Ice Core Community Data Standards

Andrei Kurbatov[1]

[1]Climate Change Institute & School of Earth and Climate Sciences
University of Maine

February 15, 2022

Contents

1 *Introduction*

Over the last three decades, internet-based technologies have modified many aspects of how research data are shared. The development of new information networks has revolutionized how research data are collected, stored and distributed. It is now possible to publish research results and ensure comprehensive access to all underlying data. Several new initiatives have been facilitated by the recognition that open, unrestricted access to meaningful research data provides an additional quality control mechanism, extends data usability into different disciplines, and ensures equal access to resources for all investigators.

> - 17th Century Open Science meant publishing articles describing science results
> - 21st Century Open Science means "sharing" all science results: articles, data, software, workflow, etc.
>
> George O. Strawn, Plenary talk at Virtual SciDataCon 2021.

The Open Research Data (ORD) policy term and idea was embraced by the National Science Foundation (NSF) as a crucial part in fostering scientific discoveries[8]. In 2013 the Office of Science and Technology Policy (OSTP) directed major U.S. Federal agencies "to develop a plan to support increased public access to the results of research funded by the Federal Government." On May 9, 2013, President Obama signed an Executive Order (EO) M-13-13 entitled "Making Open and Machine Readable the New Default for Government Information," which requires all Federal agencies to comply with a new Open Data Policy. One of the requirements was that "...any datasets in the agency's enterprise data inventory that can be made publicly available must be listed at www.[agency].gov/data in human– and machine–readable format."

The Findable, Accessible, Interoperable and Reusable (FAIR) data term and Guiding Principles [9] were introduced in parallel to the concept of open data. While both these terms are overlapping they have different meanings: "Open data can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike." In the "Earth in Time" document, the National Academies of Sciences, Engineering, and Medicine [7] in 2020 recommended that: "the Division of Earth Sciences at the NSF should develop and implement a strategy to provide support for FAIR practices within community-based data efforts."

Currently there is a requirement for recipients of many U.S. Government agencies research awards to timely share all data sets generated by funded projects. Most research proposals also have to include a data management plan that identifies the type of data the project will generate and how data sets will be archived, under which conditions (data use licenses), and when (data set release date) data will be publicly shared. It is also a requirement

at the end of the project cycle, in the final report, to include information on a long term data repository where all project generated data will be publicly available. In addition, most leading universities, research groups, and publishers also independently endorse ORD management and distribution practices.

This review briefly summarizes the current state of ice core data sharing and introduces possible steps that will help to make the ice core data sets better aligned with the principles of ORD policies and drive future innovations and discoveries of US ice core research. In the future, standardization of ice core data collection, recording, and distribution will simplify and semi automate (with minimum human intervention) access to data generated by various research teams and ultimately will benefit collaboration among different research communities and foster the future generation of multidisciplinary paleoclimate data products (e.g., [5]) that is crucial for further understanding of complex, non-linear climate system operation, thresholds and forcing.

2 *The current state of research data sharing*

Currently there are several initiatives relevant to the ice core community that are leading the global effort to open and improve access to research data. While these initiatives are aligned with the ORD policies, some endorsed details on data access and reusability are varied.

- The American Geophysical Union (AGU) has an online portal dedicated to Data & Software for Authors. It has an information resource page on Open and FAIR Data and Software and its Data Leadership web page summarizes how current and future activities within the AGU help to reach out to other scientific communities and the public.
- In the U.S., several government agencies commonly fund ice core research: e.g., the NSF, National Aeronautic Science Administration (NASA) and National Science Foundation (NOAA). All these agencies embrace the FAIR Guiding Principles for scientific data management and stewardship and have dedicated web sites and documents related to this topic. For example, since 2016 the NSF Office of Polar Programs (OPP) has had a data policy document, NSF 16-055, that requires funded investigators to release all project generated datasets.
- The United States Geological Survey (USGS) is about to publish a report with 88 recommendations developed during a 3-day 2019 workshop [3].
- The European Geophysical Union (EGU) has a slightly different approach that evolved from an "open access" to research results (scientific publications) initiative. See the latest Open Access 2020 document here. The "open access" initiative is aiming to make access to all research results and underlying data open and transparent. Recently, to

provide all European researchers, innovators, companies and citizens with a federated and open multi-disciplinary environment, a new European Open Science Cloud (EOSC) initiative started.

- Institutional examples would be the Open Access Initiative at Berkeley, community supported DRYAD data sharing repository, the Harvard Dataverse platform, and Digital Commons platform from bepress
- The NASA and the European Space Agency (ESA) provide open access to collected data via the EarthData portal.

The presented list of examples is definitely incomplete but it shows an ongoing adaptation of ORD by research communities and makes a strong case that more and more research data are becoming publicly available globally. While it is relatively intuitive to support FAIR ideas and to some degree implement data sharing protocols by a single investigator or small research teams working in specific research areas, multiple data sources and standards could relatively quickly become confusing, especially for investigators new to ice core science. It is not always intuitive or simple to use or synthesize data produced by different projects and teams. Another issue could arise when shared data sets are not fully documented (e.g., not all steps in method descriptions are documented) or only post-processed data are available (e.g., ice core data resampled to an arbitrary time scale used at the time of publication with very limited information on how the time scale was developed or how data were resampled). Lastly the key words: "best practices and standards for open data", the "findability', and"datasets reusability" could have different meaning for different investigators in the same community or when the understanding of what is important to capture or record is evolving along with data collection. It is not uncommon that some valuable information is not properly captured or recorded (e.g., physical ice core properties such as density, melt layers were not recorded before sampling or melting for gas or glaciochemical work).

It is time to review the existing state of ice core data collection by the the US ice core community, evaluate data reporting practices used in different laboratories, and establish a set of recommendations that will improve the current data collection and archiving ecosystem.

Several very minor, incremental improvements would help to guide individual research teams on possible standards for collection, organization, and distribution of ice core data. The goal is to share meaningful datasets that promote data reusability and, if needed, replication of research results and also *minimize burden* on the specific investigator to reformat data or fill out peculiar information before submission to a data repository or research journal.

3 *Currently established ice core data sharing practices*

The US ice core research community has a long tradition of distributing ice core data sets with publications.

In the mid 1990's, participants of the US ice coring program made pioneering efforts to share the GISP2 data set with the public. Most of the data were reported in tabulated form and accompanied by a data description and list of relevant publications. The data set was published by the University of Colorado, NSIDC User Services, CIRES, in 1997. Currently, the data set is not easy to find but it is available from NOAA's National Centers for Environmental Information (NCEI) web site. A copy of the original CD-ROM [1] archived in a single zip file is also available from the PANGAEA data repository and a copy of the GISP2 project notebooks is hosted by the University of Maine.

While it is great that multiple repositories are hosting this dataset, it could be confusing for a new investigator to navigate GISP2 original and new data or use new data sets or adapt the best version of the time scales.

The number of data repositories that currently archive and distribute ice core data is increasing.

- The NSF hosts the Open Data at NSF web site. A search on the U.S. Government open data portal for "ice core" returned 813 datasets, while the data consolidator site Climate.gov returns "No documents found" for the same query.
- The NSF Arctic Data Center ADC holds data sets from the Arctic but occasionally Antarctic data sets are also archived there.
- The U.S. Antarctic Program Data Center (USAP-DC) currently is the go to place for ice core data sets collected within the last decade.
- NOAA's National Centers for Environmental Information (NCEI) stores a large collection of ice core data and derived paleoclimate data products.
- The National Snow and Ice Data Center (NSIDC) archived a number of older ice core project data sets.
- The European Earth & Environmental Science community is storing most ice core data at PANGAEA web-based portal. Some US ice core projects are also archived there.
- The research centers, universities and large ice core projects also maintain repositories for sharing data, methods and frequently updated or legacy documents that are not normally archived at the data centers.

- IsoLab UW
- CCI, UM
- Centre for Ice and Climate, Niels Bohr Institute
- Byrd Polar, OSU
- South Pole Ice Core (SPICEcore) project
- The West Antarctic Ice Sheet (WAIS) Divide ice core project

4 *Ice core data standards*

Funding agencies, research journals and other open access data portals have different data and metadata standards stored in multiple and isolated databases. Several data products are using ice core data or sharing ice core related metadata, for example:

- QGreenland [6]
- Quantarctica [4].
- Berkeley Earth
- Commercial Data Observation Network for Earth DataONE
- Originally funded by Scientific Committee on Antarctic Research (SCAR) a list of ice cores that have been collected as part of the International Trans-Antarctic Science Expedition (ITASE) and other initiatives are available from Ice-Reader database
- The NSF-ICF inventory page has a lot of very useful information about ice core projects and site coordinates, in addition to a list of samples that are currently available.
- The National Science Foundation National Ice Core Facility (NSF-ICF) has an inventory of ice cores and samples.

Over the years, the ice core community has adapted some "soft" ice core site naming and data distribution conventions (e.g., WDC-06A or SPC-14) but there is no single comprehensive list of all major ice core projects that is approved by the community or has been evaluated by the peer review process. The most comprehensive list of ice cores is available from the Wikipedia Ice core page but it is not maintained by the research community. For example, it is common to see the abbreviation WD used for the West Antarctic Ice Sheet (WAIS) Divide project deep ice core, or SPICEcore for the 1751 meter deep South Pole ice core (SPC-14) recovered in January 23, 2016 in reports, maps and publications. During the active phase of both WAIS Divide and SPICEcore projects, Science Management Office (SMO) teams would recommend the use of project accepted standards, but after the end of the project,

new investigators that work with samples from the NSF-ICF do not always follow these past SMO recommendations.

While experienced investigators can relatively easily navigate the lack of standards in ice core naming and data/metadata reporting, the situation creates an unnecessary barrier for new investigators, researchers from other disciplines, the general public and potentially impacts the quality of future data synthesis.

5 *Data discoverability*

The significant realization that it is important to make publicly available data discoverable led to the creation of Antarctic Master Directory (AMD). The AMD collection holds more than 7700 Antarctic dataset descriptions from 25 countries. The EARTHDATA portal SNOW/ICE managed by NASA allows one to search for specific data sets using Global Change Master Directory (GCMD) keywords. Several other organizations are trying to improve data discoverability, e.g., DataCite organization or Commercial Data Observation Network for Earth DataONE.

Currently there is an ongoing effort to assign Digital Object Identifier (DOI) to published papers and all other research products:

- Samples, using the International Geo Sample Number (IGSN) IGSN.
- Methods, for example UMaine CCI Stable Isotope Laboratory procedures posted on zenodo: https://doi.org/10.5281/zenodo.4721044
- Research communities, for example the International Ocean Discovery Program (IODP) page on zenodo server.
- Funding agencies, for example NSF
- Data sets IODP Expedition 361 Section summary

- Ice core metadata search does not prioritize and organize results. Some additional steps are required to select and use "correct" data.
- GCMD keywords do not fully reflect the needs of ice core science, and are very confusing to new investigators.
- While a lot of data are stored in numerous data centers, it is not a trivial task to create a simple collection of all data generated from a single ice core (e.g., WDC-06A) using publicly available data or to maintain an automatically updated list of measurements from generated data sets for specific ice cores using automatic data retrieval tools.

6 *Data formats*

Currently a wide range of digital data formats are utilized for ice core specific data. Most common formats are listed below but there are more proprietary or rare formats that the community has been using over the years,.

- Spreadsheet specific files: e.g., MS Excel (XLS, XLSX), Open Document Format (ODS).
- Comma Separated Values (CSV) format is relatively simple and widely accepted but is not easy to use with complicated and multidimensional data sets. For example it is not easy to organize in a single table, ice core location, time scale and ice core data files.
- Binary MATLAB® MAT-files
- Network Common Data Form (NetCDF or HDF5) is a framework of libraries and self-describing data formats.
- Linked Paleo Data (LiPD) is a data format relatively unfamiliar to the ice core community [5, 2]. It is well developed but does not fully capture ice core domain specifics and needs new software libraries and tools to beused by the ice core community.

7 *Programmatic access to ice core data*

In recent years most data repositories have provided various ways to access data products and metadata programmatically using Application Programming Interface (API). API allows

access, download and work with data programmatically from computer programs. The following web pages are developed by data centers to help users with the API interfaces.

- NSIDC.
- USAP-DC API.
- Arctic Data Center uses DataONE REST API and encodes science metadata in the Ecological Metadata Language (EML).
- NCEI has a user documentation page for supported Data Service API and a separate page for the NCEI Paleo Web Service

All data centers listed above use different formats and API protocols but with the right software tools all data sets are potentially reachable programmatically.

With a set of standards and recommendations and next generation software tools, ice core or cross-disciplinary data users would have a relatively easy time to access and retrieve all required information from these data centers.

8 *Preliminary conclusions*

The information summarized in this document is an attempt to briefly outline the current state of US ice core research community practices, standards and mechanisms of data sharing. The situation is quite dynamic and many very positive signs are clearly visible. Community and data centers are trying to use standards that are evolving and more and more data sets are becoming publicly available.

9 *Items to consider in the future*

Several items listed below should be also considered so changes in data sharing mechanisms will not negatively impact:

- Equality in data access
- Data ownership

10 *Acknowledgement*

This document incorporates some of the preliminary results from the "NSF Summer 2019 workshop: Computing Arctic Data" that took place in May 2019 at the University of Maine. The workshop was sponsored by the NSF award 1848747.

Credits Cover page phtoto was taken by Andrei Kurbatov at Tupungatito volcano, Chile.

Bibliography

[1] GRIP/GISP and GRIP Members. *Greenland Summit Ice Cores CD-ROM as zip-archive*. data set. 2017. DOI: 10.1594/PANGAEA.870454. URL: https://doi.org/10.1594/PANGAEA.870454.

[2] Christopher Heiser and Nick McKay. *LiPD.net*. Version 1.0.0. Apr. 2018. DOI: 10.5281/zenodo.1218057. URL: https://doi.org/10.5281/zenodo.1218057.

[3] Leslie Hsu. *U.S. Geological Survey Community for Data Integration 2019 Workshop Proceedings—From big data to smart data*. Report. Reston, VA, 2021. DOI: 10.3133/ofr20201132. URL: http://pubs.er.usgs.gov/publication/ofr20201132.

[4] Kenichi Matsuoka et al. "Quantarctica, an integrated mapping environment for Antarctica, the Southern Ocean, and sub-Antarctic islands". In: *Environmental Modelling & Software* 140 (June 2021), p. 105015. DOI: https://doi.org/10.1016/j.envsoft.2021.105015.

[5] N. P. McKay and J. Emile-Geay. "Technical note: The Linked Paleo Data framework a common tongue for paleoclimatology". In: *Climate of the Past* 12.4 (2016), pp. 1093–1100. DOI: 10.5194/cp-12-1093-2016. URL: https://cp.copernicus.org/articles/12/1093/2016/.

[6] T. Moon et al. "QGreenland (v1.0.1)". [software]. Available from https://qgreenland.org. https://doi.org/10.5281/zenodo.4558266. 2021.

[7] National Academies of Sciences, Engineering, and Medicine. *A Vision for NSF Earth Sciences 2020-2030: Earth in Time*. Washington, DC: The National Academies Press, 2020. ISBN: 978-0-309-67600-7. DOI: 10.17226/25761. URL: https://www.nap.edu/catalog/25761/a-vision-for-nsf-earth-sciences-2020-2030-earth-in.

[8] NSF. *Public Access Plan: Today's Data, Tomorrow's Discoveries: Increasing Access to the Results of Research Funded by the National Science Foundation*. Tech. rep. nsf15052. The National Science Foundation, Mar. 2015.

[9] Mark D. Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016), p. 160018. DOI: 10.1038/sdata.2016.18.

11 *Acronyms*

API Application Programming Interface.

NOAA National Oceanic and Atmospheric Administration

AGU American Geophysical Union

ADC The NSF Arctic Data Center

AMD Antarctic Master Directory

DOI Digital Object Identifier

GCMD Global Change Master Directory

EGU European Geophysical Union

EO Executive Order

EOSC European Open Science Cloud

EU European Union

ESA European Space Agency

FAIR Findable, Accessible, Interoperable and Reusable

GISP2 Greenland Ice Sheet Project 2

IGSN International Geo Sample Number

IGY International Geophysical Year

IODP International Ocean Discovery Program

ITASE International Trans-Antarctic Science Expedition

NASA National Aeronautic Science Administration

NASEM National Academies of Sciences, Engineering, and Medicine

NCEI National Centers for Environmental Information

NOAA National Science Foundation

NSF National Science Foundation

NSF-ICF National Science Foundation National Ice Core Facility

OPP Office of Polar Programs

OSTP Office of Science and Technology Policy

ORD Open Research Data

SCAR Scientific Committee on Antarctic Research

SMO Science Management Office

USGS United States Geological Survey

WAIS West Antarctic Ice Sheet

WDC World Data Centre

WWW World Wide Web