ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-THPDP048

SARAO SCIENCE REPOSITORY: SUSTAINABLE USE OF MeerKAT DATA

Z. Kukuma, G.L. Coetzer[1], R. S. Kupa, C. Scholar
South African Radio Astronomy Observatory, Cape Town, South Africa
[1]also at University of Pretoria, Pretoria, South Africa

Abstract

The South African Radio Astronomy Observatory (SARAO) is excited to announce the forthcoming release of its digital repository for managing astronomical data of the MeerKAT Radio Telescope. The repository, built using DSpace software, will allow researchers to catalogue and discover research data in a standardised way, while Digital Object Identifiers (DOIs) minted through the DataCite service, will ensure the unique identification and persistent citation of this research data. In this paper, we discuss the design of the repository as well as the use of DataCite for DOI minting.

INTRODUCTION

In recent years, mechanisms for data discovery and retrieval have been established. Due to these mechanisms, new and unforeseen uses of data are being discovered [1]. With the Square Kilometre Array (SKA) radio telescope project and the new Multi-Purpose Reactor (MPR) soon coming online, organisations such as SARAO which facilitates these projects will need to make the management, analysis, publication and curation of big scientific data a priority. The recent draft of the Department of Science and Innovation (DSI) on Open Science (OS) policy, requires Open Access (OA) to both scholarly publications and scientific data [2]. It also endorses ingest, discovery and dissemination of data and metadata in a manner consistent with Wilkinson's [3] 'FAIR principles' - making data Findable, Accessible, Interoperable and Reusable (FAIR) [4]. SARAO recognises the importance of repositories and Digital Object Identifiers (DOIs) as mechanisms which can improve the FAIRness of data [5], and follows the National Research Foundation (NRF)'s vision 2030 of putting "science and research into service for a better society [6]. With this in mind, SARAO's librarian and software engineers developed a digital repository.

OPEN SCIENCE, OPEN ACCESS AND DIGITAL OBJECT IDENTIFIERS (DOIs)

The SA draft National OS policy defines OA as "a set of principles through which research outputs are distributed online, free of cost or other access barriers" [2]. Moreover, OS includes both OA to research output and unhindered access to accompanying raw data and software used to analyse the data. OS is particularly important because it [7]:

- promotes accurate verification of research outputs;
- reduces duplication;
- promotes innovation and increased consumer choice;
- promotes citizen's trust in science;
- promotes public participation in research.

To make OS a reality, UNESCO [7] designed a set of principles to ensure that there is:

- transparency and reproducibility of research outputs;
- enhance impact of science on society;
- collaboration which transcends geography, language and resources barriers;
- solve problems of great social importance
- acknowledgement that there there does not exist a one-size-fits-all method for practising OS
- encourage pathways to practice
- sustainability by building on long-term practises, services, infrastructures and funding models.

DOIs are vital for OS, OA and FAIR data practises. A DOI is an alphanumeric used to uniquely identify objects. Objects can be articles, documents, software, networks, scientific data and products, et cetera. A resolvable DOI consists of a resolver, prefix and suffix as shown in Fig. 1.

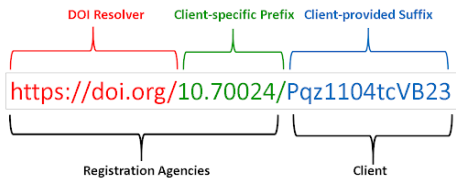

Figure 1: Composition of a DOI (adapted from Novacescu [8]).

A DOI is assigned to a web address (URL) where an object (e.g. dataset) and its metadata can be located permanently [9]. DOIs help to ensure the integrity of digital objects - with a persistent identifier that remains unchanged even if the location changes, users can always find and access the correct version of the digital object. A DOI's value as a persistent identifier is lost if the DOI metadata is not updated when the object it references changes physical location (i.e. the URL changes). The publisher or responsible party must update the URL in the metadata record to ensure that the DOI continues to redirect users to the object. Metadata collected before data release are generally used for formal citation, thus facilitating credit and acknowledgement for data creation and usage. DOIs allow different platforms, databases and information systems to exchange information consistently and unambiguously.

Examples of softwares that encourages sustainable information systems are DuraSpace (DSpace), Figshare and Zenodo amongst others [9]. They are technical frameworks

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-THPDP048

that are used for creating digital repositories [10]. Components constituting a digital repository at its most basic level, include a database, a retrieval mechanism and user interface (UI). Digital repositories have their own functional requirements. The requirements are based on the needs of users and governing organisations. In most cases they cater for the creation, integration, structuring, processing, storage and retrieval of data, and support management of DOIs [11].

When a DSpace repository is configured it must be registered with a DOI Registration Agency (RA), e.g. DataCite or CrossRef. RAs provides the repository with a Handle prefix, in order to identify the "naming authority". In the example given on Figure 1, the naming authority is doi.org. In Fig. 1, the prefix is the `10.70024` number which identifies the local name. The suffix is `Pqz1104tcVB23`, an alphanumeric number which identifies the scientific content. After content has been registered with a RA, users will be able to retrieve identifiers and create links and landing pages. DOIs should resolve to unique landing pages and not directly to the content, e.g. PDF, data file [12]. The DOI landing page (the URL) is used to denote the website to which a DOI resolves [13].

SARAO'S IMPLEMENTATION OF OS PRINCIPLES

The current mechanism used for managing SARAO's scientific outputs is through a DOI RA DataCite Fabrica. This organisation is responsible for minting a DOI link that retrieves identifiers. Datacite Commons is responsible for storing the digital assets that are minted through Datacite Fabrica.

DATACITE COMMONS

Datacite Commons launched in 2020 [14] is one of Datacite Fabrica's services. It is a discovery tool that allows searches via persistent identifiers (PIDs), e.g. DOIs, individual's Open Researcher and Contributor Id (ORCIDs), organisations unique identification codes and repositories unique identifiers. DataCite Commons is a global consortium formed to provide digital accessibility to research data, increase accepted level of research data as a resource type, increase citation of research data and to provide data management that will allow results to be reused and verified. The most significant feature of Datacite Commons is the ability to scan all DOIs regardless of whether they were registered with DataCite Fabrica. Data Commons can also reveal the relationships between DOIs, content, people, research, organisations and funders. That being said, drawbacks of DataCite Commons include: it is a relatively new platform and has limited features (e.g. advanced search capabilities).

DSpace SOFTWARE

The OA movement has played a defining role in the evolution of digital repository software. Examples of repository tools which can easily be deployed in different scenarios include DSpace, Figshare and Fedora [15].

SARAO's current repository (i.e. DataCite Commons) has several drawbacks. A new digital repository for SARAO's astronomical data was designed, using DSpace. The choice of software is based on the fact that many academic and research as well as non-profit and commercial organisations have been using DSpace for more than a decade. Examples of these organisations Raman Research Institute(RRI) in Physics, National Centre for Radio Astrophysics (NCRA), Centre for International Climate Research (CICERO), Lunar and Planetary Institute, and University of Pretoria (UP) [16–18].

DSpace is an open-source web application which allows researchers and organisations to capture, store, index, preserve and publish data, metadata, documents, video, etc. [19]. The software serves a specific need as a digital archive system which is completely customisable to fit the information needs of organisations and focuses on long-term storage and preservation of digital content. It is used widely for the creation of OS digital repositories. Benefits of using DSpace are that: it is free and easy to install; has no financial requirements beyond the cost of the hardware; has no recurring licence fees; is scalable; and both Linux or MS Windows can be used as operating systems. It is interoperable with other open standards-based systems and supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). DSpace supports all file formats (e.g. Pdf and JPEG) [20] and can organise, describe and store content easily through built-in structures. I can link to search engines (e.g. Google) and contain persistent network identifiers. DSpace can manage access to collections [21]. In addition, it provides an out-of-the-box user interface which provides basic and advanced search options. Disadvantages of DSpace include: customisation of the software requires technically skilled people; it has record editing limitations, and thus authorisation can be difficult and cumbersome and its web standards limit its interface features. DSpace does not exist in isolation but depends on a raft of Java projects (e.g. Cocoon) [20]. Administration and documentation of upgrades can be taxing and difficult for the novice to understand. DSpace's developments and upgrades rely on the good-will of the OS community which can be frustrating and lead to lengthy delays in problem solving. Paid support is available but is expensive. Also, the fact that DSpace is a large and complicated system can be seen as a limitation. Nonetheless, DSpace's advantages outweigh its disadvantages.

The General Architecture of DSpace

DSpace provides the basis for a digital repository. The functionality of its database can be extended and configured to suit a user. DSpace architecture stems mainly from three sources namely the Framework for Distributed Digital (FDD) services [22, 23] and the Consultative Committee for Space Data Systems' Reference Model [24].

The information model or flow of an institution is based on the idea of the institution's subunits. In DSpace these subunits are called Communities (e.g. if an institution is

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-THPDP048

a university then its communities would be the different schools, departments, labs, etc.). Most institutions are multidisciplinary, so they will most likely have different communities which deposit digital assets. The differing communities will also have distinct groupings of digital assets from various departments relevant to a specific subject (e.g. Physics or Mathematics). These groupings are referred to as Collections. The digital assets that are indexed into various collections are referred to as items. Following the example used above, these items will be different publications or datasets.The single files that constitute an item are referred to as bitstreams (e.g. png files, tables or audio files). The technical architecture of DSpace is discussed below.

The Frontend/User Interface

DSpace's User Interface (UI) is web-based. It is built on an angular framework and is easy and flexible to customise and modify to the organisation's specifications. A DSpace angular repository is readily available for installation with necessary local dependencies. It can be installed using Yarn. Documentation to customise DSpace is available. DSpace has two broadly used UIs: the JavaServer Pages UI (JSPUI) and Extensible MarkUp Language UI (XMLUI). JSPUI provides a user-friendly interface for both administrators and end users, while XMLUI provides different themes and aspects (e.g. web pages, metadata forms, user interactions, dynamic responses, media filtering, statistics tools). The UI consists of: rest APIs for communication applications, Resource Description Frameworks (RDFs) for representing data that is interconnected on the web and a protocol that facilitates the remote deposit of items into repositories, referred to as Simple Web-service Offering Repository Deposit (SWORD). DSpace allows systems to connect and communicate with one another.

DSpace has a modular architecture. It can be extended or reduced by adding or removing new plugins/modules. Modules/plugins can add features and functionality to platforms and can be integrated with different systems (e.g. databases and content management systems). The public facing interface referred to as the application layer allows for finding and accessing items and bitstreams. It contains components that create a platform for users to interact with the content in the storage layer or database. However, the frontend can not work without a backend. DSpace offer's fictional data when installed so that users can install and connect via HTTP/S, to the backend on the same or different computers.

The Backend (Server)

The UI and server of DSpace require different technologies. The DSpace code base is developed in Java, and requires the UNIX-like (Linux, HP/UX, Mac OS X, etc.) or Microsoft Windows operating systems. The server- side web application is developed via Node.js. Web pages are provided through Apache Tomcat. For secure communication between applications a OpenSSL/mod_ssl protocol is used. The ability to demonstrate structured, full-text search across pipelines that are of high-dimension is made possible by Apache Lucene. The storage layer or database is where the content and its corresponding metadata are physically stored. DSpace is a data intensive application. Data are stored in relational databases (e.g. postgreSQL or Oracle). Databases store metadata for items, bitstreams, configurations and administrative information. This form of storage allows for different types of data, in different formats (e.g. text, images, videos and audios) to be stored.

DSpace business layer is responsible for mediating the interactions between the database and UI. It handles authorisation and authentication processes, and manages the workflow of the system. This component contains the administrative toolkit and group manager for E-persons.

Installation and Configuration

DSpace's code base, download, installation and operational instructions are available on Github [25]. This webpage provides an extensive elaboration on how to set up and configure files, to customise and modify files as per organisational requirements. DSpace backend technology consists of the server, postgreSQl, SOLR containers, and the frontend consists of the UI container. These four containers are configured with fictional data to create a digital repository. They are containerised using docker images that can be pulled from dockerhub. To then customise the repository for an organisation, the images can be pulled, configured and rebuilt. Some configurations can change the database and URLs. Different organisations will have different data that they are storing in the repository. Following the example repository mentioned previously, the communities would be changed to the different schools, departments and labs, cascaded with the different collections that they want to store in the different communities. Customising the UI is relatively easy. UIs can be branded to reflect an organisation's identity [15].

THE SARAO REPOSITORY

SARAO installed DSpace and customised it to suit the organisation's needs for a science repository. Benefits such as a configurable operating system project, a community of developers and researchers constantly improving DSpace, made the selection of this type of software appealing for the design of a digital repository for SARAO. The aim of SARAO's digital repository is to be a platform to make SARAO's scientific data FAIR and support OS.

For SARAO, we have adapted and customised the architecture of DSpace to suit the needs of the organisation. In Fig. 2 the architecture of the system is illustrated. The architecture shows the system's three tiered layer. The application layer was customised to interact with users. The business Layer is where the content management, search and authentication resides, while the database layer is used to retrieve and store data from PostgresQL and Solr.

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-THPDP048

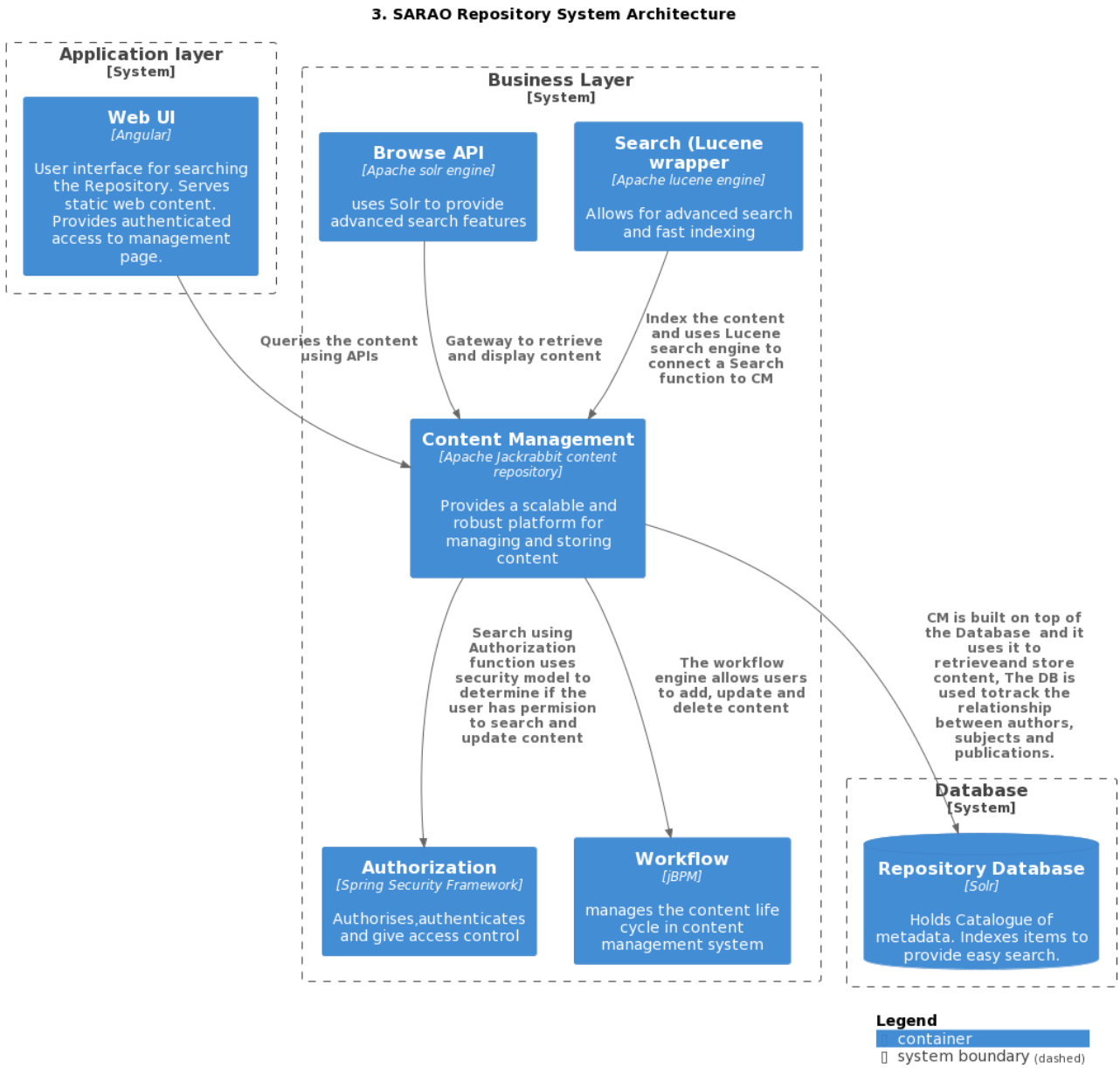

Figure 2: Architectural structure of The SARAO Repository [26].

FUTURE DEVELOPMENTS

Areas of improvement of the SARAO repository have been identified. For example the automation of DOI creation instead of the current tedious manual process. Regular organisational needs assessments will be conducted in future. These assessments will inform future developments of the system.

CONCLUSIONS

With the increasing demand for OS, OA and digital preservation, SARAO recognises the importance of a repository as tools to accelerate the organisation's data visibility, discovery, accessibility, interoperability, usability as well as reporting and acknowledgement [5]. The organisation is excited to announce the release of this digital repository for managing and preserving astronomical data. We believe that SARAO's digital repository will set a standard to be followed by other astronomical institutions.

REFERENCES

[1] S. Stall *et al.*, "Advancing FAIR data in Earth, space, and environmental science. Eos, Earth and Space Science News", Eos Transactions American Geophysical Union, vol. 99, p. 109301. doi:10.1029/2018eo109301

[2] H. Pienaar, "Draft National Open Science Policy. Department of Science and Innovation", *Front. Res. Metr. Anal.*, vol. 8, p. 1233867, Jun. 2023. doi:10.3389/frma.2023.1233867

ISBN: 978-3-95450-238-7 ISSN: 2226-0358 doi:10.18429/JACoW-ICALEPCS2023-THPDP048

[3] M.D. Wilkinson *et al.*, "The FAIR guiding principles for scientific data management and stewardship", *Sci. Data*, vol. 3, p. 160018. doi:10.1038/sdata.2016.18

[4] T. Hey, "Open science and Big Data in South Africa", *Front. Res. Metr. Anal.*, vol. 7, 2022. doi:10.3389/frma.2022.982435

[5] G. Coetzer, R. Botha, C. Schollar, and K. Elger, "An institutional research data repository and digital object identifier for SARAO radio astronomy, fundamental astronomy, astrometry and geodesy", *Bulletin of the AAS*, vol. 54, no. 2, Apr. 2022. doi:10.3847/25c2cfeb.66ee866c

[6] National Research Foundation (NRF), https://www.nrf.ac.za/wp-content/uploads/2021/03/NRF-Vision-2030_0.pdf

[7] Organisation for Economic Co-operation and Development, https://www.oecd.org/sti/inno/open-science.htm

[8] J. Novacescu *et al.*, "A model for data citation in astronomical research using DOIs", *Astrophys. J. Suppl. Ser.*, vol. 236, no. 1, p. 20, Dec. 2017. doi:10.3847/1538-4365/aab76a

[9] DOI Foundation, https://www.doi.org/

[10] C. Curdt and D. Hoffmeister, "Research data management services for a multidisciplinary, collaborative research project", *Program: electron. Lib. Inf. Syst.*, vol 49, no. 4, pp. 494-512, Sep. 2015. doi:10.1108/prog-02-2015-0016

[11] S. Kramer, "Matrix of use cases and functional requirements for research data repository platforms", *Res. Data Alliance*, Sep. 2016. doi:10.15497/rda00033

[12] DataCite, https://support.datacite.org/docs/landing-pages

[13] Open Research, https://www.openresearch.org/wiki/Property:DOI_landing_page

[14] Datacite, https://support.datacite.org/docs/datacite-commons#:~:text=DataCite%20Commons%20is%20work%20in,commons.datacite.org%2F

[15] Nils Körber and Hussein Suleman, "Usability of digital repository software: a study of DSpace installation and configuration", https://core.ac.uk/download/pdf/232196078.pdf

[16] B.M. Meera, S. Krishnamurthy, and K. Manjunath, "Institutional Repository Know-how of a decade in Managing Digital Assets", *Eur. Phys. J. Conf.*, vol. 186, p. 07001. Jan. 2018. doi:10.1051/epjconf/201818607001

[17] Lyrasis, https://registry.lyrasis.org/

[18] University of Pretoria, https://repository.up.ac.za/handle/2263/371

[19] DSpace Lyrasis, https://dspace.lyrasis.org/about/

[20] DSpace-Institutional repository software, https://disa.ukzn.ac.za/sites/default/files/presentations/DisaWorkIR1.pdf

[21] DSpace, https://wiki.lyrasis.org/pages/viewpage.action

[22] Virginia Dons FEDORA: A Prototype for a Digital Object Repository, https://www.dlib.org/dlib/july00/staples/07staples.html

[23] S. Khan, "Dspace or Fedora: Which is a Better Solution?", *SRELS J. Inf. Manage.*, vol. 56, p. 45-50, Feb. 2019. https://journals.indexcopernicus.com/api/file/viewByFileId/519601#:~:text=Clearly%20Dspace%20has%20some%20advantages,installations%20as%20compared%20to%20Fedora

[24] Model for an Open Archival Information System (OAIS), CCSDS 650.0-R-2, http://ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf

[25] Github, https://github.com/DSpace/DSpace

[26] M.J. Bass *et al.*, "DSpace–A sustainable solution for institutional digital asset services–spanning the information asset value chain: ingest, manage, preserve, disseminate", https://web.mit.edu/dspace/live/implementation/design_documents/architecture.pdf