination of the archives, some even point out the exact specifics of the required data, tools, and services.

For the first question, most users listed well-known archives like Set of Identifications, Measurements and Bibliography for Astronomical Data (SIMBAD), Vizier, Hubble Legacy Archive (HLA), NASA/IPAC Extragalactic Database (NED), European Southern Observatory(ESO), and Centre des Données Astronomiques de Strasbourg(CDS). The reason for the use of these archives is because of the types of data that they are interested in. Some are interested in large datasets for machine learning experiments. Some users use certain archives because of the ease of use and functionality they provide. Of particular note was the mention of using multiple archives to complement the data they already have, like spectra, if an image is already available or images in a particular field in different wavelengths or to find new data for future projects. One response was to use Vizier and SIMBAD to start with but use the IVOA registry to discover data resources.

5.1.2 Tools and Services

The second question which focused on the tools and services that users look for in an archive seems like some are interested in the easy access, search, and data retrieval functions of certain archives coupled with the scientific relevance of the data to the user's research. Some are interested in more advanced search options, like cone search, multiple-object search, and name resolution searches. Others would like to have tools like image previews, basic information about the data, interactive nature for finding data, well-documented APIs for scripting, and having the ability to use SQL-based services for tabular data.

5.1.3 Multi-wavelength Support

For the third question, most responses were yes to multi-wavelength support from archives to complement their current wavelength range. Some of the respondents want the archive to provide information upfront about the availability of data in other wavelengths. One remark was to possibly include the wavelength as one of the optional search parameters on

the search or query interface. To eliminate multiple searches it might be useful if the archive cloud provide a discovery tool that would return data from multiple wavelengths even if not requested, but allow the user to decide whether or not they need it. Combining data from the different archives to get the multi-wavelength representation normally seems like a manual process to retrieve the data.

5.1.4 Standard Tools

In the last question, a few users mentioned IVOA standard tools and applications. MAST uses the European Space Agency's (ESA) discovery portal, called ESASky, which provides full access to the entire sky. It is an application that allows users to visualise cosmic objects near and far across the electromagnetic spectrum via web browsers on any device with an internet connection. Others are using Python libraries like PyVO, Astroquery, and Astropy's VOTable reader and also other IVOA based IVOA based tools and applications like TOPCAT, Aladin, VOTable, and Hierarchical Progressive Surveys (HiPS), which describe astronomical images and provide a solution for managing large amounts of data.

5.2 Archive Support Staff Feedback

The interaction with support staff from various archives yielded interesting remarks in the support staff questionnaire. The aim of the staff questionnaire is to get information about the features and what type of support they provide to the astronomy community. The questions were set up to get answers in fields such as metadata, data formats, storage technologies, authentication protocols, user capacity building, data privacy policies, interoperability, IVOA services, support challenges, and user feedback.

5.2.1 Most Important Tools and Services

For the question about the what they deem as the most important tools and services for an astronomy data archive should provide, the respondents had the following remarks. An archive should provide tools and services that makes it easy to find and access data thus providing an example of the FAIR principles. A good practice would be to offer both interactive and programmatic interfaces and to publish its services in the IVOA Registry. Visualization tools and Jupyter notebook style of working with data, thus using the data in the archive without having to download it, i.e. moving the code to where the data is stored in the archive. Another priority is to improve search speeds by creating data catalogues. An archive must make data discovery and access possible via different interfaces like a web interface or an API. A good description of what is in the data, location, and any limitations, thus lowering the expectations from the user. It is important to have parameter search options such as spatial, temporal, energy, and polarization. The archive should provide science platforms for visualization and further processing, user storage, and user databases.

It should have an authentication and authorization (AA) system that handles access to proprietary metadata, proprietary data, user storage, processing, and database allocations.

One of the archive staff at NRAO explained that having updated and accurate documentation of APIs allows access to data of known provencance in a format compatible with people's research methods. For the NOIRLab it is important to serve raw data and its associated calibration reduced as soon as possible after observations are complete and to serve the data with complete and accurate metadata. At ASTRON it is important to have UI and API capabilities for data discovery (searching to make appropriate selections given a science related objective) and data access (standard methods and protocols for users to retrieve and work with data from the archive). ASTRON provides long-term preservation of, and access to, curated astronomical data collections by following the FAIR principles.

5.2.2 Additional Metadata

In question two I'm interested in learning more about the extra metadata that their individual archives supply in addition to the information found in the file headers. Extra metadata provided by some archives is file size, version of the reduction software, and data release dates. At CDS additional metadata include IVOA metadata such as Uniform Content Descriptors (UCD), coordinates, and photometric system metadata. The CADC is using the Common Archive Observation Model (CAOM) to augment when the metadata is not present in the file header. They also work with their data sources (telescopes) to ensure the headers contain as much of the CAOM information as possible, including the FITS WCS standards. ASTRON's Apertif and LOFAR data archives include a catalog (database) of structured metadata is captured. They have descriptive metadata of the data products. In general, these are also in and extracted from, file headers. Comprehensive provenance includes descriptions and settings plus where applicable input data products used by, processes that contributed to the generation of data products. The catalogue registers all levels of contributing data products even if these are not stored in the archive themselves, e.g. certain raw or intermediate data products.

5.2.3 Data formats

For the next question, I'm interested in whether their archive offers a variety of data types and if there have been any requests to support other data formats. A variety of data formats are used by most archives but consistent among them is the use of the FITS data format. FITS is common among most responses because of its compression ability, tabular format, and image extension. The CDS makes use of a variety of data formats which include FITS, VOTable, TSV, XML+CSV, HTML, and ASCII. MeerKAT created their format called Visibility Format (MKv4) which provides access to the Measurement Sets. FITS are used for some of the observations and the PSRCHIVE data format for beamformer products. HTML, PDF, and PNGs formats are used for reports and thumbnails for quick views. Although in some cases the underlying formats are the same the data product could be different because

of the different metadata schemas that are created for each new format or report. At MAST the use of FITS files continues because of the IMAGE or BINTABLE extensions that it provides for images, spectra, and light curves. Image formats such as JPEG, GIF, and PNG are provided for product preview and display purposes. New missions such as the James Webb Space Telescope (JWST) are adopting the new ASDF format provided [JWST, 2019]. The CADC uses native telescope file formats, predominantly FITS but also TAR, SDF, PNG, CSV, and more recently HDF5. FITS is NOIRLab's primary format but they makes use of a miscellaneous file interface that can handle any file format since some visitor instruments produce things besides FITS files. At ASTRON the archives contain data products based on various data formats including Measurement Sets, FITS, and HDF5. Even though the archives are designed to support specifically defined data products, there is no strict requirement on the format. In particular, the LOFAR archive allows ingesting 'unspecified' data products for which minimal metadata is available and collected, and no specific data format is required. For obvious reasons this type of data is not desirable and the intention is to follow up at some point with a data curation effort to properly describe the data and offer defined/standard data product types.

5.2.4 Storage Technologies

Question four enquires about the storage technologies they have deployed at different archives and if other technologies have been considered for use and which ones they considered and the reasons behind the decision. For storage technologies, IUCAA uses an NFS-based file system with level 5 RAID configuration but also considered using Object stores and is recommending research into the use of HDFS in an astronomy context. The CDS uses RAID systems with duplicate servers and external mirror sites. They are considering using cloud infrastructure because this may give them a more flexible solution, and the benefit of the generic data infrastructures that are rapidly developing now (e.g. EOSC). SARAO uses a combination of storage technologies such as Ceph with an S3 gateway for large data store, which consist of approximately 20Petabyes (20PB) of spinning disks. A 20PB tape storage library for backup [MeerKAT, 2019]. For metadata search they use Solr which is a nosql DB. SARAO considered using elasticsearch, but Solr has a built in lattitude and longitude type and search which translates well to the celestial sphere and gives them a location search for their observations. Most of the archive data at MAST is stored and served from an Isilon appliance. The NRAO deployed the Next Generation Archive System (NGAS) software, on the enterprise Redhat Linux operating system, using a variety of hardware. The CADC uses Ceph Object Store, and Ceph Filesystem, both with middleware to manage the inventory. NOIRLab makes use of Amazon Web Services to provide cloud storage [Toro Rivera, 2021]. For data management middleware ASTRON use iRODS, for the Apertif archive and dCache (dcache.org) for the LOFAR archive. Both archives include RAID-based disk/online storage pools and tape robot storage backends.

5.2.5 Authentication Protocols

The question about authentication resulted in information about the protocols that the archive support and to know if other types are considered or requested. IUCAA used standard web-based authentication requiring username and password but also uses token-based authentication for APIs. They are considering using an Open Authentication framework but still want to do some research to help them understanding which Open Authorization (OAuth) providers (Github, Google, etc.) are preferred by the community. Data at CDS are fully open but there is still a CDS login with username and password authentication to download data. SARAO uses Security Assertion Markup Language (SAML2) and JSON Web Token (JWTs) to authenticate the archive. SAML is similar to OAuth, it provides SSO for their web portal. They then generate tokens to provide direct access to any data through the S3 endpoint [Armin, 2020]. They would like to join SAFIRE and eduGain for Identity Provision, both using SAML, so their technology would not have to change. INAF kept authentication protocols at a minimum. Usernames and passwords are kept in an administrative MySQL table. The archives contain data products based on various data formats including Measurement Sets, FITS, and HDF5. Even though the archives are designed to support specifically defined data products, there is no strict requirement on the format. In particular, the LOFAR archive allows ingesting 'unspecified' data products for which minimal metadata is available and collected, and no specific data format is required. For obvious reasons this type of data is not desirable and the intention is to follow up at some point with a data curation effort to properly describe the data and offer defined/standard data product types.

I discovered that the archives employ a variety of authentication mechanisms, including OAuth, SAML, Central Authentication Service (CAS), and traditional web-based authentication with usernames and passwords. OAuth2 is the industry-standard authorisation protocol that enables specific authorization flows for web apps, desktop applications, and mobile devices [Parecki, 2022]. SAML provides cross-domain web-based single sign-on (SSO), which reduces the administrative expense of providing numerous authentication tokens to the user [Ragouzis et al., 2008]. CAS is a web-based single sign-on mechanism used by NRAO. The goal of CAS is to allow a user to access various applications while just entering their credentials (user ID and password) once. It also enables online apps to verify users' identities without requiring access to a database [Apereo CAS, 2017].

MAST uses SAML2 and JWTs to authenticate to the archive. SAML is used in authentication to manage users centrally and OAuth is used for authorisation and to allow you to access one service from another without having to input your login credentials twice. Tokens are generated to provide direct access to any data through the S3 endpoint. At some point, they would like to join SAFIRE and eduGain for Identity Provision, because both are using SAML, so their technology would not have to change. MAST's archive applications support OAuth2 authentication. The authentication step is handled by the Shibboleth identity

provider. The OAuth2 layer makes it easier for them to allow users to generate tokens for programmatic access to their services. Usernames and passwords are kept in an administrative MySQL table. The tomcat interface talks jdnc to the MySQL server using a fixed internal username. The CADC uses OIDC, token, X509 certificates, and cookies. They follow the IVOA SSO and CDP standards for the implementation of authentication protocols. At ASTRON the native data management authentication protocols are supported which include username and password and tokens for iRODS, and X509 certificates and tokens for dCache. Both archives also offer user/pass authentication in web services that have been specifically developed for their archives. Integration with federated AAI is in progress.

5.2.6 Human Capacity Building

For this section the question is about the mechanisms that are in place to develop and strengthen user skills to use the archives. For complex archives that are serving instrument-specific data sets, it is useful to consider the use of Jupyter notebooks, Simple videos, and workshops for reduction-related aspects. The CDS supports training events in the context of VO Schools and workshops, and the services are also publicised in conferences and on social media. Demonstration booths are provided on an annual basis at AAS, EAS, and ADASS conferences for interaction with the community. CDS maintains a Help Desk email hotline. CDS provides documentation on the tools and services. SARAO has a helpful guide for using the web interface and provides a cookbook that shows how to perform common data access. They also provide a summer school where users will get to interact with the archive with help from SARAO staff while going through the cookbook and video tutorials. For MAST there is written documentation for all services and data collections and a small set of instructional videos. Recently there has been a drive to provide Jupyter notebooks which illustrate data discovery, access, and use via Python. The NRAO provides training initiatives led by observatory support staff. INAF provides documentation in the form of tutorials to the public and email correspondence for consortium users. The CADC's contribution to user skills improvement is supported by the documentation and responses to user queries through the Help desk, Slack and Github Issues. NOIRLab provides documentation that is available at the archive site (https://archive.gemini.edu/help/index.html). ASTRON offers Web-based documentation and tutorials, an online helpdesk for user support, and regular data schools for using the LOFAR archive.

5.2.7 Public Data Policy

In terms of proprietary periods and the right of access to certain data, I could like to know what the archive's policies are around making data public. The policy around making data public seems to be a practice among most archives that make use of proprietary periods policies except for CDS which doesn't have any restriction on access to the data. At the

SARAO archive, all data can only be accessed by the members of the proposal who requested a particular observation for a defined proprietary period. Membership of the proposal group is managed by the primary investigator or a delegated administrator who manages the data access. At the end of the proprietary period, the data becomes publicly accessible. Data can be made public before the proprietary period on request from the PI. The proprietary period which is normally one year is agreed upon between the head of science at SARAO and each group. The CADC archives data from multiple facilities and every facility has its proprietary policy, which is enforced by the archive. Most data has a release date as part of the metadata and as long as a release date is before the current date, the data automatically becomes publicly accessible. NOIRLab follows the observatory policy which states that standard queue data have a 1-year proprietary period. Fast Turnaround programs only have 6 months of proprietary period. In general, data in the archive is made public following a one-year proprietary period. Apertif is a survey instrument for which the associated science teams work on regular data releases of fully processed and curated data collections. Data included in published releases are made publicly (and openly, i.e. anonymously) available.

5.2.8 Interoperability Standards

This section is seeking an understanding of the standards are employed to support interoperability at the archives.

Interoperability standards deployed at the archives are Simple Cone Search (SC), and Simple Image Access Protocol (SIA). In addition to supplying data in FITS standard format, MAST makes use of some IVOA standards to support interoperability. CDS uses the following IVOA services, Table Access Protocol, Conesearch, SIAP, SSAP, HIPS, and SAMP. The MAST archive uses a subset of the CDS services and adds the DataLink protocol too. The standards used by SARAO are the protocols and formats like AWS S3, HTTP, SAML, JSON. SARAO has not implemented anything to make the archive VO compliant or produce VOTables and the like. The vo.astron.nl, which is registered in the IVOA registry and provides TAP, SIAP, ADQL, HIPS. An ObsCore compliant metadata schema is used for released data collections. INAF supports MySQL, FITS, and VOTable. The NRAO deployed NGAS, ASDM, and some people talked about IVOA but it was not clear from the answers if the future development will be resourced. The CADC deploys a variety of IVOA services.

5.2.9 Main Challenges

This section answers the question about the main challenges staff are facing in terms of supporting the archive. Astronomy-related data archives are under pressure to respond to many concurrent inquiries, which might put a lot of strain on their processes and resources

and cause serious problems with archive maintenance. In terms of the challenges faced by the archive, IUCAA, mentioned the lack of manpower and the need for dedicated staff. Normally the contractual staff who create the archives are generally different from those who need to support and maintain the archive. It is also difficult to get access to a talent pool that can leverage technologies such as distributed systems, cloud-native services, containerization, orchestration, etc to create self-healing and self-maintaining systems. At CDS the rapidly increasing data volume presents challenges for the physical infrastructure and costs. CDS is addressing this by the innovation of hierarchical systems for big data, e.g. HiPS. The increased complexity of data also presents a challenge, and this is addressed by choosing the appropriate level of standardisation. As astronomy and data science are rapidly evolving, it is a challenge to keep up with all of the new developments, so there needs to be a strategy of research and development so that new technologies and techniques can be adopted, not too soon, and not too late. At SARAO the main challenge is a lack of human resources to develop, maintain and support the archive. Only one person is working on the archive, web interface, and metadata catalogue, meaning one person has all the responsibilities. The commissioning scientists sometimes assist with user queries as they have direct relationships with many of the proposals. MAST has a mix of technology infrastructures that have evolved over many years without retiring the older ones. Maintaining these services in multiple stacks takes a very long time. Their biggest current challenge is ensuring that their services scale both to the increasing size of the datasets, but also the increased use of programmatic access to the archive. Programs can put much more load on the services than individual users of websites. There is a limit on manpower which means that software development, data ingest, and general system maintenance is handled by one single person. At INAF, there is a lack of support staff which is halting the software development process, data generation, and maintenance of the archive. A big challenge for CADC is to make sure that metadata quality and operational reliability standards are adhered to. For NOIRLab the biggest challenge is to retain support staff.

ASTRON's main challenges are data curation, which is to ensure that data in the archive is completely and correctly described. On ingest, providers do not always have or know complete information, or find it complex and time-consuming to collect it in an appropriate format, while on the operational side there are limited resources available to support/enforce this. Efforts are underway to complete missing information, remove obsolete/insufficient quality data, and also to add value by determining and annotating data quality but this is a significant undertaking. Storage capacity at the petabyte scale is a concern because infrastructure costs become significant. Creating awareness of data capacity being a scarce resource and finding and applying methods to either limit or reduce the size of data, with minimal impact on science content, is an ongoing effort. Infrastructure uniformity in a distributed archive involving multiple data centers, and a transparent and uniform user experience is a big challenge.

5.2.10 User Feedback Platforms

To get a view of how the archive is performing, I asked what mechanisms they provide to collect user feedback on their use of the archive. At IUCAA every request is logged into a database. A contact form is available which when filled triggers an e-mail to designated support staff. The CDS Helpdesk, user interaction at large community events (AAS, EAS, IAU, etc.), training events, social media. They intend to make user surveys in the future. I also just have discussions with people when they have new requirements. IUCAA, annually host a meeting of a standing Users' Group as well as send a survey to all known users. The help desk and having a booth at AAS meetings help to get feedback from the users.

The use of tools like Help desk, Slack, Github Issues, and a booth at the annual astronomy conference helps to gather user feedback that allows the archive teams to fix bugs, add features, simplify interfaces and improve documentation at the CADC. A user satisfaction survey goes out to all users every semester but is not specific to the archive. Feedback on LOFAR use, including the archive, is solicited approximately yearly at community meetings. For the Apertif archive, an online user questionnaire is offered to collect feedback but this did not result in any responses. Also for the Apertif archive, an assessment panel of experts in the astronomical archive domain has been formed in 2020. At regular intervals, usage statistics and developments are reported to the panel which provides the panel feedback on options to improve the user experience for the provided services and interfaces. In the future, this form of assessment and collecting feedback will include the LOFAR instrument data archive.

5.3 Summary

Based on the responses of the questionnaires', the major conclusions drawn from this are that archives are crucial. The majority of archives employ various methods, with certain tools and service deployments showing overlap. The security and authorisation tools and methods employed by archives share a number of similarities. The majority of archives use IVOA tools and services to facilitate interoperability. Every archive has a mechanism for user feedback via different platforms that helps to enhance the experiences of archive users. Surveys, help desk requests, user feedback booths at workshops and conferences, and any other kind of contact to address problems are examples of how to get feedback.

Chapter 6

Conclusions

6.1 Summary

The objective of this research is to identify best practices for developing astronomical data archives that exhibit interoperability between observatories and multi-wavelength data. Astronomy science should be inherently interoperable, not just in terms of scientific data and models, but also in the tools and interfaces used to exploit them. User surveys were conducted with archival support employees as well as archive users as part of the research. A series of questions were presented in order to get firsthand knowledge about the various archives that users tend to use, what tools and services archives give, and if any of the tools or services aid with interoperability.

Astronomy data archives are important for storing copies of observations at several wavelengths. Access to data archives has increased throughout time, allowing for more research to be undertaken. According to the literature, the IVOA tools and services increased data accessibility. Staff who specialize in large data storage, information security, database administration, and general information technology construct and manage archives that give platforms for access to research data to the broader astronomical community. The archive users approach their expectations and desired interactions with stored material differently. The ability of an archive to allow different methods of data access is an indicator of how popular the archive becomes among users.

The IVOA architecture is examined and explained in order to demonstrate the progression of data interoperability across wavelengths, which is beneficial to astronomy research across wavelengths. It allows astronomers to view all archived astronomy data through a single interface. The IVOA's tools and services enable data to be discovered, compared, and retrieved by merging multispectral and multiple instrument data from several archives. Each institution hosts its own data archives, however data may be exchanged by complying to extensible metadata standards and interchange protocols. Distributed virtual observa-

tory operations will be necessary, along with the core set of standards, interoperability, and administration services.

The key findings taken from the questionnaire replies are that archives are critical to multi-wavelength astronomy. Currently archives use a variety of strategies, with some tools and service deployments overlapping. There are certain parallels between the security and authorization tools and processes used by archives. To enable interoperability, the vast majority of archives employ IVOA technologies and services. Every archive provides a system for collecting user input via various platforms, which helps to improve archive users' experiences. How to obtain feedback includes surveys, help desk inquiries, user feedback booths at seminars and conferences, and any other type of communication to solve concerns. Multi-wavelength investigations have resulted in the discovery of countless new objects in space, allowing us to explore the Universe in ways that were not possible before.

The findings of the exploratory study, indicate that there are several barriers to providing improved interoperability and online access to the archives and metadata associated with them. Interoperability is a priority for some institutions, but more has to be done to prioritise it in practice. There are many approaches used, as well as organization of resources and collaboration across different institutions to achieve common goals. The breadth and availability of data will be expanded throughout the scientific community as a result. The use of data from numerous archives elevates multi-wavelength research to provide a deeper explanation of the findings in astronomy. The VO provides the tools and resources required to conduct multi-wavelength investigations.

Future research might include investigations into a range of views. More conceptual work, modeling of linkages between institutions (observatories/archives), authentication, and authorisation are required particularly for handling alerts generated at high frequencies to develop next-generation standards for distributed and interoperable data storage systems. In order to achieve technical and system levels of interoperability, investigations into the paths for institutions to take in order to participate in the federated archives are required. While the findings of this study gave some techniques for unlocking best practices for being interoperable across multiple wavelengths, the subject of interoperability remains a rich area of discussion for future research.

Bibliography

[AaaS, 2019] AaaS (2019). *The Canadian Astronomy Data Centre - Archive as a Service.* Last accessed 21 January 2019 https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/doc/AaaS/.

[Adam et al., 2018] Adam, D., Cristiano, D., Ignacio, V., Vuong, M., Thomas, B., Vincenzo, F., Nathalie, F., Christophe, M., and Stefano, Z. (2018). *ESO Archive Data and Metadata Model.* ESO.

[Aladin, 2020] Aladin (2020). *Aladin Sky Atlas.* Aladin.

[ALMA, 2021] ALMA (2021). *Where is ALMA data stored?* Last accessed 1 October 2021 https://www.almaobservatory.org/en/about-alma/how-alma-works/how-alma-observations-are-carried-out/.

[ANDS and NCRIS, 2017] ANDS and NCRIS (2017). *10 Astronomy Research Data Things.* National Research Infrastructure for Australia.

[ANU, 2021] ANU (2021). *SkyMapper.* Last accessed, 5 December 2021 https://skymapper.anu.edu.au/.

[Apereo CAS, 2017] Apereo CAS (2017). *Apereo Central Authentication Service project.*

[Armin, 2020] Armin, H. (2020). *Usage of external JWT in CDS Services.*

[Arviset and Gaudet, 2010] Arviset, C. and Gaudet, S. (2010). *IVOA Architecture Version 1.0.* IVOA.

[Arviset et al., 2006] Arviset, C., Ortiz, I., Osuna, P., and Salgado, J. (2006). *ESA Scientific Archives and inter-operable Virtual Observatory systems.*

[Astron, 2021] Astron, S. (2021). *Making Discoveries in Radio Astronomy happen!*

[ASTRON Editorial Team, 2018] ASTRON Editorial Team (2018). *ASTRON joins ESCAPE project to build open science cloud.*

[Ball, 2010] Ball, N. (2010). *KDD-IG and CANFAR.* CADC, Herzberg Institute of Astrophysics, Victoria.

[Berriman, 2011] Berriman, B. (2011). *HOW WILL ASTRONOMY ARCHIVES SURVIVE THE DATA TSUNAMI?*

[Berriman, 2022] Berriman, B. (2022). *A Brief History of the IVOA*. Last accessed, 2 April 2022:https://wiki.ivoa.net/twiki/bin/view/IVOA/NewcomersIntroIVOABasics.

[Bertocco, 2017] Bertocco, S. (2017). *ASTERICS DADI ESFRI Forum*.

[Bertocco et al., 2017] Bertocco, S., Major, B., Dowler, P., Gaudet, S., Taffoni, G., and Pasian, F. (2017). *Interoperable geographically distributed astronomical infrastructures: technical solutions*. Vera C. Rubin Observatory project.

[Bertoccoa et al., 2018] Bertoccoa, S., Dowler, P., Gaudet, S., Major, B., Pasiana, F., and Taffonia, G. (2018). *Cloud access to interoperable IVOA-compliant VOSpace storage*.

[Bonnarel et al., 2008] Bonnarel, F., Dowler, P., Noddle, K., and Tody, D. (2008). *HDF5 Advanced Topics*. HDF and HDF-EOS Workshop XII.

[Bonnarel et al., 2015] Bonnarel, F., Dowler, P., Noddle, K., and Tody, D. (2015). *IVOA Data Access Layer: Goals, Achievements and Current Trends*.

[Breitenfeld et al., 2020] Breitenfeld, M., Pourmal, E., Byna, S., and Koziol, Q. (2020). *Achieving High Performance I/O with HDF5*.

[CADC, 2020] CADC (2020). *Canadian Astronomy Data Centre*. Last accessed,9 March 2020 https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/.

[CADCDoc, 2019] CADCDoc (2019). *CADC Documentation*. Last accessed, 3 March 2019 https://www.cadc-ccda.hia-iha.nrc-cnrc.gc.ca/en/doc/.

[CANFAR, 2021] CANFAR (2021). *CANFAR Storage Management*. Last accessed, 5 December 2021 https://www.canfar.net/en/docs/storage/.

[Chandler, 2014] Chandler, C. (2014). *Data Policies* . Last accessed, 21 March 2014 https://science.nrao.edu/observing/proposal-types/datapolicies.

[Chapman et al., 2015] Chapman, J., Dempsey, J., Miller, D., Heywood, I. Pritchard, J., Sangster, E., Whitting, M., and Dart, M. (2015). *CASDA: The CSIRO ASKAP Science Data Archive*. ASP Conference Series, Vol. 512.

[Chenzhou et al., 2012] Chenzhou, C., Dongwei, F., Yongheng, Z., Ajit, K., Boliang, H., Zihuang, C., Jian, L., and Deoyani, N. (2012). *Enhanced Management of Personal Astronomical Data with FITSManager. Enhanced Management of Personal Astronomical Data with FITSManager*.

[Dashofy et al., 2002] Dashofy, E., van der Hoek, A., and Taylor, R. (2002). *Towards architecture-based self-healing systems*. WOSS '02: Proceedings of the first workshop on Self-healing systems.

[Databricks, 2021] Databricks (2021). *https://parquet.apache.org/documentation/latest/*. PARQ.

[Dowler et al., 2021] Dowler, P., Evans, J., Arviset, C., Gaudet, S., and IVOA Technical Coordination Group (2021). *International Virtual Observatory Alliance - IVOA Architecture Version 2.0*.

[Drakos, 1998] Drakos, N. (1998). *How to use the MIDAS Manual*. Last accessed, 3 March 2019 https://www.eso.org/sci/software/esomidas/doc/user/18NOV/vola/node5.html.

[Dubois-Felsmann et al., 2019] Dubois-Felsmann, G., Economou, F., Lim, K., Mueller, F., Pietrowicz, S., and Wu, X. (2019). *Large Synoptic Survey Telescope (LSST) Data Management Science Platform Design*.

[EGI, 2017] EGI (2017). *Final CANFAR Integration Release*. https://documents.egi.eu/public/ShowDocument?docid=3038.

[ESOArchive, 2010] ESOArchive (2010). *CESO Data Archive Facility*. Last accessed, 3 March 2019 http://archive.eso.org/cms.html.

[Fabio, 2009] Fabio, P. (2009). *Management of Astronomical Data Archives and their Interoperability through the Virtual Observatory, IVOA*.

[FITS Working Group, 2018] FITS Working Group (2018). *Definition of the Flexible Image Transport System*. FITS.

[Galluzzi et al., 2019] Galluzzi, V., Cavallaro, F., Costa, A., Knapic, C., Sciacca, E., Rygl, K., Liuzzo, E., and Massardi, M. (2019). *Applicability of VO framework*.

[Gosh et al., 2006] Gosh, D., Sharman, R., Rao, R., and Upadhyaya, S. (2006). *Self-healing systems — survey and synthesis*. Science Direct.

[Greene and Plante, 2008a] Greene, G. and Plante, R. (2008a). *Chapter 7: Web-based Tools—The NVO Registry: Finding Data and Services in the NVO*. ASP Conference Series, Vol. 382, © 2008.

[Greene and Plante, 2008b] Greene, G. and Plante, R. (2008b). *Chapter 41: An Overview of the Registry Framework*. ASP Conference Series, Vol. 382, © 2008.

[Greenfield et al., 2015] Greenfield, P., Droettboom, M., and Bray, E. (2015). *ASDF:A new data format for astronomy*. ASDF.

[Guy, 2018] Guy, L. (2018). *LSST Alert Streams Solar System Science*.

[Hanisch, 2007] Hanisch, B. (2007). *The Origins, Formation, and Scientific Promise of the Virtual Observatory*.

[Hanisch et al., 2016] Hanisch, B., Greene, G., and O'Mullane, W. (2016). *The National Virtual Observatory*.

[Hanisch et al., 2007] Hanisch, R., IVOA Resource Registry Working Group, and NVO Metadata Working Group (2007). *Resource Metadata for the Virtual Observatory Version 1.12*.

[Hanisch et al., 2010] Hanisch, R., Quinn, P., De Young, D., Padovani, P., and Pasian, F. (2010). *Guidelines for Participation Version 1.1*. IVOA.

[ICRAR, 2021] ICRAR, O. (2021). *Data-Intensive-Astronomy*. Last accessed, 21 October 2021 https://www.icrar.org/our-research/data-intensive-astronomy/.

[ICRARDIA, 2019] ICRARDIA, O. (2019). *DATA INTENSIVE ASTRONOMY*. Last accessed, 10 May 2019 https://www.icrar.org/wp-content/uploads/2019/05/DIA-Brochure.pdf.

[IRSA, 2020] IRSA (2020). *Infrared Science Archive*. Last accessed, 8 August 2020 https://www.ipac.caltech.edu/project/irsa.

[Iyer and DuttaDuwarah, 2018] Iyer, G. and DuttaDuwarah, S. ane Sharma, A. (2018). *DataScope: Interactive Visual Exploratory Dashboards for Large Multidimensional Data*.

[Juric, 2015] Juric, M. (2015). *LSST Data Management: Building the Data System for the Era of Petascale Optical Astronomy*. LSST Data Management Project Scientist WRF Data Science Chair in Astronomy, University of Washington.

[Juric et al., 2021] Juric, M., Kantor, J., Lim, K.-T., Lupton, R., Dubois-Felsmann, G., Jenness, T., Axelrod, T., Aleksic, J., Allsman, R., AlSayyad, A., Alt, J., Armstrong, R., Basney, J., Becker, A., Becla, J., Biswas, R., Bosch, J., Boutigny, D., Kind, M. C., Ciardi, D. R., Connolly, A., and Daniel, S. F. (2021). *LSST: Data Management*. Vera C. Rubin Observatory project.

[JWST, 2019] JWST, . (2019). *Understanding JWST Data Files*. Last accessed, 7 October 2019 https://jwst-docs.stsci.edu/getting-started-with-jwst-data/understanding-jwst-data-files.

[Kavelaars, 2016] Kavelaars, J. (2016). *CADC/CANFAR An Integrated Science Platform for JCMT Users*.

[Keck, 2020] Keck (2020). *The Keck Observatory Archive*. Last accessed, 25 August 2020 https://www.ipac.caltech.edu/project/keck-archive.

[Kellermann et al., 2021] Kellermann, K., Bouton, E., and Brandt, S. (2021). *Open Skies: The National Radio Astronomy Observatory and Its Impact on US Radio Astronomy*. Springer.

[Kern et al., 2019] Kern, J., Robnett, J., and Glendenning, B. (2019). *The Science Ready Data Products Revolution at the NRAO*. NRAO Doc. : 530-SRDP-043-MGMT Version: 1.0.

[Knowk and Tody, 2008] Knowk, S. and Tody, D. (2008). *Chapter 47: Introduction to DAL: Simple Image Access Protocol.* ASP Conference Series, Vol. 382, © 2008.

[KOA, 2020] KOA (2020). *W. M. Keck Observatory Archive.* Last accessed, 25 August 2020 https://koa.ipac.caltech.edu/UserGuide/about.html.

[La Plante et al., 2021] La Plante, P., Williams, P., Kolopanis, M., and Zheng, H. (2021). *A Real Time Processing system for big data in astronomy: Applications to HERA.*

[Lacy and Bonnarel, 2020] Lacy, M. and Bonnarel, F. (2020). *Radio Interest Group.* International Virtual Observatory Alliance.

[Lacy et al., 2020] Lacy, M., Kern, J., and Tobin, J. (2020). *The NRAO Science Ready Data Products Pilot Program.* ASP Conference Series, Vol. 527.

[Louys, 2020] Louys, M. (2020). *Radio Astronomy visibility data discovery and access using IVOA standards.*

[Luminousmen, 2022] Luminousmen (2022). *Big Data file formats.* Last accessed, 23 February 2022 https://luminousmen.com/post/big-data-file-formats.

[MAST, 2015] MAST (2015). *Multimission Archive at STScI.* Last accessed, 19 February 2021 http://archive.stsci.edu/manuals/archive$_h$*andbook/chap*1.*html*.

[MASTAccess, 2020] MASTAccess (2020). *How To Access MAST Data.* Last accessed: 28 August 2020 https://archive.stsci.edu/publishing/data-use.

[MASTAstroQuery, 2020] MASTAstroQuery (2020). *MAST Queries.* Last accessed, 28 August 2020 https://astroquery.readthedocs.io/en/latest/mast/mast.html.

[MeerKAT, 2019] MeerKAT, . (2019). *SKA, MeerKAT and other Radio Astronomy Projects: Overview and Update.*

[Mink et al., 2020] Mink, J., Diaz, R., Fernique, P., Michel, L., Louys, M., and Landais, G. (2020). *Data Formats Bof.* ADASS.

[Mink et al., 2014] Mink, J., FITS Technical Group, Mann, B., and Astronomy and Computing (2014). *Astronomical Data Formats Past, Present Future.* ADASS2014.

[Molinaro and Dempsey, 2020] Molinaro, M. and Dempsey, J. (2020). *IVOA Data Access Layer: roadmap as of year 2020.*

[MWA, 2021] MWA (2021). *The Murchison Widefield Array.* Last accessed, 6 December 2021: https://www.mwatelescope.org/telescope/data/.

[NASA, 2013] NASA (2013). *Multiwavelength Astronomy.* Last accessed, 21 April 2021 https://imagine.gsfc.nasa.gov/science/toolbox/multiwavelength1.html.

[National Research Council, 2007] National Research Council, Division on Engineering and Physical Sciences, S. C. (2007). *Portals to the Universe: The NASA Astronomy Science Centers*. The National Academies Press.

[Nebot, 2022] Nebot, A. (2022). *Intro to the IVOA - Interop meeting 25-29 April 2022*.

[NGAS, 2020] NGAS (2020). *The Next Generation Archive System*. Last accessed, 27 August 2020 https://github.com/ICRAR/ngas.

[NOIRLab, 2019] NOIRLab (2019). *National Optical-Infrared Astronomy Research Laboratory*. Last accessed, 14 October 2019 https://noirlab.edu/public/about/faqs/.

[NRAO, 2021a] NRAO (2021a). *NRAO Available Hardware Resources*. Last accessed, 6 December 201 https://info.nrao.edu/computing/guide/cluster-processing/appendix/available-hardware-resources.

[NRAO, 2021b] NRAO (2021b). *NRAO cluster-processing*. Last accessed, 6 December 2021:https://info.nrao.edu/computing/guide/cluster-processing.

[NRAO, 2021c] NRAO (2021c). *NRAO Data Storage and Retrieval*. Last accessed, 6 December 2021:https://info.nrao.edu/computing/guide/cluster-processing/data-storage-and-retrieval.

[O'Toole and Sealy, 2019] O'Toole, S. and Sealy, K. (2019). *Bringing Together the Australian Sky - Coordination and Interoperability Challenges of the All-Sky Virtual Observatory*. Astronomical Data Analysis Software and Systems XXVIII.

[Parecki, 2022] Parecki, A. (2022). *OAuth 2.0*. Last accessed, 15 December 2022:https://oauth.net/2/.

[Pasian and Molinaro, 2017] Pasian, F. and Molinaro, M. (2017). *The FAIR paradigm as a key to Open Astronomical Data*. INAF Osservatorio Astronomicodi Trieste, Italy ASTERICS and EOSCpilot projects.

[Pasion, 2013] Pasion, F. (2013). *Data Management and Archiving in Astronomy: Status and Challenges for the Future*.

[PAWSEY, 2021] PAWSEY (2021). *Introduction to Pawsey*. Last accessed, 2 January 2021 https://pawsey.org.au/about-us/the-pawsey-centre/.

[Price et al., 2015] Price, D. C., Barsdell, B. R., and Greenhill, L. J. (2015). *HDFITS: porting the FITS data model to HDF5*. https://arxiv.org/abs/1505.06421.

[Radziwil, 2011] Radziwil, N. (2011). *End to End Operations at the National Radio Astronomy Observatory*.

[Ragouzis et al., 2008] Ragouzis, N., Hughes, J., and Philpott, R. and Maler, E. and Madsen, P. and Scavo, T. (2008). *Security Assertion Markup Language (SAML) V2.0 Technical Overview*. OASIS Open.

[SAAO, 2022] SAAO (2022). *Data Discovery Tools.* Last accessed, 16 May 2022 https://www.saao.ac.za/astronomers/data-discovery-tools/.

[Seaman et al., 2017] Seaman, R., Williams, R., Allan, A., Barthelmy, S., Bloom, J., Brewer, J., Denny, R., Fitzpatrick, M., Graham, M.and Gray, N., Hessman, F., Marka, S., Rots, A., Vestrand, T., and Wozniak, P. (2017). *Sky Event Reporting Metadata (VOEvent)Version 2.0.* Internatoinal Virtual Observatory Alliance.

[Shupe et al., 2020] Shupe, D., Desai, V., and Groom, S. (2020). *Working with large catalogs using the Astronomy eXtensions for Spark (AXS) framework.* ADASS.

[Simbad, 2020] Simbad (2020). *Simbad .* Last accessed, 27 August 2020:https://simbad.u-strasbg.fr/guide/simbad.htx.

[Swinbank, 2021] Swinbank, J. (2021). *The ASTRON Science Data Centre.* Last accessed, 15 June 2021 https://www.astron.nl/astronomy/.

[TAP, 2019] TAP (2019). *SkyMapper Southern Sky Survey.* Last accessed, 6 December 2019 https://skymapper.anu.edu.au/.

[Thomas et al., 2015] Thomas, B., Jenness, T., and Economou, F. (2015). *Learning from FITS: Limitations in use in modern astronomical research.* NOAO.

[Toro Rivera, 2021] Toro Rivera, E. (2021). *NOIRLAB: International Gwimini Observatory - AMREN Project Meeting.* AMREN Project Meeting - June 2021.

[UEDIN et al., 2022] UEDIN, LU, ESO, and CNRS DR10 (2022). *The European Virtual Observatory - VO Technology Centre.*

[Vizier, 2020] Vizier (2020). *Vizier - library of published astronomical catalogues.* Last accessed, 7 August 2020 https://vizier.cds.unistra.fr/index.gml.

[Witz et al., 2019] Witz, S., Lyons, D., Plank, J., Hausman, C., Lively, R., Arora, J., and Benson, J. (2019). *Towards a Self-healing Archive.* ASP Conference Series, Vol. 521.

[Wynholds et al., 2011] Wynholds, L., Fearon, D., Borgman, C., and Traweek, S. (2011). *When Use Cases Are Not Useful: Data Practices, Astronomy, and Digital Libraries.*

[Young, 2004] Young, D. (2004). *Software Archive and NVO.* NOAO, 2024.