DATA SCIENCE JOURNAL

Aquino, J et al 2017 Motivation and Strategies for Implementing Digital Object Identifiers (DOIs) at NCAR's Earth Observing Laboratory – Past Progress and Future Collaborations. *Data Science Journal,* 16: 7, pp. 1–7, DOI: https://doi.org/10.5334/dsj-2017-007

PRACTICE PAPER

Motivation and Strategies for Implementing Digital Object Identifiers (DOIs) at NCAR's Earth Observing Laboratory – Past Progress and Future Collaborations

Janine Aquino, John Allison, Robert Rilling, Don Stott, Kathryn Young and Michael Daniels

Earth Observing Laboratory, National Center for Atmospheric Research, Boulder, Colorado, US

Corresponding author: Janine Aquino (janine@ucar.edu)

In an effort to lead our community in following modern data citation practices by formally citing data used in published research and implementing standards to facilitate reproducible research results and data, while also producing meaningful metrics that help assess the impact of our services, the National Center for Atmospheric Research (NCAR) Earth Observing Laboratory (EOL) has implemented the use of Digital Object Identifiers (DOIs) (DataCite 2017) for both physical objects (e.g., research platforms and instruments) and datasets. We discuss why this work is important and timely, and review the development of guidelines for the use of DOIs at EOL by focusing on how decisions were made. We discuss progress in assigning DOIs to physical objects and datasets, summarize plans to cite software, describe a current collaboration to develop community tools to display citations on websites, and touch on future plans to cite workflows that document dataset processing and quality control. Finally, we will review the status of efforts to engage our scientific community in the process of using DOIs in their research publications.

Keywords: data management; digital object identifiers (DOIs); persistent identifiers; data stewardship; reproducibility

Introduction

Over 50 years ago, the National Center for Atmospheric Research (NCAR) was established to serve the broader atmospheric research community and extend their capabilities for research beyond what a single university could provide. Chief among these capabilities was the development and deployment of observational research platforms that allow scientists to sample data around the globe. The Earth Observing Laboratory (EOL) at NCAR is responsible for maintaining and deploying the majority of these research platforms, including aircraft, radars, profilers, flux systems, and sounding systems, all equipped with state-of-the-art instrumentation developed in collaboration with the university community. Referred to as Lower Atmosphere Observing Facilities (LAOF), these National Science Foundation (NSF)-funded facilities can be requested by the community (EOL 2013a). The primary deliverable of campaigns that make use of these research platforms is the data that EOL collects, quality controls, archives and distributes to scientists.

Increasingly, it has become important to provide metrics quantifying the services that NCAR provides to its community, and to link datasets and software to the publications derived from them (Yale 2010: 8; Peng 2011: 1226–1227). Toward that end, an NCAR-wide initiative to employ Digital Object Identifier (DOI) technology was formed in 2010 that produced an important technical note describing how DOIs could best be implemented across a complex organization such as NCAR (Mayernik 2012).

Within EOL, implementation required facing and solving several challenges in terms of metadata consistency, granularity of datasets, proper attribution and automation of the DOI creation process. To aid in tracking the usage of research platforms, a mechanism for assigning DOIs to physical objects such as aircraft, radars, and other instrumentation was decided upon. The EOL implementation procedures were established and approved by EOL management in 2014 (EOL 2014). This paper describes details about the development and implementation of DOI policies and procedures within EOL. We hope it will serve as a useful example for others in the geosciences doing the same.

Timeliness

DOIs are at the forefront of sustainable data management practices. A 2013 Office of Science and Technology Policy (OSTP) memorandum (Holdren 2013) states:

> '. . . the results of unclassified research that are published in peer-reviewed publications directly arising from Federal funding should be stored for long-term preservation and publicly accessible to search, retrieve, and analyze in ways that maximize the impact and accountability of the Federal research investment.'

Within the geosciences community, EOL has collaborated with EarthCube (http://earthcube.org), the Federation for Earth Science Information Partners (ESIP) (http://www.esipfed.org/), Earth Science Data System Working Groups (ESDSWG) (https://earthdata.nasa.gov/community/earth-science-data-system-working-groups-esdswg) , and Working Towards Sustainable Software for Science Practice and Experiences (WSSSPE4) (http://wssspe.researchcomputing.org.uk/wssspe4/), all of which have current initiatives toward developing guidelines for using DOIs. In September 2015, NCAR formed a Data Stewardship Engineering Team (DSET 2015) to expand DOI usage across NCAR. Their efforts led to the realization that training is needed to help data providers understand the level of expertise and time needed to maintain DOIs.

As a continuation of our engagement with the NSF EarthCube effort, EOL hosted a Geosciences Paper of the Future training seminar at NCAR on September 28, 2015 (http://www.ontosoft.org/gpf/training-sessions). The seminar, given by members of the OntoSoft EarthCube Building Block and spanning many geosciences areas, pointed out the importance of documenting the datasets, software and scientific workflow of a research publication so that data and research results are reproducible. With access to the data, software and scientific workflow, a user should be able to begin with the most 'raw' data and reproduce derived data used to support scientific conclusions reported in a publication. EOL followed up by hosting a series of discussions among its own scientists to initiate better practices in this regard, ultimately benefitting scientists within the Laboratory who wish to build upon their own results. Forward-thinking scientists in our lab have taken these practices further by demonstrating a direct linkage of software and data with the derived publication, and documents that describe the workflow (Cooper 2016: 147).

Data Services at EOL

The EOL Metadata Database and Cyberinfrastructure (EMDAC) system (EOL 2013b) is a comprehensive metadata database and integrated cyberinfrastructure that is the hub of all EOL data services. Versions of EMDAC date back to the early 1990's. EMDAC now hosts over 11,000 datasets with confirmed metadata from over 500 field campaigns going back to 1967. Implemented to follow community metadata standards, EMDAC contains file, dataset, and campaign-level metadata. Stewardship of, and access to, this valuable community resource is a primary responsibility of EOL.

Development of DOI Implementation Guidelines

During 2014, an internal committee consisting of data managers from across EOL developed guidelines for implementing DOIs within EOL. Scope was limited to datasets collected using EOL-hosted facilities during NSF-funded field campaigns ('EOL datasets') and EOL-managed LAOF. By following the DataCite metadata schema, the content of some DOI metadata fields were readily agreed upon. Areas that required additional thought are discussed below. The deliverable for this committee was the EOL Implementation Procedures Document (EOL 2014), accessible online to our users and the community in general.

DOIs versus other persistent identifiers

The decision to use DOIs over other persistent identifiers was made at the NCAR committee level (Mayernik, 2012). Two major considerations in this decision were (1) the familiarity of the academic and scientific communities with DOIs, and (2) the appropriateness of DOIs for resources that are expected to have long-term community interest and value. See Mayernik (2012) for a full discussion.

Authorship

Authorship provides a primary metric by which scientists are evaluated. Authorship can include any individual making significant contributions to creation and stewardship of the dataset. Determining authorship, contributor type, and citation order can be difficult and time-consuming. The development of clear guidelines mapping the roles of contributors to DataCite fields is a critical component of accurate authorship assignment.

As EOL instruments are owned by NCAR and operated by EOL staff, EOL is given authorship for all EOL datasets with individuals given other contributor roles. For non-EOL (external) datasets housed at EOL, which constitute the bulk of EOL's archive, authorship is determined in collaboration with the instrument PI. Composite datasets contain many component datasets, which likely each have a different author. For composites formed at EOL, EOL is listed as the author and component datasets available via EMDAC are assigned DOIs and linked to the composite dataset via the DataCite 'relatedIdentifier' field (DataCite 2014). An ORCiD (http://orcid.org) or other author registration service is not required, but is recorded with contact metadata if available.

Granularity of DOI assignment

There is no community standard for the granularity at which DOIs are assigned. Within EOL, this decision was made to align with EMDAC, which is built around collections at the dataset level (a collection of files forms a dataset; a collection of datasets constitutes a field campaign; datasets are delimited by instrument or platform, and field campaign). The decision was made NOT to assign a new DOI for a new dataset version. Instead, version indicator and creation date are tracked within EMDAC, the version is included as part of the 'Title' and 'Version' metadata fields registered with the DOI, and the version and access date are given in the citation displayed on the dataset homepage and mailed to users with each data download:

> EXAMPLE: *UCAR/NCAR – Earth Observing Laboratory. 2016. NSF/NCAR GV (HIAPER) QC Dropsonde Data,* ***Version 4.0****. UCAR/NCAR – Earth Observing Laboratory. http://dx.doi.org/10.5065/D67H1GSR.* ***Accessed 15 Apr 2016****.*

Authors should always follow journal citation requirements. If the version number is omitted, the version can still be recovered from the access date. Best practices concerning versioning a DOI vs assigning a new DOI are an open area of discussion in the community. We follow Mayernik (2012) versioning guidelines: Any changes to the data require assignment of a new version; While metadata versions are not explicitly tracked, a new version may be assigned if metadata changes are deemed integral to the data. We also keep meticulous notes on versioning decisions within EMDAC.

Use of a single file requires reference to the entire dataset with a 'Subset used' designation listed after the version. 'Suffix passthrough', a concept implemented by the University of California EZID service, cites a single file by appending a suffix to the identifier. The resulting id is resolvable and the file citation suffix is "passed through" to the target (EZID 2017). In order to reference an entire field campaign, multiple datasets must be cited leading to repetitive citations in the text, e.g., (UCAR/NCAR – Earth Observing Laboratory [EOL] 2016a), (EOL 2016b), (EOL 2016c). Although this could be perceived as a burden on users, we believe it allows for better metrics and clearer citation counts for dataset authors.

Implementing DOIs at EOL

Infrastructure improvements

The EMDAC relational database is accessed through an interactive web application that steps users through the process of creating datasets and adding metadata. In preparation for assigning and tracking DOIs for datasets, the schema of the relational database was updated with the addition of tables for dataset versions and DOI information. Once data and metadata accuracy and completeness have been verified, minting a DOI for a dataset using EMDAC triggers an automated submission to EZID. The returned handle is added to a new record in the DOI table along with the current dataset version, and the new record is linked to the dataset.

In order to increase access to citation information containing DOIs, each dataset landing page displays the DOI, a permanent link to the dataset, a link to the DataCite XML metadata, as well as ISO 19115 conforming metadata that includes the dataset DOI. A sample citation following ESIP (2012) guidelines is displayed on the page, and a large number of additional citation styles may be pulled from the CrossCite DOI citation service and displayed. For an example, see EOL (2016c).

Metadata cleanup

Metadata for approximately 11,000 historical datasets contained within EMDAC were reviewed, and updated where necessary. Where versioning had not previously been applied, version 1.0 was assigned. Almost 3,000 EOL datasets were identified as final versions available online and DOIs were successfully assigned. Other non-EOL data require additional authorship assignment before DOIs can be assigned, a task that has yet to be tackled in bulk.

Data Curation

As computer systems and community standards evolve, datasets become inaccessible due to obsolete formats and outdated software tools. Efforts to make these 'dark' data available include compiling lists of field campaigns completed and determining their current availability, locating data, determining the level of existing quality control, updating metadata, performing often complex format conversions from obsolete to modern formats, adding data to the data archive, and assigning DOIs to both original legacy-formatted data and reformatted data. In addition, relatively large EOL datasets historically were distributed via offline media (such as exabyte and VHS tapes). The advent of inexpensive removable disks, high speed networks, and a very large centralized NCAR storage system have allowed the migration of old data from offline media to online archives.

Approximately 215 'dark' datasets from 70 field campaigns which took place between 1975 and 2016 have been made accessible again and assigned DOIs.

Web page development

Complementary metadata not contained within a DOI, such as field campaign dates, locations, objectives, summaries of instrument functionality and quality control of datasets, are crucial for full understanding and use of these data. Where metadata for historical data resided on inaccessible web pages, a project page consolidating this information was developed (or updated if extant) and linked to the dataset.

DOIs for physical objects

EOL assigned DOIs to 23 NCAR-managed instruments and platforms (LAOF) and one software service, the EOL Field Catalog (EOL 1995), to facilitate the generation of usage metrics. The DataCite resource type 'PhysicalObject' was used to identify these facilities. 'PublicationYear' was set to the year the facility became available, and 'Date' was left blank. As many individuals contribute to the upkeep and deployment of a facility, authorship was assigned to EOL, and as with datasets, other DataCite roles were used to give credit to individuals associated with the facility.

Drupal DOI module

For each Physical Object assigned a DOI above, a digital representation in the form of a web-hosted landing page was created within the EOL Drupal content management system (CMS) hosted website. A Drupal module to read metadata from DataCite and generate a citation to display on the page was developed, ensuring DOI metadata remain up-to-date. This module is available on GitHub for community use and contributions to development (EOL 2016a).

'First use' papers

Historically, to cite a facility, authors would cite a paper detailing the 'first use' of the facility:

> EXAMPLE: *Lutz, J., P. Johnson, B. Lewis, E. Loew, M. Randall and J. Van Andel, 1995: NCAR SPol: Portable polarimetric S-band radar. Preprints, Ninth Symp. on Meteorological Observations and Instrumentation, Charlotte, NC, Amer. Meteor. Soc., 408–410.*

With the addition of DOIs for facilities, (e.g., for S-Pol http://dx.doi.org/10.5065/D6RV0KR8), there is a potential loss of citation count to the engineers and scientists who wrote the descriptive 'first use' paper for the instrument. EOL suggests authors continue to include these publications in their citations, as well as citing the relevant DOIs. In order to facilitate this dual citation, EOL provides clear guidance in the form of an example citation containing both components (if extant) at the bottom of each facility DOI landing page.

Guidance to authors

When EOL began advertising the existence of DOIs to in-house authors, it became apparent that citing DOIs was not intuitive. Authors were unclear how to reference the citation in the text, and how to handle multiple in-text citations of different datasets from EOL-hosted instruments deployed during a single campaign.

To answer these and other questions, and to facilitate and ease community use of DOIs, EOL has developed a document providing guidance to authors on citing physical objects and datasets, both within the body of a publication and in the references section. This document is available on the EOL website (EOL 2016b).

Ongoing Efforts

EOL has assigned DOIs to 90% of EOL datasets collected between 10/1/2004 (FY 2005) and early 2016. As we approach 100%, EOL has begun to turn its attention to documenting two other areas required for reproducible research: software and scientific workflows.

While working to demonstrate enhanced research reproducibility by linking publications to the data, software, and workflows used in analysis, EOL has developed a standard workflow template that documents all components of data processing and quality control used to create a given dataset. For an example, see Mahoney (2008). This basic information has been documented for all EOL-hosted instruments and we will now work to develop guidelines for citing these workflows.

Through multiple current community initiatives, the geosciences community is working to develop guidelines for citing software that meet the needs of reproducible research. EOL is collaborating with the combined ESDSWG Software Citations Working Group and ESIP subgroup on the 'Role of Identifiers in the World of Software'. In addition, EarthCube has multiple software citation Building Blocks (EarthCube 2016) which we are watching closely.

Lessons Learned

(1) DOIs are the appropriate identifier for resources that are expected to have long-term community interest and value.
(2) Authorship is a high-profile area that must be carefully considered when assigning DOIs.
(3) A DOI assignment granularity of "dataset" (a collection of files) hits an appropriate balance between good metrics for dataset authors and the burden of multiple citations per publication.
(4) Creating a digital landing page and assigning DOIs to Physical Objects facilitates credit for instrument PIs.
(5) Complete, accurate metadata is critical for dataset discovery. To keep metadata up to date, automation of the linkage between local metadata and DataCite is preferred.
(6) To ease and facilitate the use of DOIs, provide guidance to authors on how to cite DOIs in their publication, and provide example citations. Include a reference to the first use paper so the DOI does not replace this valuable citation.
(7) Citing data and instrumentation is not enough to enable reproducibility – software and workflows must also be cited.

Acknowledgements

The authors wish to thank Linda Cully, Steve Williams, and Eric Dattore for their assistance developing guidelines and implementing DOIs at EOL. We acknowledge Julie Anne Jubera, Matt Paulus, Linda Echo-Hawk, Brooks Snyder, and Tyler Lugger for their countless hours loading data, developing documentation, and reviewing metadata. Without their dedication and careful work, DOI implementation at EOL would be much less complete.

Thanks to Matt Mayernik, Brigitte Baeuerle, and Linda Cully for their thoughtful comments and suggestions for this manuscript.

This work could not have been completed without the financial support of the National Science Foundation under Cooperative Agreement No. AGS-0856145. Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Competing Interests

MD is on the board of directors of the Foundation for Earth Science Information Partners (ESIP). MD is also the lead PI for the Cloud-Hosted Real-time Data Services for the Geosciences (CHORDS) Earth Cube Building Block and Co-PI of the Enabling Scientific Collaboration and Discovery through Semantic Connections (EarthCollab) EarthCube Building Block. All other authors have no competing interests.

Author Information

Priding themselves with over 140 years of combined experience, the authors have been providing data management services to the Earth sciences community for over two decades. With expertise in atmospheric sciences, computer science, system administration, statistics, physics, and art, they bring a broad range of expertise and a creative perspective to solving the challenges of modern data management. As a team they have been on the cutting edge of developing services and implementing new ideas across the data management landscape.

References

Cooper, W A, et al. 2016 *Characterization of Uncertainty in Measurements of Wind from the NSF/NCAR Gulfstream V Research Aircraft.* NCAR Technical Note NCAR/TN-528+STR, 175 pp, DOI: https://doi.org/10.5065/D60G3HJ8

DataCite 2014 *DataCite schema V3.0* (August 2015). DOI: https://doi.org/10.5438/0010 (Accessed November 2014).

DataCite 2017 *Our Mission*. Available at: https://www.datacite.org/mission.html (Accessed January 2017).

EarthCube 2016 *EarthCube Funded Projects – Building Blocks* (August 2016). Available at: http://earthcube.org/group-type/funded-projects-building-blocks (Accessed 28 August 2016)

EZID 2017 Suffix Passthrough Explained. Available at: http://ezid.cdlib.org/learn/suffix_passthrough (Accessed January 2017).

Foundation for Earth Science Information Partners (ESIP) 2012 *Interagency Data Stewardship/Citations/provider guidelines*. Available at: http://wiki.esipfed.org/index.php/Interagency_Data_Stewardship/Citations/provider_guidelines (Accessed 23 January 2017).

Holdren, J P 2013 (22 Feb) *Increasing Access to the Results of Federally Funded Scientific Research*. OSTP public access memo. Available at: https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013. pdf (Accessed 21 September 2016).

Mahoney, M J and **UCAR/NCAR – Earth Observing Laboratory** 2008 Microwave Temperature Profiler (MTP) for HIAPER. UCAR/NCAR – Earth Observing Laboratory. DOI: https://doi.org/10.5065/D6251G7K (Retrieved December 20, 2016).

Mayernik, M, et al. 2012 *Data citations within NCAR/UCP*. NCAR Technical Note NCAR/TN-492+STR. DOI: https://doi.org/10.5065/D6ZC80VN (Accessed 14 November 2014).

Peng, R D 2011 Reproducible Research in Computational Science. *Science*, 334(6060): 1226–1227. DOI: https://doi.org/10.1126/science.1213847

UCAR/NCAR – Data Stewardship Engineering Team 2015 *NCAR Data Stewardship Engineering Team (DSET)*. Available at: https://ncar.ucar.edu/data-stewardship-engineering-team-dset (Accessed 28 August 2016)

UCAR/NCAR – Earth Observing Laboratory 1995-present EOL Field Catalog. Available at: https://doi.org/10.5065/D6SQ8XFB (Accessed 28 August 2016).

UCAR/NCAR – Earth Observing Laboratory 2013a *Digital Guide to the National Science Foundation's Lower Atmospheric Observing Facilities (LAOF)*. Available at: https://www.eol.ucar.edu/laof-digital-guide (Last accessed 28 August 2016).

UCAR/NCAR – Earth Observing Laboratory 2013b *The EOL Metadata Database and Cyberinfrastructure (EMDAC)*. Available at: http://data.eol.ucar.edu (Last accessed 28 August 2016).

UCAR/NCAR – Earth Observing Laboratory 2014 *EOL Digital Object Identifiers (DOI) Implementation Procedures*. Available at: https://www.eol.ucar.edu/content/implementation-procedures (Last accessed 28 August 2016).

UCAR/NCAR – Earth Observing Laboratory 2016a *Drupal DOI module* (31 August 2016). Available at: https://github.com/NCAR/drupal_DOI (Last accessed 28 August 2016).

UCAR/NCAR – Earth Observing Laboratory 2016b *Digital Object Identifiers (DOI) Guidance to Authors*. DOI: https://doi.org/10.5065/D6K64GGV (Accessed 2 September 2016).

UCAR/NCAR – Earth Observing Laboratory 2016c Low Rate (LRT – 1 sps) Navigation, State Parameter, and Microphysics Flight-Level Data. Version 1.1. UCAR/NCAR – Earth Observing Laboratory. DOI: https://doi.org/10.5065/D65Q4T96 (Accessed 23 January 2017).

Yale Law School Roundtable on Data and Code Sharing 2010 (June 17) *Reproducible Research: Addressing the Need for Data and Code Sharing in Computational Science,* from Stanford University. Available at: http://web.stanford.edu/~vcs/Conferences/RoundtableNov212009/RoundtableOutputDeclaration.pdf (Retrieved 28 August 2016).

How to cite this article: Aquino, J, Allison, J, Rilling, R, Stott, D, Young, K and Daniels, M 2017 Motivation and Strategies for Implementing Digital Object Identifiers (DOIs) at NCAR's Earth Observing Laboratory – Past Progress and Future Collaborations. *Data Science Journal,* 16: 7, pp. 1–7, DOI: https://doi.org/10.5334/dsj-2017-007

Submitted: 22 September 2016 **Accepted:** 21 February 2017 **Published:** 22 March 2017

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

OPEN ACCESS