ESIP Data Citation Guidelines

Ruth Duerr
National Snow and Ice Data Center
Outline

• Purpose
• Background
• General format
• Mandatory fields
• Optional fields
Purpose of data citation

- To aid scientific **reproducibility** through direct, unambiguous reference to the precise data used in a particular study. (This is the paramount purpose and also the hardest to achieve).
- 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 authors and stewards.
- To aid in tracking the **impact** of data set and the associated data center 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.
ESIP Data Citation Guidelines

Background

• ESIP Preservation and Stewardship cluster has spent several years working on the identifier and citation topic

• Guideline heritage:
  • Examination of guidelines from ESIP member data centers
  • IPY citation guidelines
  • Tested during application at meetings such as GeoData 2011 and ESIP meetings

• Draft citation guidelines were approved by the General Assembly at Jan. 2012 winter meeting
General format

• Cite a data set as if it were a book!
• For example,

Mandatory Fields

- **Author(s)**—the people or organizations responsible for the intellectual work to develop the data set. The data creators.
- **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
- **Version**—the precise version of the data used. Careful version tracking is critical to accurate citation.
Mandatory Fields

- **Archive and/or Distributor**—the organization distributing or caring for the data, ideally over the long term.

- **Locator/Identifier/Distribution Medium**—this could be a URL but ideally it should be a persistent service, such as a DOI, Handle or ARK, that resolves to the current location of the data in question.

- **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.
Author

• Data stewards should work with data providers to determine who gets credit and takes responsibility for the data set
• Small groups and organizations may be authors
  • Be as specific as possible for accountability and credit purposes

• A data set that is a collection of several smaller, independent data sets will not have an author
  • The individual data sets would have their own specific citations with author
  • The collection would likely have an editor or compiler, though:

Release date

• For a completed data set, the release date is simply the year of release.


• More precise dates can be used if needed to indicate exactly when the data became available and citable
Release date (continued)

• If detailed versioning information is lacking, try and capture when updates occurred.
  • For infrequently or irregularly updated data, list the first year of released followed by "updated" with the current update information.

• For an ongoing and regularly updated data set, list the first year of release followed by the last update.
Title

Best Practices

• Try to avoid having version or edition information in the title.
• Should not be the title of a project or a related publication.
• A data set should have an identity and title of its own.
Version

- Careful versioning and documentation of version changes are central to enabling accurate citation.
- Include version as part of the citation for any version greater than 1.
- It may be appropriate to track major and minor versions.

Archive and/or distributor

• The organization that maintains and manages the release or distribution of the data set.
  • Often implies responsibility for stewardship of the data set.
  • Often considered the data "publisher."
  • DataCite describes this role as:
    "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."
• May be appropriate to recognize a major sponsor of the data here.
Locator, Identifier, or Distribution Medium

• If there is one fixed distribution medium, include it - DVD or CD-ROM
• If the data is available on the internet or via multiple media a persistent locator is required
  • PURLs, ARKs, DOIs, Handles, or any other persistent equivalent will do
  • DOIs are favored by publishers
  • Thompson Reuter is working on including data sets with DOIs and possibly ARKs in the Web of Science
Locator, Identifier (continued)

• Best practices
  • Locators should point to a landing page for the data set
  • Do not include the name of an organization in the locator
  • Use the http form of the locator for human usability

Optional Fields

- **Subset Used** - The logical equivalent of citing a passage in a book

- **Editor, Compiler, or other important role** - Other roles may need to be recognized especially if there is no author

- **Archive or Distributor Place** - The city, state (if necessary), and country of the archive or distributor

- **Distributor, Associate Archive, or other Institutional Role**

- **Data within a larger work**
Subset used

• Data stewards should suggest reasonable ways to specify a subset of a data set
• Specifying a temporal and/or spatial subset
• Subsets based on sub-collections or representations of the data set
• In the future it it is conceivable that systems may be able to record the exact data that corresponds to a particular data access query
Editor, Compiler, or other important role

- **Editor** - a person or team who is responsible for creating a value-added and possibly quality-controlled product from the data.

- **Compiler** - a person who is responsible for compiling a product from the data albeit with minimal scientific or technical input.

Editor, Compiler, or other important role (cont.)

- Editors and compilers may often be responsible for a larger work that includes multiple data sets from different authors

Editor, Compiler, or other important role (cont.)

- Occasionally there may be both a compiler and editor as well as other roles

- Only acknowledge at most two roles in the citation

- Record other roles in metadata or documentation

Distributor, associate archive, or other institutional role

- There may be multiple organizational roles that need to be recognized
  - This can be a way to recognize a funding source


Data within a larger work

- Cite data sets within a larger work similarly to how a chapter in a book is cited