Research Data Citation for the IVS: Data Repositories and Persistent Identifiers

Glenda Coetzer[1,2], Roelf Botha[1,2]

Abstract Data producers often do not receive sufficient credit for their work. A solution is provided via data citation, with data repositories and persistent identifiers fulfilling an important role towards this—it supports Wilkinson's FAIR data principles and has rapidly become a key practice in research and scholarly publication. The IVS data providers also experience an increasing requirement to provide data usage statistics to operating institutions and funders. Data citation principles are summarized and recommendations for the implementation of these principles for IVS-generated datasets and products are presented.

Keywords Data citation, Data repository, PID, DOI, FAIR data, research products

1 Introduction

Bibliographic citation was formalized in scholarly publishing more than a century ago (Stathis *et al.*, 2023), and the theory and application thereof have rapidly become a key practice in supporting Wilkinson's FAIR data principles (Coetzer *et al.*, 2024). Similar to citation of other resources, the citation of data is part of good research practice and the scholarly ecosystem (IVS, 2022); therefore reliable and reproducible research output depends upon a strong foundation of robust, accessible, and citable data. Data citation gives credit and attribution to the creator, encourages sharing, collaboration, and re-use, enables verification of research results, and allows for tracking usage and impact.

In 2014, the Joint Declaration of Data Citation Principles (JDDCP), endorsed by scholarly and academic organizations, articulated a consensus of data citation principles, focusing on purpose, function, and attributes (Stathis *et al.*, 2023):

- **Importance:** Data should be considered legitimate, citable products towards or resulting from research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.
- **Credit and attribution:** Data citations should facilitate in providing scholarly credit and normative and legal attribution to all contributors of the data, recognizing that a single style or mechanism of attribution may not be applicable to all data.
- **Evidence:** In scholarly literature, whenever and wherever a claim relies upon data, the corresponding data should be cited.
- **Unique identification:** A data citation should include a persistent method for identification which is machine actionable, globally unique, and widely used by a community.
- **Access:** Data citations should facilitate access to the data and associated metadata, documentation, code, and other materials required for both humans and machines to make informed use of the referenced data.
- **Persistence:** Unique Persistent IDentifiers (PIDs) as well as metadata describing the data and its disposition, should exist even beyond the lifespan of the data described.
- **Specificity and verifiability:** Data citations should facilitate identification of, access to, and verification of the specific data supporting a claim. Cita-

1. South African Radio Astronomy Observatory (Hartebeesthoek site)
2. University of Pretoria, South Africa

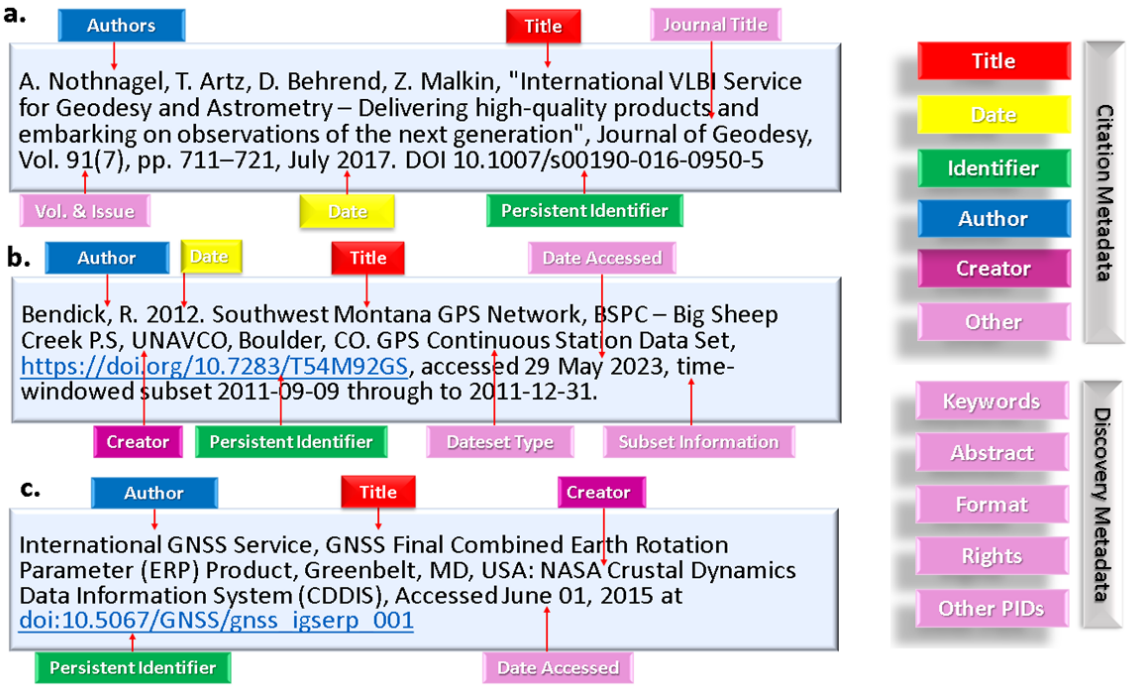

Fig. 1 Examples of citations for (a) a published article, (b) data, and (c) product as well as accompanying metadata.

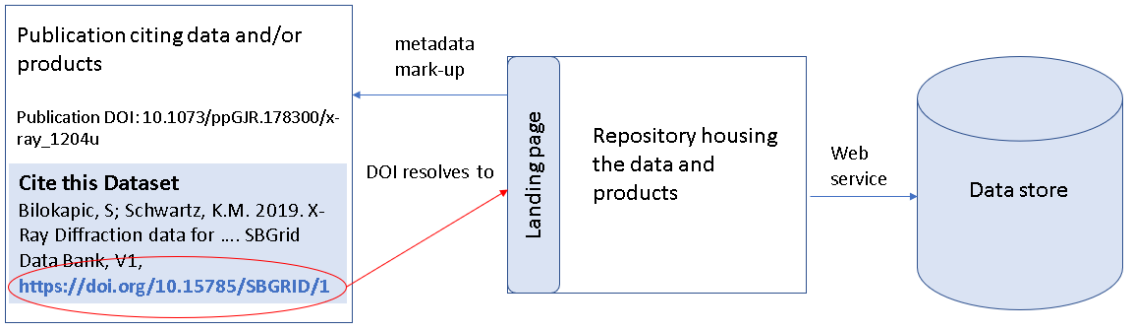

Fig. 2 Genetic data citation using a DOI (adapted from Altman, Borgman, and Crosas, 2015).

tion metadata should include sufficient information about provenance and fixity to be able to verify that the specific time slice, version, and/or granular portion of data retrieved subsequently, is the same as was originally cited.

- **Interoperability and flexibility:** Data citation methods should be sufficiently flexible to accommodate the variant practices among communities, but should not differ so much that interoperability of data citation practices across communities is compromised.

Data repositories play a vital role in data citation, as they:

- enable data stewardship and discovery services to find and give access to the data;
- can incorporate PIDs (e.g. Digital Object Identifiers (DOIs) for publications and data);
- provide for metadata required for data citation (Stathis *et al.*, 2023).

Therefore, repository role players need to work closely with a variety of stakeholders, including data

Table 1 Guidelines for the implementation of the JDDCP data citation principles (Fenner *et al.*, 2019).

Level	JDDCP Nr	Guideline
Required	1	All datasets intended for citation *must* have a globally unique persistent identifier that can be expressed as an unambiguous URL.
	2	Where appropriate, PIDs for datasets *must* support multiple levels of granularity.
	3	The PIDs expressed as an URL *must* resolve to a landing page specific for that dataset, and that landing page must contain metadata describing the dataset.
	4	The PID *must* be embedded in the landing page in machine-readable format.
	5	The repository *must* provide documentation and support for data citation.
Recommended	6	The landing page *should* include metadata required for citation, and ideally also metadata facilitating discovery, in human- and machine-readable format.
	7	The machine-readable metadata *should* use schema.org markup in JSON-LD format.
	8	Metadata *should* be made available via HTML *meta* tags to facilitate use by reference managers.
	9	The persistent identifier *should* be embedded in the landing page in machine-readable format.
Optional	10	Content negotiation for schema.org/JSON-LD and other content types *may* be supported so that the PID expressed as URL resolves directly to machine-readable metadata.
	11	HTTP link headers *may* be supported to advertise content negotiation options.

service providers, publishers, end-users, researchers and, of course, librarians.

During the research lifecycle, metadata (bibliographic information) are created or collected continuously, either manually or automatically. For metadata to be useful, it requires correct temporal and spatial extent, relevance, understandability, sufficiently descriptive keywords (standardized), and contain basic elements (Le Roux *et al.*, 2018). Metadata can be organized and structured according to different standards, such as Dublin Core, DataCite, and the International Organization for Standardization (ISO) (Manu and Bhakti, 2019; ISO, 2019). However, in some scientific disciplines there are still not agreement on the appropriate metadata schema, nor are there uniformity in the application of metadata elements. The lack of data citation standards has therefore resulted in data citations that are highly variable, as is apparent from Figure 1.

Many scientific publications today feature a '*Cite this Dataset*' or a '*Data Availability*' statement and link which is in effect a citation. Citations are useful in that they generally feature metadata and PIDs, which are ideal tools for linking users to papers, papers to data and data to the originators of the data, as illustrated in Figure 2. A PID, such as a DOI, should resolve to the data custodian's landing page, from where users can gain access to the data. Unfortunately, some organizations and institutions mostly still apply URLs for linking users to data, introducing, for instance, the possibility of link rot when there are changes in domains.

2 Recommendations

We present recommendations to the IVS data services for the implementation of the JDDCP and PIDs to the IVS community, specifically for alignment or uptake. The guidelines for IVS data repositories are summarized in groups of required, recommended, and optional in Table 1.

The following are the recommended concepts to consider:

- **Globally Unique PIDs:** A data citation must include metadata describing the data as well as a persistent method for identification that is machine actionable beyond the data's lifespan (JDDCP, principle 6), globally unique and widely used by a community (JDDCP, principle 4) and, ideally, well-organized according to standard metadata formats. The use of the PID should follow community best practices.
- **Granularity:** PIDs must be understood, resolvable, and machine-readable and must support multi-level granularity (e.g., versioning and grouping/clustering of items).
- **Landing pages:** HTTP URL PIDs must resolve to a specific landing page and not resolve to the data itself (i.e., only link to the data source), for single entries, sets, entire repositories, or curated databases (Altman, Borgman, and Crosas, 2015). Ideally, landing pages should provide metadata with additional information, such as how to cite the dataset, licensing as well as persistence policies (Figure 3). The PID can be found on the

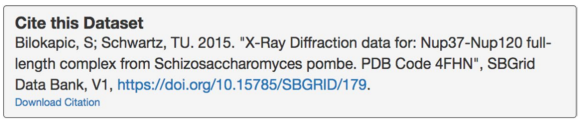

Fig. 3 Example of a citation for a dataset with a download/resolvable link (in BibTex or other standard bibliographic reference manager format).

Fig. 4 Example of a PID embedded in the metadata record.

landing page, but it should ideally be embedded in schema.org markup and/or using HTML meta tags (Figure 4).

- **Documentation and author support:** Documentation should follow the community recommendations, such as how data should be cited, how metadata can be obtained, and who to contact for more information, and should also address the specifics of that particular data repository.

Table 2 Citation metadata elements for data repositories. Key: * name of ID field depends on schema.org serialization format; ** not all datasets will have 'the main researchers involved in producing the data', in which case the more generic 'An entity primarily responsible for making the resource' is recommended, and this can also be an organization.

Citation Metadata	Dublin Core	Schema .org	DataCite	DATS
Dataset Identifier	identifier	@id*	identifier	identifier
Title	title	name	title	title
Creator**	creator	author	creator	creator
Data repository or archive	publisher	publisher	publisher	publisher
Publication Date	date	date-Published	publication-Year	date
Version	*not available*	version	version	version
Type	type	type	resource-TypeGeneral	type

- **Metadata on landing pages:** Landing pages should provide metadata required for data citation in both human- and machine-readable format, and should be accessible without requiring authentication. Ideally it should provide formatting in one or more citation styles common to the community in a '*Cite this Dataset*' field as well as providing means of copying/downloading the citation as text. The landing page should also show all versions, or link to a page with version information. The metadata elements required for data citation are displayed in Table 2.

In addition to the metadata required for citation, it is recommended that additional metadata, again in human-readable and machine-readable formats, be provided on landing pages to assist with data discovery, as indicated in Table 3.

Table 3 Discovery metadata for data repositories. Key: *Related datasets can have part/whole relations (IsPartOf, etc.), version relations (IsVersionOf, etc.) or reference relations (references); ** Related publications reference a dataset published previously, reference a dataset published in parallel with the publication, or otherwise document a dataset.

Discovery Metadata	Dublin Core	Schema .org	DataCite	DATS
Description	description	description	description	dataType dimension Material...
Keywords	subject	keywords	subject	keywords
Licence	licence	licence	rights	licence
Related Dataset*	isPartOf-isVersionOf references	isPartOf citation	related-Identifier	isPartOf
Related Publication**	bibliograph-icCitation	citation	related-Identifier	publication

3 Conclusions

The application of data citation principles (e.g., JDDCP), the attribution of PIDs (e.g., DOIs) for data and product citations, and the use of descriptive metadata have broad impact on research, scholarly publishing, and recognition. Data citation is further supported by tools such as data and metadata repositories. Researchers should cite data when communicating their scholarly or scientific findings in the same manner they cite articles, books, and other sources. It is imperative that the IVS establish structured and well-documented mechanisms for geodesy data, products, software, equipment/instruments, networks and

stations, which can assure discovery, retrieval, and citation of data used in scientific publication, and to give recognition to individuals, institutions, and funders for the creation and storage of data and products.

References

1. Stathis, K., Puebla, I. and Hirsch, M. 2023. DataCite Training: Citations in DataCite Metadata. http://www.datacite.org
2. International VLBI Services for Geodesy and Astrometry. 2022. IVS data and products. https://ivscc.gsfc.nasa.gov/products-data/products.html
3. Coetzer, G., Botha, R.C., van Deventer, M.J., Ball, L.H., Bothma, T.J.D. 2024. Metadata Infrastructure for the South African Radio Astronomy Observatory Data and Products: A Conceptual Framework. CODATA Data Science Journal. [in review]
4. Fenner, M., Crosas, M., Grethe, J.S., Kennedy, D., Hermjakob, H. 2019. https://www.nature.com/articles/s41597-019-0031-8
5. Le Roux, J., Bugbee, K. Dixon, V., Sisco, A., Staton, P., Garcia-Solera, I., Ramachandran, R. 2018. A metadata curation approach to improve the discoverability and accessibility of NASA Earth Science data. https://impact.earthdata.nasa.gov/pubs/ARC_AGU_Fall18_Presentation.pdf
6. Altman, C., Borgman, C., Crosas, M. 2015. An introduction to the joint principles for data citation. https://scholar.harvard.edu/files/mercecrosas/files/febmar15_rdap_altman_etal.pdf
7. Manu, T.R., Bhakti, G. 2019. Research Data Management lifecycle: an overview. In the Trends, Challenges and Opportunities for LIS Education and Practice (Festschrift Volume in Honour of Prof. Muttayya M. Koganuramath)