Arctic Data Center Training

Arctic Data Center

2018-02-23

Contents

Chapter 1

How this book was made

This is a book written in **Markdown**. To create your own book, the **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Arctic Data Center

The Arctic Data Center is the primary data and software repository for Arctic NSF Polar Programs. The Arctic Data Center was first funded by NSF in 2016, and currently serves as the repository of record for NSF-funded Arctic data. The Arctic Data Center builds upon a long history of NSF investments in data preservation for Arctic communities, starting first with CADIS in the International Polar Year (2007), and continuing with ACADIS which expanded its mission.

2.1 Strong partnerships for preservation

The NSF Arctic Data Center helps the research community reproducibly preserve and discover all products of NSF-funded science in the Arctic including data, metadata, software, documents, and provenance that link these in a coherent knowledge model. Key to the initiative is the partnership between NCEAS at UC Santa Barbara, DataONE, and NOAA's NCEI, each of which bring critical capabilities to the Center. Infrastructure from the successful NSF-sponsored DataONE federation of data repositories will enable data replication to NCEI, providing both offsite and institutional diversity that is critical to long term preservation.

2.2 Commitment to training

The Arctic Data Center conducts training in data science and management, both of which are critical skills for stewardship of data, software, and other products of research that are preserved at the Arctic Data Center.

Each year, our training and outreach staff provide hands-on training at Arctic research conferences and in dedicated training sessions targeting early-career and under-represented populations. Training and outreach will focus on effective means for long-term data management, following a curriculum being developed and refined by the open science community. We will also be offering a Data Science Fellowship Program, launching Fall 2017. Learn more about current offerings at arcticdata.io

2.3 Further Reading

For more on the Arctic Data Center, read our proposal summary online

2.4 Acknowledgements

Work on this package was supported by: NSF award #1546024 to M. B. Jones, S. Baker-Yeboah, A. Budden, J. Dozier, and M. Schildhauer.
Additional support was provided for working group collaboration by the National Center for Ecological Analysis and Synthesis, a Center funded by the University of California, Santa Barbara, and the State of California.

Chapter 3

NSF Standards and Policies

For current, up-to-date information on NSF's data policies as they apply to all NSF-funded research, see NSF: Dissemination and Sharing of Research Results. Please contact your NSF Program Manager if you have questions about what to submit or what is required for any particular award.

3.1 Who should submit?

3.1.1 Arctic Research Opportunities (ARC)

If funded by NSF under ARC, researchers are required to submit:
* Complete metadata and all appropriate data and derived products
* Within 2 years of collection or before end of award, whichever comes first.

Data sets from ARC-supported scientific research should be deposited in long-lived and publicly-available archives appropriate for the specific type of data collected (by default, the NSF supported Arctic Data Center at arcticdata.io, or others where appropriate). Metadata for projects, regardless of where they are archived, should be submitted to this Arctic Data Center for centralized access and discoverability.

For all ARC supported projects, see the NSF ARC-programs data policy, which includes the following conditions:
> Complete metadata must be submitted to a national data center or another long-lived, publicly accessible archive within two years of collection or before the end of the award, whichever comes first.
> All data and derived data products that are appropriate for submission (see exceptions below), must be submitted within two years of collection or before the end of the award, whichever comes first.

Please contact your NSF Program Manager if you have questions about what to submit or what is required for any particular award.

3.1.2 ARC Arctic Observing Network:

If funded by NSF under ARC-AON, researchers are required to submit:
- Complete metadata and all data
- Real-time data made public immediately
- Within 6 months of collection

For all ARC supported Arctic Observing Network projects, NSF also requires:
- Real-time data must be made publicly available immediately. If there is any question about what constitutes real-time data, please contact the appropriate NSF program officer.

- All data must be submitted to a national data center or another long-lived publicly accessible archive within 6 months of collection, and be fully quality controlled. All data sets and derived data products must be accompanied by a metadata profile and full documentation that allows the data to be properly interpreted and used by other researchers.

Please contact your NSF Program Manager if you have questions about what to submit or what is required for any particular award.

3.1.3 Arctic Social Sciences Program (ASSP):

If funded by NSF under the ASSP, researchers follow a different set of guidelines to submit:
- NSF policies include special exceptions for ASSP and other awards that contain sensitive data.

NSF policies include special exceptions for Arctic Social Sciences awards and other awards that contain sensitive data, including human subjects data and data that are governed by an Institutional Review Board policy. These special conditions exist for sharing social science data that are ethically or legally sensitive or at risk of decontextualization.

In these cases, NSF has requested that a metadata record be created that documents non-sensitive aspects of the project and data, including the title, contact information for the data set creators and contacts, and an abstract and methods description summarizing the data collection methodologies that does not include any sensitive information or data.

For sensitive social science data, resarchers should submit the following: * Metadata record that documents non-sensitive aspects of the project and data
+ Project Title
+ Data Tite
+ Contact information for creators
+ Abstract for dataset
+ Methods for collection and analysis of data

Please let us know when submitting your record that your data contains sensitive information so that we can adjust our review process accordingly.

Chapter 4

Data Management Planning

4.0.1 NSF Data Management Planning Tool for Arctic research

(Coming in 2018)

Chapter 5

Digital Object Identifiers and Licensing

5.1 DOIs

5.1.1 What is a DOI?

Learn more about digital object identifiers, or DOIs, at www.doi.org.

5.1.2 How does the Arctic Data Center provide DOIs?

Once data have been submitted to the Arctic Data Center, our metadata staff will review and provide suggestions for improvement. Once everything is set, we will make the data publicly accessible and publish it with a DOI. This will allow you and other researchers to cite the data set directly in NSF reports, publications, and other venues. The DOI is registered with DataCite using the EZID service, and will be discoverable through multiple data citation networks, including DataONE and others.

Once you have published your data with the Arctic Data Center, it can still be updated by providing an additional version which can replace the original, while still preserving the original and making it available to anyone who might have cited it. To update your data, return to the data submission tool used to submit it, and provide an update.

Any update to a data set qualifies as a new version and therefore requires a new DOI. This is because each DOI represents a unique, immutable version, just like for a journal article. DOIs and URLs for previous versions of data sets remain active on the Arctic Data Center (will continue to resolve to the dataset landing page for the specific version they are associated with), but a clear message will appear at the top of the page stating that "A newer version of this dataset exists" with a hyperlink to the latest version. With this approach, any past uses of a DOI (such as in a publication) will remain functional and will reference the specific version of the dataset that was cited, while pointing users to the newest version if one exists.

5.1.3 Data versioning with DOIs

Each update to the dataset has a unique identifier (in this case, a DOI). This allows researchers to cite the exact version of the data set that they used. Newer versions are clearly indicated.

5.1.4 DOI Summary

- We assign a DOI to each published data set
- Researchers should cite data they use
- We are working with DataCite to track the citations to data

5.2 Licensing at the Arctic Data Center

Chapter 6

Preparing a Submission

6.1 Preparing data

To prepare for upload, it's good to have your files in order. You might want to take a look at some best practices for managing your data files. For a given project, perhaps you have 6 data files, and one document that describes the methods that you used to collect or analyze your data. Collect these files into a single directory, and name them with short but descriptive names. Try to avoid spaces in your file names, but rather use dashes "-" or underscores "_".

6.1.1 File Formats

While the Arctic Data Center supports the upload of any data file format, sharing data can be greatly enhanced if you use file formats that can be used by a variety of softwares, including free or open source software.

6.1.1.1 File formats that support data sharing

File type	Proprietary format	**Preferred open data format**
Tabular	.xls	**.csv**
Text	.doc	**.txt**
Image	.psd	**.png**, **.tiff**, **.jpeg**, **.pdf**
Video	.wma	**.avi**
Audio		**.mp3**, **.aac**, **.ogg**
/	.matlab	**.NetCDF**
Geospatial	/	**.shp**, **.geojson**, **.kml**, **.wms**, **.wmt**
Archiving or Compression	/	**.tar**, **.zip**
Others	/	**.xml**, **.html**, **.css**, **.json**, **.yaml**

For instance, while Microsoft Excel files are commonplace, for tabular data your data is much more accessible as a Comma Separated Values (CSV) text file, which can be read on any computer without requiring Microsoft products. Data submitted in Excel workbooks will undergo conversion to CSVs by our staff before being made public. Other proprietary formats will also be converted to plain-text formats when possible. Likewise, for image files, use common formats like PNG, JPEG, TIFF, etc. Most all browsers can handle these. GIS data can be exported to ESRI shapefiles, and data created in Matlab or other matrix-based programs can be exported as NetCDF (an open binary format).

6.2 Why metadata?

Metadata helps davelopers of data, users of data, and organizations funding data creation in a variety of ways.

6.2.0.1 Metadata supports data developers

- Helps avoid data duplication
- What has been collected already?
- Communicates reliable information
- What method was used?
- What methods are in common use in my field?
- Publicizes your work (Hey, I made this!)
- Save time the next time (Hey, I've already done this!)

6.2.0.2 Metadata supports data users

- Makes relevant data findable
- Helps you easily evaluate if the data is suitable for your work
- Helps you understand if and how to actually use

6.2.0.3 Metadata supports organizations

- Avoids duplication in new work
- Maximizes funding dollars
- Good metadata can be a great advertisement for the organization
- Transcents people and time (data is not lost when researchers or labs leave the organization)

6.2.1 Researcher concerns when creating metadata

Even if the value of data documentation is recognzied, researchers are often concerned about the effort required to create metadata that effectively describes their data.

Concern	Solution
Workload required to capture accurate robust metadata	Incorporate metadata creation into data development process – distribute the effort
Time and resources to create, manage, and maintain metadata	Include in grant budget and schedule
Readability / usability of metadata	Use a standardized metadata format
Discipline specific information and ontologies	Use a standard 'profile' that supports discipline specific information

6.2.2 Authoring quality metadata

Overarching goal: A reasonable make sense of your data in 10 or 20 years without contacting you.

Think of metadata creation as “data reporting”
- Who created the data?
- What is the content of the data?
- When were the data created?
- Where are the data from?
- How were the data developed?
- Why were the data developed?

Other things to consider:
- When in doubt, be more specific. For example, spell out acronyms, use full names and full email addresses.
- Include as much information as possible directly in the metadata record.
- When creating metadata target multiple user groups, including those: looking directly for your data, who do not know about your data (but should), looking to scrutinize your work, looking to reproduce your work, looking to give you credit for your work.

6.3 Preparing your metadata

Gather together metadata that describes your data, including information about the name and identity of the data, the geospatial coordinates where it was collected, when it was collected, and by whom.

Things to keep in mind:
Title - A good title includes pertinent context, often describing the who, what, when, where, and, why of your data. Remember that the title is often the first way a potential user will evaluate your dataset.

Good Dataset Title	“Not So Good” Dataset Title
Greater Yellowstone Rivers from 1:126,700 U.S. Forest Service Visitor Maps (1961-1983)	River Data

Abstract - The data or dataset abstract is sistinct from scientific abstract. It should provide more context for the title and a high-level summary of methodologies, data formats, coverage, etc. used to create the dataset.
People - Name alone is not enough. To make sure that contributors and creators of the datasets are properly credited it is best practice to include the person’s email address, and, if possible, the person’s ORCiD.
ORCiD - Is a free, unique identifier for researchers that enables unambiguous reference to humans. ORCiD is becoming a norm in the research community, and it allows for journals and other organizations like DataCite, FundRef, Zenodo, and CrossRef to link researchers and their work in public and easily referenced ways. If you don’t have one, you can get an ORCiD at https://orcid.org

6.3.1 Metadata Standards

There are many metadata standards, make sure you choose a standard that fits your field of research and the nature of your dataset. Below are a few widely recognized metadata standards. The Arctic Data Center uses the Ecological Metadata Langauge (EML) standard.

- Dublin Core (emphasizes publications)
- Darwin Core (emphasizes collections)
- FGDC (emphasizes spatial)
- ISO19115 (emphasizes spatial & services)
- Ecological Metadata Language (general, but emphasis on filesystem artifacts, attributes, taxonomy)

6.3.2 Contributors

For people, you'll want to have their names and contact information, and an ORCID identifier for them. You'll want to have a good complete set of text describing the methods used to collect the data, as well as experimental design and sampling layouts. Finally, you'll need the data files themselves. Once you've gathered this information, choose a data submission tool and get started!

6.4 Data Submission Methods with the Arctic Data Center

There are multiple ways to submit data package to the Arctic Data Center. If you choose to use R or MATLAB (details below), remember that EML Metadata still needs to be created separately or by using the limited EML R library.

6.4.1 Data Editor

(launching 2018)

6.4.2 R Data Submission

6.4.3 MATLAB Data Submission