An Institutional Research Data Repository and Digital Object Identifiers for SARAO Radio Astronomy, Fundamental Astronomy, and Geodesy Datasets

Glenda Coetzer\textsuperscript{1}, Roelf Botha\textsuperscript{1}, Christopher Schollar\textsuperscript{1}, Kirsten Elger\textsuperscript{2}

\textsuperscript{1}South African Radio Astronomy Observatory (SARAO),
\textsuperscript{2}German Research Centre for Geosciences (GFZ)

Published on: Apr 27, 2022
DOI: 10.3847/25c2cf6b.6ee866c
License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
ABSTRACT

In recent years, researchers, librarians, publishers and funding bodies have come to realise the importance and potential of using Digital Object Identifiers (DOIs) for data in support of Wilkinson’s FAIR principles of data. The DOI system was originally developed to provide persistent linking for citable and traceable referencing to static datasets in scholarly literature. Nowadays, DOIs and other persistent identifiers can also be assigned to dynamic datasets and data products to recognise, acknowledge and reward the originators of the data. Metrics available for data citation allow data providers to demonstrate, justify, motivate and account for the value of the data they have collected. The South African Radio Astronomy Observatory (SARAO) has become interested in using dataset DOIs as a tool to accelerate its data visibility, discovery, usability, usage reporting and acknowledgement. A pilot project for the attribution of DOIs to SARAO’s datasets in radio astronomy, fundamental astronomy and geodesy is currently underway. Objectives of this project are to develop user-friendly systems towards data discovery and visibility. This will ensure usability and acknowledgement via the DOI-linked citation, whilst also providing SARAO with a usage reporting tool. In addition, methods of linking our publications with our datasets are being devised. We present progress made with the pilot project. We also wish to create awareness of the advancement of open data and open science platforms in radio astronomy, fundamental astronomy and geodesy, both locally and internationally, by making use of DOIs as persistent identifiers.

1 Introduction

Data have always been the substrata of scientific progress — without it one cannot test any assertions. The ultra-rapid expanding universe of online digital data holds promise for scientific scrutiny and its integration into new forms for scholarly publishing. Long-term mechanisms have been established for data discovery and retrievability, and due to these mechanisms, new and unforeseen uses of data are being made. Some of these mechanisms are being used as a means to recognise, acknowledge and reward the originators of the data. Delivering traceability and accountability to the scientific community and general public, for whom the data were created, is imperative [1], and persistent identifiers, such as DOIs, assist in identifying and citing scientific data as well as preventing link rot [2]. More recent initiatives, like the Coalition for Publishing Data in the Earth and Space Sciences [3], the Joint Declaration for data citation principles [4] and the Enabling FAIR Data project [5], have significantly improved the
acceptance of data citations in journal articles and motivated journals in requiring the publication of data underlying scientific articles.

2 SARAO’s Data-rich Environment

SARAO’s scientific instruments and techniques generate extensive amounts of various types of scientific data (Figure 1), including static (e.g. raw, pre-processed and reprocessed data) and dynamic (e.g. expanding) datasets (e.g. time series of rapid/ultra-rapid or final data products. Data are produced with an hourly, sub-daily, daily, weekly, monthly and annual frequency.

Most of the data generated are sent to international Data Correlators (DCs) e.g. MIT Haystack (Westford, Massachusetts. U.S.A.) and Astro/Geo Correlator at MPIfR (Bonn, Germany) and Analysis Centres (ACs) e.g. Federal Agency for Cartography and Geodesy BKG. (Leipzig, Germany.) and Goddard Space Flight Center (Greenbelt, Maryland. U.S.A.) (Figure 2).
Data Centres (DCs) and ACs, provide data and products to the scientific community and the general public under open licences. However, some HartRAO data are stored locally by the observatory (e.g. single-dish observations and some geodesy data) and used by SARAO’s researchers.

3 Digital Object Identifiers — the Magic Tool

Similar to uniquely identifying a published online article, Digital Object Identifiers (DOIs) for datasets were originally developed as a tool for providing permanent identification, access and citable and traceable reference to (static) datasets described in scholarly literature. Today, DOIs are also assigned to dynamic datasets, products (derived from the data), equipment, instruments, ground-based stations, institutions.
and networks (Figure 3) – given that the general rules for DOI-referenced data, i.e. their long-term archiving and accessibility, are not violated.

4 FAIR data

The development of criteria and guidelines to make data FAIR is a long-term international commitment and process [6], [7]. An example of an interesting outcome are the FAIR Data Object Assessment Metrics [8], that describe several levels of ‘FAIRness’ (Table 1).

Table 1. FAIRness assessment metrics [8]

<table>
<thead>
<tr>
<th>Findable</th>
<th>Accessible</th>
<th>Interoperable</th>
<th>Re-usable</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3. DOIs — the persistent linking tool, now also in use by SARAO.
Institutional Research Data Repository and DOIs for SARAO

The development of the RDR began in 2010 with the design of a Geodetic Research Data Management System (GRDMS) [9]. Following the merger between HartRAO and the SKA SA project, the project was adapted and expanded to cater for all of SARAO’s research data management needs (Figure 4).

5.1 Institutional RDR development

A pilot project for developing an institutional Research Data Repository (RDR) and DOI minting service for SARAO’s scientific data and data products was initiated based on the FAIR data principles. Development of the RDR began in 2010 with the design of a Geodetic Research Data Management System (GRDMS). After the merger between HartRAO and the SKA SA, the RDR development project was adapted and expanded to cater for all of SARAO’s research data management needs (Figure 4).

Project objectives include the development of user-friendly systems with a view towards the FAIR data principles and to increase data usage. This will also enable
tracking and acknowledgement via citations with DOI, whilst also providing SARAO with a usage reporting tool.

To address the research problem, ‘Is SARAO’s LIS able to design a system that existing and future unknown users would be able to use?’ a case study was conducted to determine the data management needs of SARAO. Discussions were held with stakeholders (e.g. scientific staff/users). SARAO’s science teams assisted with identifying data structures to be incorporated in the design of the repository. An inventory of data types was conducted and metadata were collected.

![Figure 4. Conceptual model of adapted GRDMS](image)

*DuraSpace (DSpace)* open source software was used to construct the RDR. A prototype of different ‘Communities’ and ‘Collections’, as per typical *DSpace* functionality, was created for all identified data types. Hierarchical structure access paths were created for the data types. The design of a graphical user interface and portal is ongoing.

### 5.2 DOI Minting

In addition to typical concerns of project management and software development, several other aspects had to be considered in planning to mint DOIs for SARAO’s data and products. The local situation at SARAO raised further questions - in particular, more than one division (e.g. Science, Engineering, Business Strategies, etc.) at SARAO
are interested in minting DOIs. It was therefore decided that SARAO’s Library and Information Services (LIS) and Information Technology (IT) will adopt the role of minting DOIs in the interim. Discussions with *DataCite* were initiated in 2020, followed by a licence agreement towards membership of DataCite. SARAO’s first DOI (https://doi.org/10.48479/I1db-b763) was minted on the 19th of February 2021.

### 5.3 Going forward

Some additional preparation is still required before the full implementation of the RDR, e.g. a simple “Cite this dataset” for the SARAO DOI service landing page has to be designed. This feature allows users to copy-paste the pre-generated reference (via a citation formatting service), assisting in citing the resource/dataset and guaranteeing inclusion of DOIs. There are also some matters to resolve that remain, e.g.:

- Who will be responsible for DOIs in each SARAO division in future?
- If LIS maintains its current role, how will it deal with diverse technologies (e.g. different databases) and different needs (e.g. diversified landing page appearances) of divisions?
- How will LIS translate each division’s metadata - describing different kinds of data - into a common format?
- Should SARAO consider using name spacing and extensibility (e.g. starting suffix of SARAO DOIs with namespace) for next-consecutive-integer DOI naming?
- Should `<meta>`- tags with Dublin Core attributes be considered for landing pages?
- Should SARAO DOI landing pages contain JSON-LD, which enhances search engine discoverability?

### 6 NEW: GGOS DOI Working Group for geodetic data

In 2019, the International Association of Geodesy’s (IAG) Global Geodetic Observing System (GGOS) established the first GGOS DOI Working Group (WG). The WG comprises in excess of twenty international members (including SARAO) from all IAG services and relevant members. The WG is designated to establish best practices and advocate for the consistent implementation of DOIs across all IAG Services and in the greater geodetic community as follows:

- Data providers can demonstrate the value of the data collected and analysed by institutions and individual scientists through the use of DOIs.
- DOIs provide a structured and well-documented mechanism which will enable citability, scientific recognition and reward (Elger et al. 2020).
Assessment of DOI minting strategies already implemented by the scientific community were conducted. Ongoing WG discussions include:

- Identification of data products and DOI minting strategies for geodetic data – static, dynamic and observational data, reprocessing products, networks, satellite data, etc.
- Recommendations for data licencing
- Granularity of DOIs (for stations, networks, ongoing time series, etc.)
- Discovery metadata standards – DataCite, ISO 19115, etc.
- Community metadata standards – IGS station logs, GeodesyML, etc. – how to harmonise them with the DOI metadata?
- Data formats – mostly community standards (RINEX, ICGEM/ISG formats, etc.)
- Learning from other communities (DOIs for seismic networks, astronomy data, etc.)

Future discussions will continue to explore metadata standards (e.g. GeodesyML) and the possibility of including PIDs, such as ORCID for researchers, Research Organisation Registry (ROR) for institutions and other DOI-related discovery metadata.

7 Conclusion

Precise identification of data allows observatories to better link to and track data, and related resources, enabling insight into how communities are accessing and using their data. The use of DOIs in original research to identify datasets allows peer reviewers, journal editors and funding agencies to more easily validate research methods, verify results and give credit to whom credit is due. The aim of SARAO’s pilot project for establishing an institutional RDR and DOI minting service is to ensure usability, citability, referencing and acknowledgement of its data and products via recognised mechanisms. To stay abreast of developments in the use of DOIs in complementary science disciplines, SARAO joined the first GGOS DOI Working Group for geodetic data, established in 2019. Knowledge gained from participating in this WG will be applied in continued development of SARAO’s RDR and data management services in years to come.

Acknowledgements

The authors would like to thank Aletha de Witt, Operations Astronomer at SARAO/HartRAO, and her team for their assistance with providing metadata for the different data types and information regarding the structuring of the data. We wish to also thank Khutso Ngoasheng and Amy Leigh Bowers of SARAO for their administrative and financial assistance.
Citations


4. Data Citation Synthesis Group. 2014. Joint Declaration of Data Citation Principles. Force11. https://doi.org/10.25490/A97F-EGYK


