Data Citation Guidelines for Earth Science Data, Ver. 2

This working document is now closed. The most current document is here: https://www.dropbox.com/s/9h0gz006sb36tn1/ESIP%20Data%20Citation%20Guidelines%20v2.docx?dl=0

Suggested Citation:

ESIP Data Preservation and Stewardship Committee. 2019. *Data Citation Guidelines for Earth Science Data. Ver. 2*. Earth Science Information Partners. https://doi.org/10…..

Table of Contents

Document Status

2019-03-20

Final review from Data Preservation and Stewardship Committee complete. Now seeking endorsement from the ESIP Assembly.

Related ESIP Documents

ESIP Software and Services Citation Guidelines and Examples. Ver. 1.
https://doi.org/10.6084/m9.figshare.7640426.v4

Data Citation Guidelines for Data Providers and Archives (Previous version of this document)
https://doi.org/10.7269/P34F1NNJ

Introduction

Data citation is an evolving but increasingly important scientific practice. Scholarly publishers in Earth science now require that data used for a publication be formally cited (Stall et al. 2018)

“Data citation is a reference to data for the purpose of credit attribution and facilitation of access to the data” (CODATA/ICSTI 2013, p. CIDCR6). Along those lines, we see several important purposes of data citation within the scholarly communication process:

- To aid scientific reproducibility through direct, unambiguous reference to the precise data used in a particular study.
- To provide fair credit for data creators or authors, data stewards, and other critical people in the data production and curation process.

- To ensure scientific transparency and reasonable accountability for data authors and stewards.
- To aid in tracking the impact of data set and the associated repository through reference in scientific literature.
- To help data authors verify how their data are being used.
- To help future data users identify how others have used the data.

This document provides guidelines for Earth science data stewards to create a data citation that is meaningful to a human reader, and resolvable and actionable by computers. Repositories should provide citation information for each of their data sets.

These guidelines are tailored to Earth science and related data, but they are built from accepted guidance by recognized data science and research communities and international standards organizations, notably the *Joint Declaration of Data Citation Principles* (Data Citation Synthesis Group, 2014). See Bibliography.

We aim to help Earth science data stewards define, maintain, and resolve precise, persistent citations for data they manage. Authors citing data should follow the recommended citation provided by the repository, but if the repository does not provide a citation, these guidelines can be used to construct an appropriate citation.

Citation Content

Overview

In general, data sets should be cited similar to books, but with some additional fields unique to data sets for completeness. (Used here is the author-date style described in "Chicago Manual of Style, 17th Edition".) When users cite data, they need to use the style dictated by their publishers, but by providing an example, data stewards can give users all the important elements that should be included in their citations of data sets. Data stewards need to work closely with data producers and science teams to develop the actual content of the citation to ensure that each data citation unambiguously refers to the data that was used in a particular work and that appropriate attribution is provided.

The core, required concepts in a citation are

> **Author/Creator:** The people or organizations responsible for the intellectual work to develop a data set. The data creator.
>
> **Public Release Date:** When the particular version of the data set was first made available for use (and potential citation) by others.
>
> **Title:** The formal title of the data set. It may also include version or edition information, but should be carefully controlled. A better alternative is to track version information

independent of the title. Note this is the title of the data set, not the project or a related publication. It is important for the data set to have an identity and title of its own.

Version ID: Careful versioning and documentation of version changes are central to enabling accurate citation. Data stewards need to track and clearly indicate precise versions as part of the citation for any version greater than 1. It may be appropriate to track major and minor versions.

Repository: The name of the entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the data. This property will be used to formulate the citation, so consider the prominence of the role. This may be an appropriate place to recognize a major sponsor of the data.

Resolvable Persistent Identifier: The unique identifier that provides the ability to access the data. Not all data have Persistent Identifiers (PIDs) or can be digitally accessed, so an alternative method to access metadata, eg. a URL or a physical address, can be provided instead.

Access Date: Because data can be dynamic and changeable in ways that are not always reflected in release dates and versions, it is important to indicate when online data were accessed.

Other fields can be added as necessary to credit other people and institutions, or to designate a particular resource type. It is also important to provide a scheme for users to indicate the precise version and subset of data that were used. Ideally this is handled with a specific PID for a query, but it could also be the temporal and spatial range of the data, the types of files used, or other ways of describing how the data were subsetted.

Citations should resolve to a structured “landing page” that is both machine and human readable. Ideally, the resolvable PID allows for machines to determine and access the precise data used and to enable appropriate credit and attribution. In practice, human judgement is often necessary. These guidelines seek to provide mechanisms for precise automation and reasonable human interpretation of credit and access while recognizing pragmatic compromises are often necessary.

An example citation:

> Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. Edited by M. A. Parsons and M. J. Brodzik. National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5060/D4MW2F23. Accessed 2008-05-14.

Details on Core Concepts

All of the following concepts should be included in a citation. There are always edge cases that challenge standards. The point, however, is to provide as much standard information as possible to help meet the purposes listed in the introduction.

As a reference, we provide a mapping of each of these concepts to several commonly used metadata standards or dialects.

Author or Creator

This is the name of the individual(s) or organization(s) whose intellectual work, such as a particular field experiment or algorithm, led to the creation of the data set. This is sometimes called the data creator. We prefer the term author because of its implied intellectual effort and its conventional use in citing traditional works. Data stewards, in close collaboration with data providers, need to determine who deserves to receive credit and to accept responsibility for the data set. Similarly, the steward needs to work closely with the providers to define the appropriate level of aggregation for the data set.

In some cases, the data set authors may have also published a paper describing the data in great detail. Data papers should be encouraged, and both the paper and the data set should be cited when the data are used.

In addition to the data author, there may be “editors” or other roles that could be included in the citation if they made significant intellectual contributions. Many people can be involved in the creation of a data set, so their roles should be credited elsewhere in data documentation or metadata.

> **Veefkind, P.**. 2012. OMI/Aura Ozone (O3) DOAS Total Column L3 1 day 0.25 degree x 0.25 degree V3, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/Aura/OMI/DATA3005, Accessed: 2018-10-10.
>
> **Freedman, R.** (2017). Smartphone recorded driving sensor data: Indianapolis International Airport to Urbana, IL. University of Illinois at Urbana-Champaign. https://doi.org/10.13012/B2IDB-4650469_V1. Accessed 2018-02-28.

A particular group or small organization may sometimes be the author, but one should be as specific as possible in accountability and crediting intellectual contribution. Naming an entire funding agency as an author is usually inappropriate. Additional credits can be provided in the documentation and on the landing page, which may be a more appropriate place to recognize funders and other sponsors (see below).

> **MODIS Characterization Support Team (MCST), MODIS Adaptive Processing System (MODAPS) and MODIS Science Data Support Team (SDST).** 2015. MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km, NASA Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC). http://dx.doi.org/10.5067/MODIS/MOD021KM.006, Accessed: 2018-10-12.
>
> **Zou, C-Z, Wang, W., and NOAA CDR Program.** 2013. NOAA Fundamental Climate Data Record (FCDR) of AMSU-A Level 1c Brightness Temperature, Version 1.0. NOAA National Climatic Data Center. https://doi.org/10.7289/V5X63JT2. Accessed 2019-02-07.

Public Release Date

This is the time (date stamp) when a particular version of the data set was released for use (and potential citation) by others. For a completed data set, the release date is simply the year of release. A more precise date can be used if needed to indicate when exactly the data became available and citable. ISO 8601:2019 standard dates are preferred.

> Moschetti, M. P. **2017**. Database of earthquake ground motions from 3-D simulations on the Salt Lake City of the Wasatch fault zone, Utah. U.S. Geological Survey data release. https://doi.org/10.5066/F7V98691. Accessed 2019-02-28.

If detailed versioning information is lacking for a data set, it may aid the reader to try and capture when updates were released. For a data set with infrequent or irregular releases, list the first year of released followed by "last updated" with the current updated release information.

> Wentz, F. J., J. Scott, R. Hoffman, M. Leidner, R. Atlas, and J. Ardizzone. **2016**. **last updated August, 2018**. Cross-Calibrated Multi-Platform Ocean Surface Wind Vector Analysis Product V2, 1987 - ongoing. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory http://rda.ucar.edu/datasets/ds745.1/. Accessed 2018-02-128.

For an ongoing data set that is updated on a regular or continual basis, list the first year of release followed by the last update. Updates could occur annually or more frequently.

> Maslanik, J. and J. Stroeve. 1999, **updated daily.** Near-Real-Time DMSP SSMIS Daily Polar Gridded Sea Ice Concentrations, Version 1. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/U8C09DWVX9LM. Accessed 2019-02-14.

> National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce. **1996**, **updated monthly**. NCEP/NCAR Reanalysis Monthly Mean Subsets (from DS090.0), 1948-continuing. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://rda.ucar.edu/datasets/ds090.2/. Accessed 2018-10-12.

A note on updates vs. new versions:

Ongoing updates to a time series do change the content of the data set, but they do not typically constitute a new version or edition of a data set. New versions typically reflect changes in sampling protocols, algorithms, quality control processes, etc. Both a new version and an update may be reflected in the public release date. The version number should also be included. See also the Sections on Versioning and Locators vs. Identifiers.

> Brodzik, M. J. and R. Armstrong. 2018, updated daily. Near-Real-Time DMSP SSM/I-SSMIS Pathfinder Daily EASE-Grid Brightness Temperatures, **Version 2.** NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/K7VT6D6Y2SO6. Accessed 2019-02-14.

Title

This is the formal title of the data set. It may also include version or edition information, but should be carefully controlled. A better alternative is to track version information independent of the title. Note this is the title of the data set, not the project or a related publication. It is important for the data set to have a unique identity and title of its own.

Veefkind, P.. 2012. **OMI/Aura Ozone (O3) DOAS Total Column L3 1 day 0.25 degree x 0.25 degree,** V3, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/Aura/OMI/DATA3005. Accessed 2018-10-12.

MODIS Characterization Support Team (MCST), MODIS Adaptive Processing System (MODAPS) and MODIS Science Data Support Team (SDST). 2015. **MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km,** Collection 6. NASA Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC)., http://dx.doi.org/10.5067/MODIS/MOD021KM.006, Accessed 2018-10-12.

National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce. 1996, updated monthly. **NCEP/NCAR Reanalysis Monthly Mean Subsets (from DS090.0), 1948-continuing**. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory. http://rda.ucar.edu/datasets/ds090.2/. Accessed 2018-10-12.

Version ID

Careful versioning and documentation of version changes are central to enabling accurate citation. Data stewards need to track and clearly indicate precise versions as part of the citation for any version greater than 1. The citation text should indicate the version used and the date accessed. See below for notes on resolving references to dynamic data.

Veefkind, P. 2012. OMI/Aura Ozone (O3) DOAS Total Column L3 1 day 0.25 degree x 0.25 degree, **V3**, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/Aura/OMI/DATA3005. Accessed 2018-10-12.

MODIS Characterization Support Team (MCST), MODIS Adaptive Processing System (MODAPS) and MODIS Science Data Support Team (SDST). 2015. MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km, **Collection 6**. NASA Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC). https://dx.doi.org/10.5067/MODIS/MOD021KM.006. Accessed 2018-10-12.

Cavalieri, D. J., P. Gloersen, and H. J. Zwally. Edited by J. Maslanik and J. Stroeve. 1999. *Near-Real-Time DMSP SSM/I-SSMIS Daily Polar Gridded Brightness Temperatures, **Version 1**.* NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/AKQDND71ZDLF. Accessed: 2019-02-06.

Repository

This is the organization that maintains and manages the release or distribution of the data set. There is often an implied responsibility for stewardship of the data set. This role is often considered that of a data "publisher," but we avoid that term because it may imply proprietary restrictions or unintended assertions of quality or peer-review. These issues are beyond the intent of data citation. DataCite describes this role well: "The entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the resource. This property will be used to formulate the citation, so consider the prominence of the role." This may be an appropriate place to recognize a major sponsor of the data.

P. Veefkind. 2012. OMI/Aura Ozone (O3) DOAS Total Column L3 1 day 0.25 degree x 0.25 degree, V3, **NASA Goddard Earth Sciences Data and Information Services Center (GES DISC)**. https://doi.org/10.5067/Aura/OMI/DATA3005. Accessed 2018-10-12.

MODIS Characterization Support Team (MCST), MODIS Adaptive Processing System (MODAPS) and MODIS Science Data Support Team (SDST). 2015. MODIS/Terra Calibrated Radiances 5-Min L1B Swath 1km, Collection 6. **NASA Level-1 and Atmosphere Archive & Distribution System (LAADS) Distributed Active Archive Center (DAAC)**. https://dx.doi.org/10.5067/MODIS/MOD021KM.006. Accessed 2018-10-12.

Lynch, L., M. Machmuller, C. Boot, T. Covino, C. Rithner, et al. 2019. Dissolved organic matter chemistry and transport along an Arctic tundra hillslope, Imnavait Creek Watershed, Alaska, 2018. **Arctic Data Center.** https://doi.org/10.18739/A2RF5KF5N. Accessed 2019-02-28.

Note: If the repository changes or the name of the repository changes, it is appropriate to change the repository name in the citation, but the PID should not change unless the data themselves have changed. The change in repository should be described in the documentation.

Resolvable Persistent Identifier (PID)

The unique identifier that provides the ability to access the data. We follow the recommendation of Starr et al 2015: "to use any currently available identifier scheme that is machine actionable, globally unique, and widely (and currently) used by a community, and that has demonstrated a long-term commitment to persistence." So DOIs, Archive Resource Keys (ARKs), Handles, Persistent URLs (PURLs), etc. are acceptable. Academic publishers in the Earth sciences, however, are most familiar with the DOI.

The suffix of the identifier should not include any semantic information such as reference to the archive or instrument because names and responsible parties often change over time, while the identifier should remain unchanged. Finally, to aid human usability the locator should include an easy access protocol. At the time of writing this is typically https. For DOIs this takes the form https://doi.org/.

Cavalieri, D. J., P. Gloersen, and H. J. Zwally. 1999. *Near-Real-Time DMSP SSM/I-SSMIS Daily Polar Gridded Brightness Temperatures, Version 1*. Edited by J. Maslanik and J. Stroeve. NASA National Snow and Ice Data Center Distributed Active Archive Center. **https://doi.org/10.5067/AKQDND71ZDLF**. Accessed: 2019-02-06.

Joughin, I., B. Smith, I. Howat, and T. Scambos. 2015, updated 2018. *MEaSUREs Greenland Ice Sheet Velocity Map from InSAR Data, Version 2*. NASA National Snow and Ice Data Center Distributed Active Archive Center. **doi: https://doi.org/10.5067/OC7B04ZM9G6Q.** Accessed: 2019-02-06.

Moschetti, M. P., 2017, Database of earthquake ground motions from 3-D simulations on the Salt Lake City of the Wasatch fault zone, Utah: U.S. Geological Survey data release. **https://doi.org/10.5066/F7V98691**. Accessed 2019-02-28.

Not all data have PIDs, so provide what information is available to access relevant metadata, eg. a URL. If there is one fixed medium, indicate it. For example, DVD.

> Hernes, P. and K. Kaiser. 2019. Vascular plant and microbial biomarkers of dissolved organic matter data from incubation experiments. Biological and Chemical Oceanography Data Management Office (BCO-DMO). data set version 2019-04-05 [if applicable, indicate subset used]. **http://lod.bco-dmo.org/id/dataset/754885.** Accessed 2019-02-21.

In the case of data on physical media, ideally there would be a PID pointing to a landing page describing how to access the data. If there is no PID, cite the object as applicable to its media as described in common style guides like Chicago Manual of Style or American Psychological Association.

> Portuguese National Meteorological Service. 1956. The Climate of Portugal: Mean Values of Climatic Elements in the National Territory in 1921-1950. Part IX. Lisbon, Portugal.

> United States Geological Survey, 1971. Parkfield, CA Historical Map GEOPDF 15X15 Grid 62500-Scale, 1961, Paper Map, Available at USGS Store, **https://store.usgs.gov/product/308306**.

See additional discussion in the Dynamic and Micro-citation section as well as the Note on Locators vs. Identifiers".

Access Date and Time

Because data can be dynamic and changeable in ways that are not always reflected in release dates and versions, it is important to indicate when on-line data were accessed. This is in keeping with common citation practice for online documents and other resources. Depending on how frequently the data change, it may be necessary to include time as well as date of access.

> P. Veefkind. 2012. OMI/Aura Ozone (O3) DOAS Total Column L3 1 day 0.25 degree x 0.25 degree, V3, NASA Goddard Earth Sciences Data and Information Services Center (GES DISC). https://doi.org/10.5067/Aura/OMI/DATA3005. **Accessed 2018-10-12.**

> Moschetti, M. P., 2017, Database of earthquake ground motions from 3-D simulations on the Salt Lake City of the Wasatch fault zone, Utah: U.S. Geological Survey data release. https://doi.org/10.5066/F7V98691. **Accessed 2019-02-28.**

Additional Considerations

The following concepts may also be considered for certain citation scenarios. We do not provide specific guidance here because these are still active topics of discussion in the community. We do provide some examples of current approaches.

Resource type

“Resource type” is a required or strongly recommended field in DataCite, ISO, and other metadata standards. We understand the motivation behind this but question whether it is helpful or misleading. The “type” of a resource is largely driven by context. For example, what may be an article or literature in one context may be data in another context such as natural language processing.

Also, humans and machines may interpret resource type differently. People can usually figure out a resource type for a particular use from other clues in the citation and the context in which it was cited. Machines need a precise understanding of resource type so they know what operations can be performed against the object (consider MIME types as an example). This suggests that we need not include resource type in the human readable text of the citation, and that repositories should provide structured information when resolving the identifier that allows machines to access and interpret the resource. This may be through a type registry or registered namespace or ontology. It may also be through a protocol with defined operations such as OpenDAP or the Digital Object Interface Protocol. See Resolving Citations.

Editor, Compiler, or other important roles

Sometimes, there are other people besides the authors who played an important role in the creation or development of a data set. Often these people can be characterized as editors or compilers, but other roles might also be identified. An editor is the person or team who is responsible for creating a value-added and possibly quality-controlled product from the data. In cases where there is minimal scientific or technical input, yet still substantial effort in compiling the product, the person may be more correctly cited as a compiler. Editors and compilers may often be responsible for a larger work that includes multiple data sets from different authors. Occasionally, there may be both a compiler and editor as well as other roles. Myriad other roles should be credited elsewhere in data documentation or metadata.

> Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. **Edited by M. A. Parsons and M. J. Brodzik**. National Snow and Ice Data Center Distributed Active Archive Center. Data set accessed 2008-05-14 at https://dx.doi.org/10.5060/D4MW2F23.

Data Within a Larger Work

A particular data set may be part of a compilation, or multiple data sets may be compiled into a larger product. Different repositories address this in different ways, but the basic concept is to cite the data set somewhat like a chapter in an edited volume.

For example, NASA archives and distributes a collection of 32 data sets consisting of carefully cross-calibrated data from passive microwave sensors on eight different satellites and four different temporal aggregations. A single DOI is used for the entire collection, while each of the 32 data sets also has a DOI, thus enabling citation of either the entire collection or one or more of the individual data sets. This approach helps reduce the number of citations in publications when large collections of data sets are to be cited.

> Citing the entire collection:
>
> Wentz,F. J., K. Hilburn and D. K. Smith. 2012. RSS SSM/I Ocean Product Grids NetCDF Collection, Version 7. NASA EOSDIS Global Hydrology Resource Center Distributed Active Archive Center. https://dx.doi.org/10.5067/MEASURES/SSMI-SSMIS/DATA301. Accessed 2019-02-28.

Citing two data sets from the collection:

Wentz, F. J, K. Hilburn and D. K. Smith. 2012. RSS SSM/I Ocean Product Grids Daily from DMSP F8 NetCDF. NASA Global Hydrology Center Distributed Active Archive Center. http://dx.doi.org/10.5067/MEASURES/DMSP-F8/SSMI/DATA301. Accessed 2019-02-28.

Wentz, F. J, K. Hilburn and D. K. Smith. 2012. RSS SSM/I Ocean Product Grids Daily from DMSP F10 NetCDF. Hydrology Center Distributed Active Archive Center. http://dx.doi.org/10.5067/MEASURES/DMSP-F10/SSMI/DATA301. Accessed 2019-02-28.

Dynamic and Micro-citation

This may be the most challenging aspect of data citation. It is necessary to enable "micro-citation" or the ability to refer to the specific data used—the exact files, granules, records, etc. from a particular version. Scientifically, this is to enable reproducibility by providing a precise reference to the data used. It may, however, impact the credit or attribution functions of a citation. Different subsets of a larger collection may have been created by different people. As discussed in Data Within a Larger Work, mechanisms for crediting at finer granularity are still being developed.

Mechanisms for referencing and providing access to precise subsets of data are more established. Ideally, the repository should provide a PID that resolves to the precise subset and version of the data used. We recommend that repositories implement the Research Data Alliance (RDA) Recommendation on Scalable Dynamic Data Citation, which provides a PID for a particular query.

We recognize, however, that not all repositories have the ability to implement the RDA Recommendation so other approaches that can work reasonably well, at least for human interpretation, may be used.

Versioning

In all cases, it is very important to carefully track and document versions of the data set. Individual stewards and data centers will need to develop and follow their own practices, but there are some suggestions on how to handle different data set versions relative to an assigned identifier.

For relatively static data sets, a simple approach is to assign a new identifier every time there is any change to the data or metadata. For more frequently changing data, the Digital Curation Centre (DCC) Data Citation Guidelines (Ball & Duke, 2015) suggest that DOIs be assigned to different data snapshots taken at regular intervals or as needed. This would work well for infrequently changed data sets. DCC also suggests a "time slice" approach where "the citable entity becomes the set of updates made to a data set during a particular time period rather than the full data set itself (e.g. the 2008 data from a series running since 1950)." Similarly, the

Zenodo repository and DataONE support the ability to cite the “Concept” of a data set with one DOI and the specific version of a data set with another DOI: http://help.zenodo.org/#versioning.

These approaches may be workable in some situations, but they are often unwieldy for the frequently updated time-series common in Earth science. Many repositories with such highly dynamic data only assign a new PID when there are major changes to the data (i.e. a major version). They then rely on documentation and timestamps to identify when minor changes have occured (minor versions). Individual stewards need to determine which are “major” vs. “minor” versions and describe the nature and range of every change. Typically, something that affects the whole data set, like a reprocessing with an improved algorithm, would be considered a major version. Ongoing additions to an existing time series need not constitute a new version. This is one reason for capturing the date accessed when citing the data. Small corrections or changes may constitute minor versions and should be explained in documentation, ideally in file-level metadata. This general approach, while workable, relies heavily on human interpretation. The RDA Recommendation provides better specificity and verifiability.

Subset Used

In addition to managing versions it is important to allow reference to the precise subset of data used. Again, the RDA Recommendation is the best approach for specificity and verifiability, but lacking that, repositories may be able to use the DCC snapshot or time-slice approach or the Zenodo concept-instance approach.

Again, these approaches are unwieldy for highly dynamic data, so it may be necessary to rely on a human-based approach or an initially less persistent approach. We see these as temporary solutions while repositories move to more precise, persistent, and machine-actionable approaches.

With the human-based approach, an example in a traditional context would be quoting a certain passage in a book, where one then references a specific page number in the citation. Alternatively, one might make reference to the "structural index" of a canonical text (e.g. book, chapter, and verse in the King James Bible). Unfortunately, data sets typically lack page numbers or canonical versions. Nevertheless, there is often a consistent structural form to how a data set is organized that can help users cite a specific subset. Data stewards should suggest how to reference subsets of their data. With Earth science data, subsets can often be identified by referring to a temporal and spatial range.

> Hall, D. K. and G. A. Riggs. 2016. MODIS/Terra Snow Cover Daily L3 Global 500m Grid, Version 6. **Oct. 2007- Sep. 2008, 84°N, 75°W; 44°N, 10°W**. NASA National Snow and Ice Data Center Distributed Active Archive Center. https://doi.org/10.5067/MODIS/MOD10A1.006. Accessed 2019-02-02.

Sometimes, the data may be packaged in different sub-collections or representations or “Archive Information Units”. If the packages have different content they should be explicitly referenced.

> Cline, D., R. Armstrong, R. Davis, K. Elder, and G. Liston. 2003. CLPX-Ground: ISA snow depth transects and related measurements ver. 2.0. **Shapefiles** Edited by M. A. Parsons and M. J. Brodzik. National Snow and Ice Data Center Distributed Active Archive Center. http://doi.org/10.5060/D4MW2F23. Accessed 2008-05-14 at

Depending on the guidelines from the journal, it may be more appropriate to include this information as part of the in-text citation. E.g., (Cline et al. 2003, shapefiles). Note: while this is a workable solution it does not fully meet the "Specificity and Verifiability" principle of the *Joint Declaration of Data Citation Principles.*

In some cases, it may be possible to include both the citation for the larger data set and specify the specific subset used through a URL where the subset can be specified entirely by parameters, such as:

> Lin, B., J. F. Campbell, J. Dobler, E. V. Browell, S. A. Kooi, S. Pal, T. Fan, W. Erxleben, D. Mcgregor, M. D. Obland, and C. O'dell. 2018. ACT-America: L2 Remotely Sensed Column-average CO2 by Airborne Lidar, Eastern USA. ORNL Distributed Active Archive Center https://doi.org/10.3334/ORNLDAAC/1649 Subset extracted using https://thredds.daac.ornl.gov/thredds/dodsC/ornldaac/1649/ACTAmerica-MFLL-lev2_C130_2016-08-08T160435_R1.nc.ascii?Cloud_Ground_flag[0:1:116742],Column_CO2[0:1:116742] on 2019-02-27.

There has also been some experimentation appending HTTP GET parameters with subset parameters in the DOI URL (e.g. https://dx.doi.org/10.123245/mydatasetid?subsetid=abc1234). This requires further experimentation and relies on the DOI resolver passing the GET parameters through to the end landing page. Note: neither of these approaches fully meet the "Persistence" principle of the *Joint Declaration of Data Citation Principles*.

Resolving Citations

Resolving persistent identifiers is more complex than it seems. The European Commission's FREYA project provides an excellent review of current "PID Resolution Services Best Practices".

We provide some additional guidance below.

Note on Locators vs. Identifiers

Identity and location are often confused or conflated. While one can often use an item's location to identify it or an item's identity to locate it, the concepts are distinct. This is easily understood when we consider a human example. A name such as "John James Doe" (Office Manager at the FOO Data Center) is an identifier. An address such as "123 Main St. #201, Peoria, IL, 12345-1234, USA" is a locator.

The locator might work as an identifier, because you might find John in his office, but he may also have retired and there is a new Office Manager who plays the same role but is not the same person. Similarly, you may be able to locate John based on his name and title, but what happens if he is telecommuting this week and is in Poughkeepsie not Peoria? It is similar with

digital objects. One might be able to identify a data set by its URL, for example, but there is no guarantee that what is at that URL today is the same as what was there yesterday,

Confusingly, a Digital Object Identifier is a locator. It is a Handle-based scheme whereby the steward of the digital object registers one or more locations (typically a URI) for the object. There is no guarantee that the object at the registered location will remain unchanged. Consider a continually updated data time series, for example. Indeed, the advantage is to separate location information from other information about the resource. The location is not part of the resource description so that location can be managed independently, thereby enabling URI changes, migrations to other repositories, etc.

While it is desirable to uniquely identify the cited object, it has proven extremely challenging to identify whether two data sets or data files are scientifically identical unless they are identical at the bit level (Duerr et al. 2011). Content-based (i.e., hash-based) identifiers can be used to identify objects identical at the bit level, and they may even be resolvable Nevertheless, careful consideration of the content used to generate identifiers is needed. For example, two bit-wise identical objects would have different chains of custody if they ended up in two different repositories. The chain of custody could be included in the hash, as appropriate. Otherwise, the two objects' metadata and derivation-provenance are likely to be the same but their usage and distribution information is likely to be very different. We, therefore, recommend that registered, resolvable identifiers (i.e. locators like DOIs) be used in citations because they require an institutional commitment to maintain the location information in the identifier.

In short, multiple identical data sets may have different locators, and one locator may identify multiple data sets. For example, two identical data sets may be provided and managed by two different repositories and have different DOIs. When they do, a suggested practice is that metadata should be provided that indicates that the two DOIs represent the same data set using schema.org's 'sameAs' or a similar property. In other situations multiple versions or representations of the same data set may be located through one DOI.

In general, registered, resolvable, persistent identifiers like ARKs, DOIs, and other Handles are useful to locate full data sets or collections by pointing to a landing page that describes and provides access to the data and its provenance. Other locators and identifiers may be more appropriate for locating or identifying individual records or files.

Landing Pages

The persistent identifier included in a citation should not resolve directly to the data but rather to a landing page describing and providing access to the data. The landing page should persist even if the data are no longer available. This is not the same as a "data paper". The landing page is a living document or end point that can be updated as data are moved or changed. This page is also a good place to provide more attribution for sponsors and other people important in the creation and stewardship of the data.

The landing page should provide structured information that allows humans and machines to access the data and make a basic assessment of the how appropriate the data are for an application. Recommended practice and mechanisms for machine readable landing pages is still evolving.

Content

We find that Starr et al. 2015 provide the best, current guidance on recommended (human readable) content for landing pages:

- **Data set descriptions**: The landing page must provide descriptions of the data sets available and information on how to programmatically retrieve data where a user or device is so authorized. (See data set description for formats.)
    - Minimally the following metadata elements should be present in data set descriptions:
        - Resolvable persistent identifier
        - All the other core citation concepts listed above.
        - Description: A description of the data set, with more information than the title.
    - Additional recommended metadata elements in data set descriptions are:
        - Creator Identifier(s): An ORCID or other unique identifier of the individual creator(s).
        - Other identifiers for projects, instruments, institutions etc.
        - Other contributors who are not credited in the documentation
- **Versions**: What versions of the data are available, if there is more than one version that may be accessed.
- **Explanatory or contextual information**: Provide explanations, contextual guidance, caveats, and/or documentation for data use, as appropriate. Note: Different versions are generally assigned different PIDs and the landing pages should include pointers to prior (and subsequent) versions as appropriate. When a version is deaccessioned, the landing page would still persist, but indicate that the version is no longer available, and point to the version that replaced it.
- **Access controls**: Access controls based on content licensing, Protected Health Information (PHI) status, Institutional Review Board (IRB) authorization, embargo, or other restrictions, should be implemented here if they are required.
- **Persistence statement:** Reference to a statement describing the data and metadata persistence policies of the repository should be provided at the landing page. Data persistence policies will vary by repository but should be clearly described.
- **Licensing information:** Information regarding licensing should be provided, with links to the relevant licensing or waiver documents as required (e.g., Creative Commons CC0 waiver description (https://creativecommons.org/publicdomain/zero/1.0/), or other relevant material.)
- **Data availability and disposition:** The landing page should provide information on the availability of the data if it is restricted, or has been deaccessioned (i.e., removed from

the archive). As stated in the *Joint Declaration of Data Citation Principles*, metadata should persist beyond de-accessioning.
- **Tools/software:** What tools and software may be associated or useful with the data sets and how to obtain them (certain data sets are not readily usable without specific software).

In addition, the **official, formatted citation** for the data set should be included on the landing page.

Different design styles and content layouts can be used to present the above information on a landing page. However, regardless of the presentation format, it is important to assess the usability of the landing page. A user-friendly format will help not only the understandability, but also the accessibility and usefulness of the content.

Actionability

Machine readability or actionability of landing pages is still an evolving practice. Currently, we recommend the approach in the appendix of Starr et al 2015 that recommends that "all versions of the landing page be resolvable from a single URI through content negotiation (Holtzman & Mutz, 1998), serving an HTML representation for humans and the appropriate form for automated agents."

In addition, to aid discovery, we recommend marking up landing pages according to the W3C schema.org specification. The ESIP Semantic Technologies Committee and the RDA Data Discovery Paradigms Interest Group are collaborating on the development of a geosci.schema.org extension to schema.org, which is intended to adequately describe both science repositories and data sets. See https://github.com/ESIPFed/science-on-schema.org.

Currently, the best guidance on how to publish schema.org markup for Earth and environmental science data is at https://github.com/earthcubearchitecture-project418/p418Docs/blob/master/publishing.md, but this may change. The ESIP Semantic Technologies Committee will have the most current guidance.

Finally, there is the question of how to link to the data itself from the landing page. Again, Starr et al. (2015) provide the best current guidance:

> If the data is available from a single file, directly available on the internet, use the DCAT downloadURL to indicate the location of the data.
>
> If the data is available as a relatively small number of files, either as parts of the whole collection, mirrored at multiple locations, or as multiple packaged forms, link to an ORE resource map (Lagoze et al., 2008) to describe the relationships between the files.
>
> If the data requires authentication to access, use the DCAT accessURL to indicate a page with instructions on how to request access to the data. This technique can also be

used to describe the procedures on accessing physical samples or other non-digital data.

If the data is available online but is excessive in volume or number of files, use the DCAT accessURL to link to the appropriate search system to access the data.

For data that are available either as bulk downloads or through sub-setting services, include both accessURL and downloadURL on the landing page.

Acknowledgements

This document was developed from the experience and discussions of the ESIP Data Preservation and Stewardship Committee. The following people contributed to direct writing of the document: Mark A. Parsons (0000-0002-7723-0950), Ruth E. Duerr (0000-0003-4808-4736), Hampapuram K. Ramapriyan (0000-0002-8425-8943), Heather B. Brown (0000-0001-9826-0270), Matthew S. Mayernik (0000-0002-4122-0910), Chung-Yi (Sophie) Hou (0000-0002-8087-1775), Matthew B. Jones (0000-0003-0077-4738), Bruce E. Wilson (0000-0002-1421-1728), and Nancy J. Hoebelheinrich (0000-0002-6797-7903).

Bibliography

Ball, A. & Duke, M. 2015. “How to cite datasets and link to publications.” *DCC How-to Guides*. Edinburgh: Digital Curation Centre. Available at http://www.dcc.ac.uk/resources/how-guides. Accessed 2018-02-10.

CODATA/ICSTI Task Force on Data Citation. 2013. “Out of cite, out of mind: The current state of practice, policy and technology for data citation.” *Data Science Journal 12*: 1-75., http://dx.doi.org/10.2481/dsj.OSOM13-043.

Cousijn, H., A. Kenall, E. Ganley, M. Harrison, D. Kernohan, T. Lemberger, F. Murphy, P. Polischuk, S. Taylor, M. Martone, and T. Clark. 2018. “A data citation roadmap for scientific publishers.” *Sci Data 5* 180259. https://doi.org/10.1038/sdata.2018.259.

Data Citation Synthesis Group. 2014. *Joint Declaration of Data Citation Principles.* Martone M. (ed.). Force11. https://doi.org/10.25490/a97f-egyk.

DataCite Metadata Working Group. 2017. *DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1*. http://dx.doi.org/10.5438/0014.

Duerr, R., R. Downs, C. Tilmes, B. Barkstrom, C.W. Lenhardt, J. Glassy, L. Bermudez, and P Slaughter. 2011. “On the utility of identification schemes for digital Earth science data: an assessment and recommendations.” *Earth Science Informatics 4* 139--60 https://doi.org/10.1007/s12145-011-0083-6.

Fenner, M., M. Crosas, J. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, G. Durand, R. Berjon, S. Karcher, M. Martone, and T. Clark. 2017. “A data citation roadmap for scholarly data repositories.” *BioRxiv* preprint. https://doi.org/10.1101/097196.

Freya Project report on PID Resolution Services Best Practices https://www.project-freya.eu/en/deliverables/freya_d2-1.pdf. Accessed 2018-02-10.

Katz, D., and N. Chue Hong. 2018. “Software citation in theory and practice.” *Arxiv* preprint https://arxiv.org/pdf/1807.08149.pdf. Accessed 2018-02-10.

Lagoze C., H. Van de Sompel, P. Johnston, M. Nelson, R. Sanderson, and S. Warner. 2008. ORE user guide—resource map discovery. Available at http://www.openarchives.org/ore/1.0/discovery. Accessed 2019-02-28.

Rauber, A., A. Asmi, D. van Uytvanck, and S. Pröll. 2016. “Identification of reproducible subsets for data citation, sharing and re-use.” *Bulletin of IEEE Technical Committee on Digital Libraries 12* (1): https://www.ieee-tcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf. Accessed 2018-02-10.

Rauber, A., A. Asmi, D. van Uytvanck, and S. Pröll. 2016. *Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC).* Research Data Alliance. https://doi.org/10.15497/RDA00016.

Stall, S., et al. 2018. “Advancing FAIR data in Earth, space, and environmental science”, *Eos*, 99, https://doi.org/10. 1029/2018EO109301.

Stall, S., P. Cruse, H. Cousijn, J. Cutcher-Gershenfeld, A. de Waard, B. Hanson, J. Heber, K. Lehnert, M. Parsons, E. Robinson, M. Witt, L. Wyborn, L. Yarmey. (2018), Data sharing and citations: New author guidelines promoting open and FAIR data in the Earth, space, and environmental sciences, *Sci. Editor 2018;41(3):83-87.* https://www.csescienceeditor.org/article/data-sharing-and-citations-new-author-guidelines-promoting-open-and-fair-data-in-the-earth-space-and-environmental-sciences/. Accessed 2018-02-10.

Starr, J., E. Castro, M. Crosas, M. Dumontier, R. R. Downs, R. Duerr, L. L. Haak, M. Haendel, I. Herman, and S. Hodson. 2015. “Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications.” *PeerJ Computer Science* 1 e1–. https://dx.doi.org/10.7717/peerj-cs.1

Worthington, S.. 2018, #GenR Software Citation Round-up CCBY4.0, https://doi.org/10.25815/z9ra-6k38