

## PRACTICE PAPER

# Identifiers for Earth Science Data Sets: Where We Have Been and Where We Need to Go

Justin C. Goldstein<sup>1,2</sup>, Matthew S. Mayernik<sup>3</sup> and Hampapuram K. Ramapriyan<sup>4,5</sup><sup>1</sup> US Global Change Research Program, Washington, DC, US<sup>2</sup> ICF, Fairfax, Virginia, US<sup>3</sup> National Center for Atmospheric Research, University Corporation for Atmospheric Research, Boulder, Colorado, US<sup>4</sup> Science Systems and Applications, Inc., Lanham, Maryland, US<sup>5</sup> NASA-Goddard Space Flight Center, Greenbelt, Maryland, USCorresponding author: Justin C. Goldstein ([justin.goldstein@noaa.gov](mailto:justin.goldstein@noaa.gov))

Considerable attention has been devoted to the use of persistent identifiers for assets of interest to scientific and other communities alike over the last two decades. Among persistent identifiers, Digital Object Identifiers (DOIs) stand out quite prominently, with approximately 133 million DOIs assigned to various objects as of February 2017. While the assignment of DOIs to objects such as scientific publications has been in place for many years, their assignment to Earth science data sets is more recent. Applying persistent identifiers to data sets enables improved tracking of their use and reuse, facilitates the crediting of data producers, and aids reproducibility through associating research with the exact data set(s) used. Maintaining provenance – i.e., tracing back lineage of significant scientific conclusions to the entities (data sets, algorithms, instruments, satellites, etc.) that lead to the conclusions, would be prohibitive without persistent identifiers. This paper provides a brief background on the use of persistent identifiers in general within the US, and DOIs more specifically. We examine their recent use for Earth science data sets, and outline successes and some remaining challenges. Among the challenges, for example, is the ability to conveniently and consistently obtain data citation statistics using the DOIs assigned by organizations that manage data sets.

**Keywords:** Digital Object Identifiers (DOIs); Persistent Identifiers; Data Citation; DataCite; Google Scholar; Provenance

## Introduction

Data curation and citation practices in the Earth sciences have emphasized assignment and citation of identifiers in order to promote asset discoverability and interoperability (Klump, Huber & Diepenbroek 2016). Persistent identifiers undoubtedly contribute to the popularity of data citation through enabling improved tracking of data set use and reuse, providing credit for data producers, and aiding reproducibility efforts through associating research with the exact data set(s) used (Parsons and Fox 2013; Parsons 2014; Katz and Strasser 2015). By directly linking publications with the resources underlying the scientific findings reported therein, persistent identifiers ensure scientific integrity, promote data discovery and management, and facilitate scientific communication (Hanson 2016). While the recognition of the benefits of such linking is not new, the systematic implementation of linking resources is not yet widespread. Some recent examples of implementation include the Global Change Information System which linked findings from the Third US National Climate Assessment with their underlying data (e.g., Ma *et al.* 2014; Wolfe *et al.* 2015) and the tracing of indicators found within Integrated Ecosystem Assessments (Beaulieu *et al.* 2016). In geology, persistent identifiers are core elements of digital metadata records, and are especially useful when space or resource requirements complicate storage of the samples described therein (McNutt *et al.* 2016). Persistent identifiers are essential for maintaining provenance: tracing back lineage of significant

scientific conclusions to the use of the entities (data sets, algorithms, instruments, satellites, etc.) that lead to conclusions (Ramapriyan *et al.* 2016). As stated by Tilmes, Yesha & Halem (2010):

'Provenance in this context refers to the source of data and a record of the process that led to its current state. It encompasses the documentation of a variety of artifacts related to particular data. Provenance is important for understanding and using scientific data sets, and critical for independent confirmation of scientific results'.

Unless each of the entities constituting the provenance has a persistent identifier, the provenance trace generated at a given time may not remain valid at a future time. Many professional societies, journals, and federal entities within the US have recently adopted commitments, recommendations, mandates, and procedures for citing research data used in research products, all of which rely on identifiers (for particular examples within the field of Earth sciences see Bloom, Ganley & Winkler 2014; GSA 2014; Hanson and van der Hilst 2014; Evans *et al.* 2015; Hanson, Lehnert & Cutcher-Gershenfeld 2015; Mayernik, Ramamurthy & Rauber 2015). The US Federal Open Data Policy mandated the accompaniment of appropriate citations and persistent identifiers with data sets resulting from federally-funded research (OMB 2013). Consequently, US federal agencies have begun implementing policies requiring, enabling and facilitating data citations – e.g., the National Aeronautics and Space Administration (NASA 2015), the National Oceanic and Atmospheric Administration (NOAA 2015), the National Science Foundation (NSF 2014), and the United States Geological Survey (USGS 2016). Having over 180 participating member organizations including such federal agencies, universities and commercial entities, the Federation of Earth Science Information Partners (ESIP) has developed data citation guidelines (ESIP Data Stewardship Committee 2012) which have been adopted by various parties.

Outside the US, the UK Digital Curation Centre (Ball and Duke 2015) has provided a detailed guide on citing data sets and linking them to publications. Egloff *et al.* (2016) have emphasized data citations as important components of their data policy recommendations for the European Biodiversity Observation Network (EU BON). During 2013–2014, the Committee on Data for Science and Technology (CODATA) and the International Council for Scientific and Technical Information (ICSTI) held several international workshops to articulate data citation principles (CODATA-ICSTI 2015). These produced the Joint Declaration of Data Citation Principles (Data Citation Synthesis Group 2014), which explicitly calls out the need for unique and persistent identifiers for data sets. To date, these principles have been endorsed by 372 entities worldwide, 114 of which are science organizations including journal publishers and professional societies (FORCE11 2016). The use of persistent identifiers in citations for non-data set assets such as software (e.g., Gent, Jones & Matthews 2015; Smith *et al.* 2016) and projects (e.g., NCAR 2016) is an emerging area of focus, and builds on the success of data citation efforts.

The Digital Object Identifier (DOI) is by far the most widely used identifier system-with 130 million persistent identifiers assigned to date (International DOI Foundation 2016a). Administered by the International DOI Foundation (IDF), it is an integral part of scientific publishing and possesses a high level of maturity (Klump, Huber & Diepenbroek 2016). Recently, many journal-wide open access publishing endeavors, e.g., the Coalition for Publishing Data in the Earth and Space Sciences (COPDESS 2015) and the Transparency and Openness Promotion (McNutt 2016) have selected the DOI as their identifier of choice. DOIs serve as a key requirement for items in DataCite, the primary DOI registration agent for data DOIs. Given the recent emphasis on DOIs for citing data sets, this paper explores the adoption of DOIs for Earth science data sets, outlines successes, and identifies some remaining challenges.

DOIs comprise a portion of the Handle System, but exceed its capabilities by providing a resolution system for identifiers and for requiring semantic interoperability, among other reasons (International DOI Foundation 2015). Various articles have further explored the distinctions between DOIs and other identifier schemes, and have shown the advantages of DOIs over the others. DOIs possess advantages compared to other identifier schemes. For example, Duerr *et al.* (2011) have compared DOIs with eight other identifier schemes for their use as Unique Identifiers, Unique Locators, Citable Locators, and Scientifically Unique Identifiers. Although they found none of the schemes to be suitable as Scientifically Unique Identifiers (which inform users that two data instances—even if in different formats—contain identical information), they found that DOIs fare equally well or better than others for the first three uses, and especially well as citable locators for data sets. Lee and Stvilia (2012) have compared DOIs with five other identifier schemes using the following 11 quality criteria: Uniqueness/Precision, Resolution, Interoperability, Persistence, Granularity/Flexibility, Complexity, Verifiability, Opacity/Clarity, Authority, Scalability, and Security. They

observed that the Handle system was the most widely used in 34 institutional data repositories of (the then) 61 members of the prestigious Association of American Universities. However, they have noted that DOIs ranked the highest in all the quality criteria except Opacity/Clarity. They also stated that a possible reason for low adoption of DOIs by the repositories could be ‘the cost of registering with and purchasing DOI prefixes from the International DOI Foundation’. The DOI syntax is described in ISO 26324:2012. It is important to emphasize that while the number of DOIs assigned is in the millions, the significantly lower numbers shown below for Earth science data sets is solely due to their relatively recent use for identifying and citing them.

For nearly 20 years, publications discussing the need for identifiers for data sets—many with particular emphasis on DOIs—have appeared in the literature (e.g., Helly *et al.* 1999; Helly, Staudigel, & Koppers 2003; Brase 2004; Paskin 2006; Klump *et al.* 2006; Williams *et al.* 2009; Piwowar 2011; Hills *et al.* 2015; James and Wanchoo 2015; Christensen *et al.* 2015; Mayernik, Phillips & Nienhouse 2016; Prakash *et al.* 2016). Although material discussing perceived benefits and drawbacks to various identifier schemes exists (see above), to our knowledge none to date has focused on evaluating DOI use. To what extent have DOIs accomplished their aims? Have game changers occurred in the field of data citation which hinder DOI use? This paper explores these questions with particular focus on Earth science data sets held by organizations in the US.

### DOI Use and Proliferation

As mentioned above, DOIs have proliferated in many domains since their advantages were first heralded during the late 1990s (see Rosenblatt 1997; Davidson and Douglas 1998). This section covers the recent growth of DOI assignments for Earth science data.

US federal agencies and affiliates have within the past few years initiated earnest efforts to assign DOIs to their data sets (see **Table 1**). In the cases of NASA, NOAA, and USGS the small to moderate percentages of data

Organization	Number of DOIs Assigned (Data Sets)	Total Number of Data Sets in Archive	%	Initial Year of DOI Assignment	Source	Valid Date
US Department of Energy (DOE) Atmospheric Radiation Measurement (ARM) Archive	797	41602		2012	ARM Climate Facility Data Discovery Search Results (2017); Prakash, pers. communication	2017-02-03
NASA Earth Observing System Data and Information System Distributed Active Archive Centers (EOSDIS DAACs)	5364	~ 10000	~ 54	2009	NASA 2017	2017-02-08
NOAA	612*	n/a*	n/a*	2013	DataCite 2017; De La Beaujardière, pers. communication	2017-02-08
National Center for Atmospheric Research (NCAR) Research Data Archive**	79	671	12	2012	UCAR NCAR 2017	2017-02-09
Rolling Deck to Repository (R2R)***	15405	20956	74	2015	Arko, pers. communication	2017-02-01
USGS Science Data Catalog	1407	8677	16	2011	Bristol, pers. Communication	2017-02-13

**Table 1:** DOI Assignment to Data Set by Selected Federal Agencies and Affiliates.

\*This is the number of DOIs minted by NOAA for datasets created by NOAA and archived at the NOAA Centers for Environmental Information (NCEI). NCEI archives also contain many other datasets for which the information about DOI assignments is not available. Therefore it is not meaningful to compute the percentage for comparison.

\*\*NSF-funded entity.

\*\*\*Sponsored by NSF, NOAA, Oak Ridge National Laboratory (ORNL), and the Schmidt Ocean Institute (SOI).

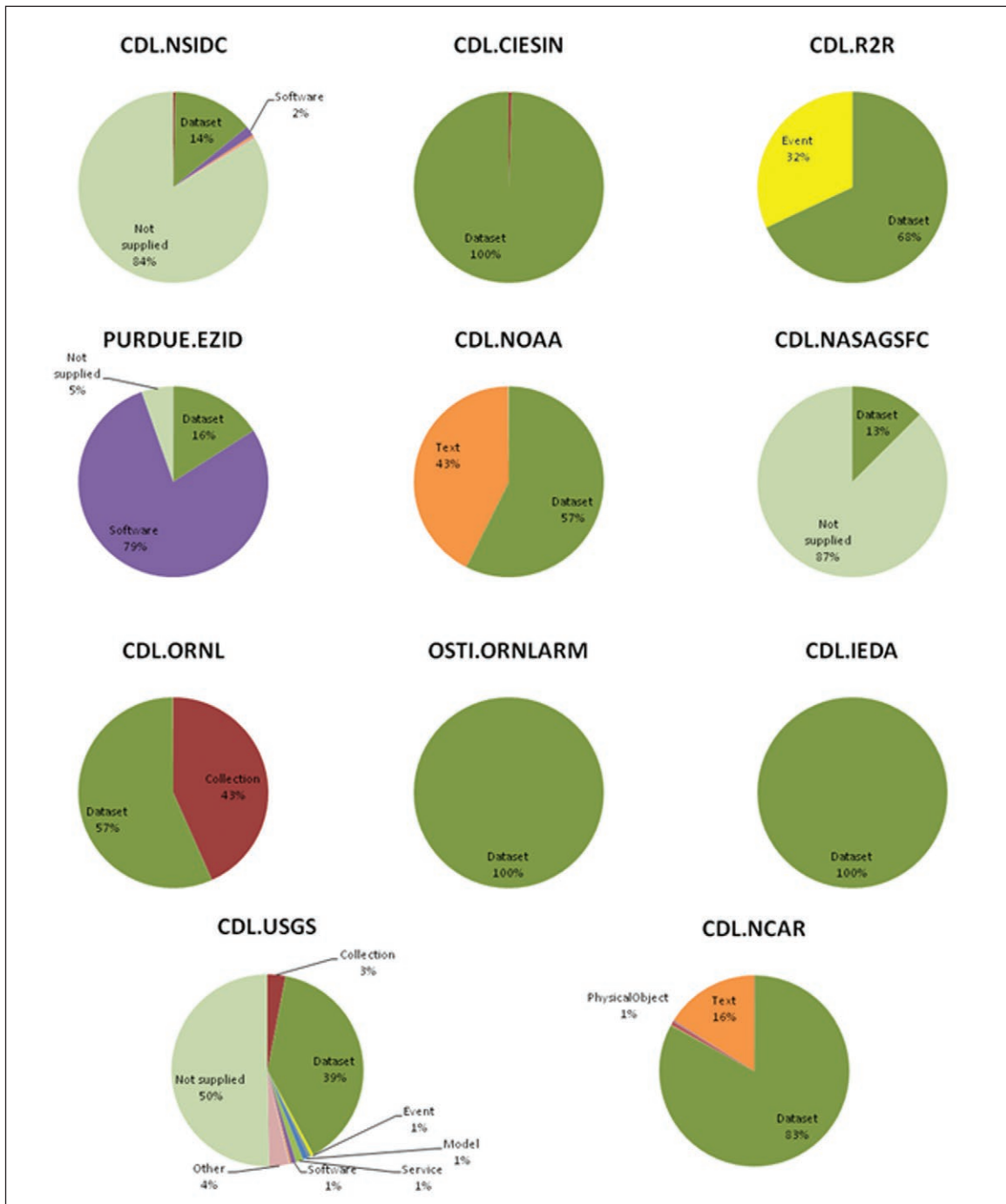
sets assigned DOIs is a function of (1) the vast size of their catalogues, and that (2) in these and other various federal entities, DOIs are minted upon the submission of the data sets to the archive and their preparation for public distribution. Note that such programs have only recently begun assigning DOIs for data sets. Looking in figshare, an independent service that supports self-archiving of data and other assets, 132,574 data sets out of 466,545 total assets (28%) were categorized as related to “Earth and Environmental Sciences”, as of January 25, 2017, (Valen, pers. communication). The DataCite organization has been influential in promoting and enabling geoscience-focused organizations to create DOIs for data sets (Klump, Huber, and Diepenbroek 2016). All organizations mentioned in **Table 1** and **2**, as well as **Figure 1**, now register DOIs for data using DataCite services. NOAA and NASA assign DOIs deliberately at the coarse, collection level rather than at the finer, “granule” level. This is to avoid a large proliferation of DOIs and to simplify citation. For example, NASA EOSDIS curates hundreds of millions of granules, compared to approximately 10,000 collections. After careful review of data sets and metadata, NOAA assigns DOIs manually, one at a time. The DOIs were initially assigned manually within NASA EOSDIS, although the recent automation of the process has resulted in significantly faster assignments (Wanchoo, James & Ramapriyan 2017). In both cases the organizations ensure that well-developed landing pages providing information about data sets exist before DOIs are assigned.

As DOIs for data sets from the above organizations have existed for a few years, it would be interesting to identify the extent to which they have been used in citations. Using information gleaned from Google Scholar (2017), **Table 2** presents the numbers of citations to objects (mostly data sets) using some DOI prefixes of interest. It should be noted that the term “include patents” was unchecked while performing the Google Scholar searches. Google Scholar returns lists of publications containing a string specified in the search window. Thus, by specifying a DOI prefix reserved for data sets, it is theoretically possible to obtain lists (and numbers) of publications that cite all data sets using such a prefix. However, the search results are sensitive to the manner in which the DOI prefixes are specified. For example, the five entries 10.7265, DOI:10.7265, doi.org/10.7265, “DOI:10.7265” and “doi.org/10.7265” return different sets of results, with some overlap. The first three entries in the search window provide several irrelevant results in which the number 7265 may have occurred in some context other than a DOI or a citation. The last two, in which the quotes are included, do provide mostly relevant results. However, differences in results exist because authors do not use a consistent format in the citations – some use DOI:10.7265 and others use doi.org/10.7265. Without a detailed analysis of the results from such searches, it is difficult to deduce the extent of overlap between the two, and thus determine the accurate number of citations of data sets with a given DOI prefix. **Table 2** shows results from two search strings for each of the DOI prefixes analyzed. Despite the issues shown above, use of DOIs for citing data sets has the potential for assessing the utility of data sets as the adoption of DOIs continues to take hold. The brackets indicate the data center code as represented in the DataCite system, e.g. CDL.USGS. “CDL” refers to the California Digital Library, which provides DOI registration services via DataCite. The DataCite codes are shown to allow comparison with **Figure 1**.

Although DataCite and related initiatives have emphasized DOI assignment primarily to data sets, they are actually being provided to a much wider set of resources. Using the DataCite (2016) search service (accessed on August 24, 2016), **Figure 1** divides DOI assignments by DataCite-identified resource type for 11 organizations. Those organizations are chosen from the ESIP (2016) member directory. The organizational names are drawn from DataCite. “CDL” refers to the California Digital Library, which provides DOI registration services.

Organization	DOI Prefix (Search string)	Number of Google Scholar results Uncovered Using their Respective Strings
NASA – EOSDIS [CDL.NASAGSFC]	“DOI:10.5067”; “doi.org/10.5067”	346; 324
National Snow & Ice Data Center (NSIDC) [CDL.NSIDC]	“DOI:10.7265”; “doi.org/10.7265”	166; 202
NCAR [CDL.NCAR]	“DOI:10.5065”; “doi.org/10.5065”	927; 415
NOAA [CDL.NOAA]	“DOI:10.7289”; “doi.org/10.7289”	578; 261
USGS [CDL.USGS]	“DOI:10.5066”; “doi.org/10.5066”	112; 178

**Table 2:** Number of Google Scholar Search Results for Resources Maintained in Selected US Data Centers (As of February 8, 2017).



**Figure 1:** DOIs, broken down by specific resource types, created by selected ESIP member organizations.

Resource types displayed in **Figure 1** are those allowed by the DataCite metadata schema, specifically according to the controlled list of types allowed in the “ResourceTypeGeneral” attribute of the “ResourceType” element. As this field was optional in the DataCite schema until 2016, not all their metadata records provide them. Records with unspecified resource types are listed as “Not Supplied”. As with those in CDL.NASAGSFC (DOI prefix 10.5067), in some cases the category “not supplied” mostly contains data sets, although this is hard to determine from the metadata alone. **Figure 1** contains one pie chart for each of the 11 organizations, showing percentages of objects belonging to various resource types to which DOIs have been assigned. Resource types with fewer than 1% of the total number of objects are not shown. The data underlying these charts are provided in the Appendix.

DOIs have proved useful for identifying, locating, and citing data sets, and the extent of their recent adoption by various organizations shows that the practice of citing data sets is catching on in scientific publications. Data set citations also can facilitate the extraction of citation metrics for those from given

organizations, for a given group of data sets or even a single data set, which would otherwise involve considerable amount of manual effort.

### **Some Remaining Challenges with DOIs**

Despite the successes with DOI implementations mentioned above, challenges remain. We organize this summary according to those identified in the literature and those based on our experience.

#### ***Challenges identified in the literature***

As the DOI system was being developed in the late 1990s, there was common sentiment within academic circles that the DOI system was too targeted to the needs of the publishing community, in-lieu of addressing the needs of the scholarly community (Cleveland 1998; Davidson and Douglas 1998; Lynch 1998). For instance, the “Armati Report” (Armati 1995), which was foundational in the development of the DOI system, focused in large part on intellectual property issues, which were (and still are) critical for academic publishers. Even in the early days, however, issues that are problematic for scientific data now were being discussed. For example, in referencing “critical issues” facing an information identification system, that report specifically calls out “flexible, granular (sub-file level) data object identification” and “real time differentiation between master data object and individual expressions of whole or parts of it” (pg. 6). Similarly, in outlining the early developments of the DOI system, Paskin (1999) described debates about the resource types to which DOIs should be assigned, the value of the metadata associated with DOIs, and the granularity at which DOIs should be assigned to various resources. These questions re-emerged when the communal practice of assigning DOIs to data sets began in earnest (Paskin 2006), and continue to inspire debate (Altman and King 2007; Wynholds 2011; Starr *et al.* 2015). Lin and Strasser (2014) have noted that the DOI system (1) lacked mechanisms for users to identify modifications to the data if needed, and (2) was designed to support references to entire data sets, not their components such as individual values or granules. Parsons and Fox (2013) note how many of these challenges in applying DOIs to data relate to the DOI system’s strong association with the concept of “publication”. In the context of scholarly publishing, DOIs are assigned after a resource (e.g. journal article) is fixed in its final state. The DataCite DOI services were built with the same model in mind (Brase, Sens, & Lautenschlager 2015). Data sets often change or are updated however, even after being made publicly accessible. Therefore, the assumptions that people may hold for resources with DOIs – that they are static and well-bounded – are not necessarily true for data.

In a Public Library of Science blog, Fenner (2011) recognized the following issues with the DOI system’s implementation: 1. Many bibliographic databases store DOIs without permitting queries or providing links to services using them; 2. DOI string syntax is often very web-unfriendly, as the unregulated format of DOI strings could require ‘escaping’ various special characters; 3. DOIs assigned to individual components such as individual figures or tables in a paper, while very useful, can ‘confuse bibliographic databases and make it more difficult to track all the links to a given article’; 4. Assignment of new DOIs to updated versions of articles complicates the tracking of all references to the articles. He noted that journals are beginning to handle these issues, however.

During a retrospective discussion of 20 years of web-based identifier management systems, Bide (2015) recognized that some of the key barriers to universally adopting identity management techniques/standards involve governance issues associated with achieving broad adherence to standards, coupled with the fact that use of identification systems is not free of charge (even though the costs may be small). In short, many challenges related to DOI adoption revolve around practices, not technical capabilities.

#### ***Challenges identified by authors’ experience***

We have focused on the use of identifiers in Earth science data. Although this scientific community is slowly recognizing the value of assigning identifiers to data products, data citations have not yet become the norm in scientific publications. It is expected that the gathering of accurate data set citation metrics will be facilitated by the emergence of this norm. However, the lag in adoption of data citations by data users and journals slows the ability to gather such metrics accurately. We have not yet found prevalent web-based mechanisms for consistently and accurately deriving citation metrics for data sets originating from given organizations. Such difficulties are:

- As shown in **Figure 1**, it is very difficult to identify the resource type to which a DOI refers without consulting its metadata. In many cases even from the metadata it is impossible to identify

the resource type. Although very recently remedied to some extent by changing the resource type to a mandatory metadata field, it will take some time for the data archives to comply with this recent requirement.

- No universally accepted recommendations of good practices for DOI syntax (e.g., random strings, clearly identifiable strings) exist. Although stated in the DOI Handbook (International DOI Foundation 2016b), the “best practice” of assigning opaque identifiers is not universally followed. Opaque DOIs refer to those with suffixes having no discernable meaning (e.g., 10.5065/D62J68XR).
- Google Scholar counts the appearance of multiple citations with a single DOI prefix (e.g., 10.5067) within a document as one record. The determination of the true number of citations using such a prefix therefore requires either manual or automated analysis of the search results. This would hold true for outputs from any search engine.
- The number of Google Scholar outputs changes daily, and sometimes decreases, complicating analysis efforts.
- No method exists for distinguishing between “internal citations” that use DOI prefixes for describing identifier assignment processes and “external citations” that cite data sets with those prefixes.
- Among data archives, variations among the definitions of “data sets” and the level of granularity at which they exist complicate overall evaluation of the extent of DOI use. For instance, one archive may consider three similar data sets as one, in contrast to that of another archive.

It is possible that these difficulties will be overcome with specialized software for scripted searches.

Our analysis of the assignment of DOIs to resources (primarily data sets) by various organizations managing Earth science data reveals significant differences in approaches. These differences occur in syntax, number of prefixes used, and the variety of resource types to which DOIs are assigned, among others (see **Figure 1** and **Table 1**). This complicates the use of these DOIs for consistently deriving or aggregating information such as usage metrics. While not problematic for using DOIs to enable persistent and unique references to individual resources, the variation in the manner in which DOIs are used in citations makes it difficult to use DOIs for gathering metrics of data usage and citation. Many data usage metrics, such as download counts and quantities of bytes delivered, are already highly repository specific, being contingent on the manner in which repositories collect, manage, and present data (Weber *et al.* 2013). DOI-based metrics are not likely to be any more comparable across organizations or repositories than other data usage metrics due to the variations in DOI implementation.

## Discussion

As demonstrated above, the utilization of DOIs within the Earth Science community has come a long way since the 1990s. It is worth noting though that much of this work has emerged only recently—as early as the beginning of this decade—owing in part to revolutions in the digital arena. This study only examined DOI metrics for US-based organizations; further investigation is needed to enable comparison with similar initiatives in other geographic areas. However, given the international interest in this topic and multi-national scope of key organizations such as DataCite and the Research Data Alliance, we would expect to see only minor, if any, differences across national borders. Increased DOI utilization will necessitate the development and adaptation of automated approaches to data citation. Along with automated approaches, the use of structured metadata having mandatory “resource type” fields will facilitate improved characterization of the importance of Earth science data sets and related information, with the appropriate linkages to related resources, other types of resources as well as the linking of them to related resources.

Within its data citation guidelines, the ESIP Data Stewardship Committee (2012) provided recommendations for citing so-called “dynamic data” (i.e., data sets containing evolving contents). More recently, the Research Data Alliance (RDA) Working Group on Data Citation has proposed a query-based approach for generating precise citations from archival databases (Rauber *et al.* 2015). Furthermore, Buneman, Davidson & Frew (2016) have developed an automated approach to query-based citation generation which returns the data along with relevant citations from databases. When used with DOIs, such approaches should cite data more precisely, assess provenance more accurately, and improve reproducibility of scientific products. In addition, the increased use of web-based persistent identifier schemes will enable scientific data systems to leverage linked data and Semantic Web technologies more broadly. These technologies, which are being

used by a number of organizations to support systems for data documentation, discovery, and integration, rely on resources having web-resolvable identifiers (Ma *et al.* 2014; Wilson *et al.* 2015). The use of persistent identifiers adds robustness to Semantic Web applications, reducing problems related to link obsolescence.

## Additional File

The additional file for this article can be found as follows:

- **Appendix.** Data underlying Figure 1. DOI: <https://doi.org/10.5334/dsj-2017-023.s1>

## Acknowledgements

Any opinions, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or funders. Goldstein's employment at the US Global Change Research Program National Coordination Office is funded through NASA Contract NNH15CN83C awarded to ICF. Ramapriyan's work is supported by NASA Contract Number NNG15HQ01C awarded to Science Systems and Applications, Inc. The National Center for Atmospheric Research is sponsored by the US National Science Foundation. We thank the following individuals for providing information about their organizations' respective DOI assignments: Robert Arko (R2R), Sky Bristol (USGS), Tom Boden and Terri Killeffer (CDIAC), Jeff de La Beaujardière (NOAA), Giri Prakash (DOE ARM Program), Dan Valen (figshare), and Lalit Wanchoo (Adnet Systems, Inc. / NASA). We also thank the two anonymous reviewers for providing suggestions which improved the quality of the manuscript. This paper is a result of the authors' participation in the Federation of Earth Science Information Partners (ESIP) Data Stewardship Committee.

## Competing Interests

The authors have no competing interests to declare.

## Author Information

Justin Goldstein served as the Advance Science Coordinator at the US Global Change Research Program (USGCRP), a confederation of the Research arms of 13 US federal agencies interested in Global Change research. He was employed by ICF. Goldstein's interests reside in Earth science informatics and applications: primarily data management and analysis, query ability, and semantics, as well as in Earth System Science, hydrology/water cycle research, and remote sensing. From 2014–2016, he co-chaired the Federation of Earth Science Information Partners (ESIP) Data Stewardship Committee (DSC). Goldstein holds a Ph.D. in geography from the University of Oklahoma and BS and MA degrees in geography from the University of Maryland, College Park. At the time of publication, Justin is now affiliated with National Oceanic and Atmospheric Administration (Washington, DC, USA), however this research in this paper was conducted during his tenure at the institutions mentioned above.

Matthew S. Mayernik is a Project Scientist and Research Data Services Specialist within the Library in the National Center for Atmospheric Research (NCAR)/University Corporation for Atmospheric Research (UCAR), located in Boulder, CO. His work is focused on research and service development related to research data curation, including metadata practices and standards, data curation education, data citation and identity, and social and institutional aspects of research data. He completed his Master's in Library and Information Science and Ph.D. in Information Studies at UCLA. As of 2017, he co-chairs the ESIP DSC. As a member of the Board on Data Stewardship within the American Meteorological Society (AMS), he led the writing of a data archiving and citation recommendation that has since been implemented by all AMS journals.

Hampapuram Ramapriyan is a Research Scientist/Subject Matter Expert at Science Systems and Applications, Incorporated (SSAI). He supports the Earth Science Data and Information System (ESDIS) Project at NASA Goddard Space Flight Center through its contract with SSAI. The ESDIS Project is responsible for archiving and distributing most of NASA's Earth science data using the Earth Observing System Data and Information System (EOSDIS). Ramapriyan's primary focus is data stewardship and preservation. Prior to his employment with SSAI, he was the Assistant Project Manager of the ESDIS Project. His responsibilities included management of Science Investigator-led Processing Systems that processed and delivered data to the EOSDIS Distributed Active Archive Centers (DAACs). He has supported the US Global Change Research



Program in the analysis of the Third National Climate Assessment (NCA3) Report for completeness of provenance of images resulting from NASA's data. He has been an active member of the Federation of Earth Science Information Partners (ESIP) since its inception in 1998. He has led development of the emerging Provenance and Context Content Standard and its adaptation to NASA's Earth Science Data Preservation Content Specification. He is currently a member of the ESIP Data Stewardship Committee and chairs the Information Quality Cluster.

## References

- Altman, M and King, G** 2007 A Proposed standard for the scholarly citation of quantitative data. *D-Lib Magazine*, 13(3/4). DOI: <https://doi.org/10.1045/march2007-altman>
- Armati, D** 1995 *Information identification: A Report to STM International Association of Scientific, Technical, and Medical Publishers*. Available at: [https://www.doi.org/topics/Armati\\_Info\\_Identification.pdf](https://www.doi.org/topics/Armati_Info_Identification.pdf) [Last accessed 22 September 2016].
- ARM Climate Facility Data Discovery Search Results** 2017 Available at: <https://www.archive.arm.gov/discovery/#v/results/s/s::> [Last accessed 01 February 2017].
- Ball, A and Duke, M** 2015 How to cite datasets and link to publications. Available at: <https://www.dcc.ac.uk/resources/how-guides/cite-datasets> [Last accessed 22 September 2016].
- Beaulieu, S E, Fox, P A and Di Stefano, M** 2016 Toward cyberinfrastructure to facilitate collaboration and reproducibility for marine integrated ecosystem assessments. *Earth Science Informatics*. DOI: <https://doi.org/10.1007/s12145-016-0280-4>
- Bide, M** 2015 The DOI – Twenty years on. *D-Lib Magazine*, 21(7/8). DOI: <https://doi.org/10.1045/july2015-bide>
- Bloom, T, Ganley, E and Winker, M** 2014 Data access for the open access literature: PLOS's data policy. *PLOS Biology* 12(2): e1001797. DOI: <https://doi.org/10.1371/journal.pbio.1001797>
- Brase, J** 2004 Using digital library techniques – Registration of scientific primary data. *Research and Advanced Technology for Digital Libraries, Lecture Notes in Computer Science* (Vol. 3232, pp. 488–494). Springer Berlin / Heidelberg. DOI: [https://doi.org/10.1007/978-3-540-30230-8\\_44](https://doi.org/10.1007/978-3-540-30230-8_44)
- Brase, J, Sens, I and Lautenschlager, M** 2015 The Tenth anniversary of assigning DOI names to scientific data and a five year history of DataCite. *D-Lib Magazine*, 21(1/2). DOI: <https://doi.org/10.1045/january2015-brase>
- Buneman, P, Davidson, S and Frew, J** 2016 Why data citation is a computational problem. *Communications of the ACM*, 59(9). DOI: <https://doi.org/10.1145/2893181>
- Christensen, T, Albani, M, Mitchell, A, Miura, S, Kudo, Y, Maggio, I and Cosac, R** 2015 Best practices for persistent identifiers in Earth observation archives, *Proceedings for the 2015 PV Conference, 3–5 November 2015, Darmstadt, Germany*. Available at: <https://elib.dlr.de/102110/> [Last accessed 8 February 2017].
- Cleveland, G** 1998 Digital libraries: Definitions, issues and challenges. *International Federation of Library Associations and Institutions, Universal Dataflow and Telecommunications Core Programme, Occasional Paper 8*. Available at: <https://www.ifla.org/archive/udt/op/udtop8/udt-op8.pdf> [Last accessed 1 February 2017].
- CODATA-ICSTI** 2015 International series of workshops to implement data citation principles. Available at: <https://www.codata.org/news/59/62/International-Series-of-Workshops-to-Implement-Data-Citation-Principles> [Last accessed 22 September 2016].
- COPDESS** 2015 COPDESS statement of commitment. Available at: <https://www.copdess.org/statement-of-commitment/> [Last accessed 1 February 2017].
- DataCite** 2016 DataCite metadata search beta. Available at: <https://search.datacite.org/ui> [Last accessed 24 August 2016].
- DataCite** 2017 DataCite metadata search beta. Available at: [https://search.datacite.org/ui?q=10.7289&fq=resourceType\\_facet:%22Dataset%22](https://search.datacite.org/ui?q=10.7289&fq=resourceType_facet:%22Dataset%22) [Last accessed 8 February 2017].
- Data Citation Synthesis Group** 2014 Joint declaration of data citation principles. Martone, M (ed.) San Diego CA: FORCE11. Available at: <https://www.force11.org/group/joint-declaration-data-citation-principles-final> [Last accessed 22 September 2016].
- Davidson, L A and Douglas, K** 1998 Digital object identifiers: Promise and problems for scholarly publishing. *The Journal of Electronic Publishing*, 4(2). DOI: <https://doi.org/10.3998/3336451.0004.203>
- Duerr, R E, Downs, R R, Tilmes, C, Barkstrom, B, Lenhardt, W C, Glassy, J, Bermudez, L E and Slaughter, P** 2011 On the utility of identification schemes for digital Earth science data: an assessment

- and recommendations. *Earth Science Informatics*, 4: 139–160. DOI: <https://doi.org/10.1007/s12145-011-0083-6>
- Egloff, W, Agosti, D, Patterson, D, Hoffmann, A, Mietchen, D, Kishor, P and Penev, L** 2016 Data policy recommendations for biodiversity data. EU BON Project Report. *Research Ideas and Outcomes*, 2: e8458. DOI: <https://doi.org/10.3897/rio.2.e8458>
- ESIP** 2012 Interagency Data Stewardship/Citations/provider guidelines. Available at: [https://wiki.esipfed.org/index.php/Interagency\\_Data\\_Stewardship/Citations/provider\\_guidelines](https://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines) [Last accessed 22 September 2016].
- ESIP** 2016 ESIP Partners. Available at: <https://esipfed.org/esip-member-list> [Last accessed 05 August 2016].
- ESIP Data Stewardship Committee** 2012 Data citation guidelines for data providers and archives. DOI: <https://doi.org/10.7269/P34F1NNJ>
- Evans, P L, Strollo, A, Clark, A, Ahern, T, Newman, R, Clinton, J F, Pedersen, H and Pequegnat, C** 2015 Why seismic networks need digital object identifiers. *Eos*, 96. DOI: <https://doi.org/10.1029/2015EO036971>
- Fenner, M** 2011 The Trouble with DOIs. Available at: <https://blogs.plos.org/mfenner/2011/10/09/the-trouble-with-dois/> [Last accessed 1 February 2017].
- FORCE11** 2016 Endorse the data citation principles. Available at: <https://www.force11.org/datacitation/endorsements> [Last accessed 01 February 2017].
- Gent, I, Jones, C and Matthews, B** 2015 Guidelines for persistently identifying software using DataCite: a JISC Research Data Spring Project. Available at: <https://purl.org/net/epubs/work/24058274> [Last accessed 1 February 2017].
- Google Scholar** 2017 Search tips. Available at: <https://scholar.google.com/intl/us/scholar/help.html#overview> [Last accessed 1 February 2017].
- GSA** 2014 GSA data policy for publications. Available at: <https://www.geosociety.org/gsa/pubs/datapolicy.aspx> [Last accessed 1 February 2017].
- Hanson, B** 2016 AGU opens its journals to author identifiers. *Eos*, 97. DOI: <https://doi.org/10.1029/2016EO043183>
- Hanson, B, Lehnert, K and Cutcher-Gershenfeld, J** 2015 Committing to publishing data in the Earth and space sciences. *Eos*, 96. DOI: <https://doi.org/10.1029/2015EO022207>
- Hanson, B and Van Der Hilst, R** 2014 AGU's data policy: History and context. *Eos*, 95(37): 337–337. DOI: <https://doi.org/10.1002/2014EO370008>
- Helly, J J, Elvins, T T, Sutton, D and Martinez, D** 1999 A Method for interoperable digital libraries and data repositories. *Future Generation Computer Systems*, 16(1): 21–28. DOI: [https://doi.org/10.1016/S0167-739X\(99\)00032-1](https://doi.org/10.1016/S0167-739X(99)00032-1)
- Helly, J, Staudigel, H and Koppers, A** 2003 Scalable models of data sharing in Earth sciences. *Geochemistry, Geophysics, Geosystems*, 4(1). DOI: <https://doi.org/10.1029/2002GC000318>
- Hills, D, Downs, R R, Duerr, R E, Goldstein, J C, Parsons, M A and Ramapriyan, H K** 2015 The importance of data set provenance for science, *Eos*, 96. DOI: <https://doi.org/10.1029/2015EO040557>
- International DOI Foundation** 2015 Factsheet: DOI System and the Handle System. Available at: <https://www.doi.org/factsheets/DOIHandle.html> [Last accessed 1 February 2017].
- International DOI Foundation** 2016a Fact sheet: Key facts on digital object identifier system. Available at: <https://www.doi.org/factsheets/DOIKeyFacts.html> [Last accessed 1 February 2017].
- International DOI Foundation** 2016b DOI handbook, Section 2.2. DOI: <https://doi.org/10.1000/182>
- James, N and Wanchoo, L** 2015 Developing data citations from digital object identifier metadata. *American Geophysical Union (AGU) Fall Meeting*; 14–18 Dec. 2015; San Francisco, CA; United States. Available at: <https://hdl.handle.net/2060/20150023484>
- Klump, J, Bertelmann, R, Brase, J, Diepenbroek, M, Grobe, H, Höck, H, Lautenschlager, M, Schindler, U, Sens, I and Wächter, J** 2006 Data publication in the open access initiative. *Data Science Journal*, 5: 79–83. DOI: <https://doi.org/10.2481/dsj.5.79>
- Klump, J, Huber, R and Diepenbroek, M** 2016 DOI for geoscience data – how early practices shape present perceptions. *Earth Science Informatics*, 9(1): 123–136. DOI: <https://doi.org/10.1007/s12145-015-0231-5>
- Lee, D J and Stvilia, B** 2012 Identifier schemas and research data. *Proceedings of the American Society for Information Science and Technology*, 49(1): 1–4. DOI: <https://doi.org/10.1002/meet.14504901311>
- Lin, J and Strasser, C** 2014 Recommendations for the role of publishers in access to data. *PLOS Biology*, 12(10): e1001975. DOI: <https://doi.org/10.1371/journal.pbio.1001975>
- Lynch, C** 1998 Identifiers and their role in networked information applications. *Bulletin of the American Society for Information Science and Technology*, 24(2), 17–20. DOI: <https://doi.org/10.1002/bult.80>


- Ma, X, Zheng, J G, Goldstein, J C, Zednik, S, Fu, L, Duggan, B, Aulenbach, S M, West, P, Tilmes, C and Fox, P** 2014 Ontology engineering in provenance enablement for the National Climate Assessment. *Environmental Modelling & Software*, 61: 191–205. <https://doi.org/10.1016/j.envsoft.2014.08.002>
- Mayernik, M S, Phillips, J and Nienhouse** 2016 Linking Publications and Data: Challenges, Trends, and Opportunities. *D-Lib Magazine*, 22(5/6). DOI: <https://doi.org/10.1045/may2016-mayernik>
- Mayernik, M S, Ramamurthy, M K and Rauber, R M** 2015 Data archiving and citation within AMS Journals. *Journal of Climate*, 28(7): 2529–2530. DOI: <https://doi.org/10.1175/2015JCLI2222.1>
- McNutt, M** 2016 Taking up TOP. *Science*, 352(6290): 1147. DOI: <https://doi.org/10.1126/science.aag2359>
- McNutt, M, Lehnert, K, Hanson, B, Nosek, B A, Ellison, A M and King, J L** 2016 Liberating field science samples and data. *Science*, 351(6277): 1024–1026. DOI: <https://doi.org/10.1126/science.aad7048>
- NASA** 2015 Open data and the importance of data citations: the NASA EOSDIS perspective. Available at: <https://earthdata.nasa.gov/open-data-and-the-importance-of-data-citations-the-nasa-eosdis-perspective> [Last accessed 1 February 2017].
- NASA** 2017 EOSDIS DOIs status and listing. Available at: <https://wiki.earthdata.nasa.gov/display/DOIsforEOSDIS/EOSDIS+DOIs+Status+and+Listing> [Last accessed 8 February 2017].
- NCAR** 2016 NCAR Command Language. DOI: <https://doi.org/10.5065/D6WD3XH5>
- NCAR UCAR** 2017 NCAR UCAR Research Data Archive. Available at: <https://rda.ucar.edu/> [Last accessed 25 January 2017].
- NOAA** 2015 NOAA Data Citation Procedural Directive, version 1.1, June 1, 2015: NOAA Environmental Data Management Committee. Available at: <https://nosc.noaa.gov/EDMC/PD.DC.php> [Last accessed 9 February 2017].
- NSF** 2014 Dear Colleague Letter – Supporting scientific discovery through norms and practices for software and data citation and attribution. Available at: <https://www.nsf.gov/pubs/2014/nsf14059/nsf14059.jsp> [Last accessed 9 February 2017].
- OMB** 2013 Memorandum for the heads of executive departments and agencies: Open-Data Policy, managing information as an asset. Available at: <https://project-open-data.cio.gov/policy-memo/> [Last accessed 9 February 2017].
- Parsons, M A** 2014 Data citation then and now. Available at: [https://tw.rpi.edu/media/latest/parsons\\_citation\\_geodata2014.pdf](https://tw.rpi.edu/media/latest/parsons_citation_geodata2014.pdf) [Last accessed 9 February 2017].
- Parsons, M A and Fox, P A** 2013 Is Data publication the right metaphor?. *Data Science Journal*, 12: WDS32–WDS46. DOI: <https://doi.org/10.2481/dsj.WDS-042>
- Paskin, N** 1999 DOI: Current status and outlook. *D-Lib Magazine*, 5(5). DOI: <https://doi.org/10.1045/may99-paskin>
- Paskin, N** 2006 Digital object identifiers for scientific data. *Data Science Journal*, 4: 12–20. DOI: <https://doi.org/10.2481/dsj.4.12>
- Piowar, H A** 2011 Who shares? Who doesn't? Factors associated with openly archiving research data. *PLoS One*, e18657. DOI: <https://doi.org/10.1371/journal.pone.0018657>
- Prakash, G, Shrestha, B, Younkin, K, Jundt, R, Martin, M and Elliott, J** 2016 Data always getting bigger—a scalable DOI architecture for big and expanding scientific data. *Data*, 1(2). DOI: <https://doi.org/10.3390/data1020011>
- Ramapriyan, H K, Goldstein, J C, Hua, H and Wolfe, R E** 2016 Tracking and establishing provenance of Earth science datasets: a NASA-Based example, *Provenance and Annotation of Data and Processes*, Mattoso, M and Glavic, B (eds.), 6th International Provenance and Annotation Workshop, IPAW 2016 McLean, VA, USA, June 7–8, 2016, Proceedings. DOI: <https://doi.org/10.1007/978-3-319-40593-3>
- Rauber, A, Asmi, A, van Uytvanck, D and Pröll, S** 2015 Data citation of evolving data: Recommendations of the Working Group on Data Citation (WGDC). Available at: [https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations\\_151020.pdf](https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf) [Last accessed September 15, 2016].
- Rosenblatt, B** 1997 The Digital object identifier: Solving the dilemma of copyright protection online. *The Journal of Electronic Publishing*, 3(2). DOI: <https://doi.org/10.3998/3336451.0003.204>
- Smith, A M, Katz, D S and Niemeyer, K E** FORCE11 Software Citation Working Group 2016 Software citation principles. *Peer J Computer Science*, 2: e86. DOI: <https://doi.org/10.7717/peerj-cs.86>

- Starr, J, Castro, E, Crosas, M, Dumontier, M, Downs, R R, Duerr, R, Clark, T, et al.** 2015 Achieving human and machine accessibility of cited data in scholarly publications. *PeerJ Computer Science*, 1: e1. DOI: <https://doi.org/10.7717/peerj-cs.1>
- Tilmes, C, Yesha, Y and Halem, M** 2010 Tracking provenance of Earth science data. *Earth Science Informatics*, 3(1): 59. DOI: <https://doi.org/10.1007/s12145-010-0046-3>
- USGS** 2016 USGS data management- Data citation. Available on-line at: <https://www2.usgs.gov/datamanagement/describe/citation.php> [Last accessed 22 September 2016].
- Wanchoo, L, James, N and Ramapriyan, H K** 2017 NASA EOSDIS data identifiers: Approach and system. *Data Science Journal*, 16. DOI: <https://doi.org/10.5334/dsj-2017-015>
- Weber, N M, Thomer, A K, Mayernik, M S, Dattore, R E, Ji, Z and Worley, S** 2013 The Product and system specificities of measuring impact: Indicators of use in research data archives. *International Journal of Digital Curation* 8(2): 223–234. DOI: <https://doi.org/10.2218/ijdc.v8i2.286>
- Williams, D N, Ananthakrishnan, R, Bernholdt, D E, Sim, A, et al.** 2009 The Earth System Grid: Enabling access to multimodel climate simulation data. *Bulletin of the American Meteorological Society*, 90(2): 195–205. DOI: <https://doi.org/10.1175/2008BAMS2459.1>
- Wilson, A, Cox, M, Elsborg, D, Lindholm, D and Traver, T** 2015 A Semantically enabled metadata repository for scientific data. *Earth Science Informatics*, 8(3): 649–661. DOI: <https://doi.org/10.1007/s12145-014-0175-1>
- Wolfe, R E, Duggan, B, Aulenbach, S M, Goldstein, J C, Tilmes, C and Buddenberg, A** 2015 Providing provenance to instruments through the US Global Change Information System. Geoscience and Remote Sensing Symposium (IGARSS), 2015 IEEE International. DOI: <https://doi.org/10.1109/IGARSS.2015.7325719>
- Wynholds, L** 2011 Linking to scientific data: Identity problems of unruly and poorly bounded digital objects. *International Journal of Digital Curation*, 6(1). DOI: <https://doi.org/10.2218/ijdc.v6i1.183>

**How to cite this article:** Goldstein, J C, Mayernik, M S and Ramapriyan, H, K 2017 Identifiers for Earth Science Data Sets: Where We Have Been and Where We Need to Go. *Data Science Journal*, 16: 23, pp. 1–12, DOI: <https://doi.org/10.5334/dsj-2017-023>

**Submitted:** 02 October 2016    **Accepted:** 30 March 2017    **Published:** 21 April 2017

**Copyright:** © 2017 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited. See <http://creativecommons.org/licenses/by/4.0/>.

 *Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

**OPEN ACCESS** 