




The Role of Metadata and Vocabulary Standards in Enabling Scientific Data Interoperability: A Study of Earth System Science Data Facilities

Matthew S. Mayernik, National Center for Atmospheric Research, University Corporation for Atmospheric Research, Boulder, CO, US, mayernik@ucar.edu 

Yauheniya Liapich, University of Denver, Denver, CO, US

Abstract

Objective: Journal publishers within many sciences are increasingly expecting data to be deposited into repositories that support the FAIR principles. Data repositories are thus needing to determine what implications the FAIR principles have on their existing services and systems. Metadata standards and controlled vocabularies are specifically called out as core components of the FAIR principles related to interoperability.

Methods: This paper looks specifically at the ways that metadata standards and controlled vocabularies are used by Earth system science data repositories. Data sets from 55 data facilities were examined to determine which metadata standards and controlled subject / keyword vocabularies were used.

Results: The findings indicate that only the ISO 19115:2003 and DataCite metadata standards are used by more than 40% of the data facilities, and the NASA Global Change Master Directory (GCMD) keywords are the only keyword vocabulary of broad use within this community.

Conclusions: These findings raise questions about the extent to which metadata standards and keyword vocabularies can facilitate interoperability beyond narrow sub-sections of the data facility communities. This study also points to systematic challenges related to migration to new standards.

Received: April 29, 2022 **Accepted:** August 18, 2022 **Published:** December 19, 2022

Competing Interests: The authors declare that they have no competing interests.

Data Availability: All data underlying this study are available at: Liapich, Y. & Mayernik, M.S. (2021). Investigation of metadata standard use by geoscience data repositories [data set]. UCAR/NCAR - DASH Repository. Version 2.0. <https://doi.org/10.5065/z9ch-wk24>

The *Journal of eScience Librarianship* is a peer-reviewed open access journal. © 2022 The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution 4.0 International License (CC-BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

See <https://creativecommons.org/licenses/by/4.0>.

∞ OPEN ACCESS

Introduction

The FAIR principles have become a point of reference for data-focused activities in the academic fields, stating that the downstream benefits of data increase when data are made Findable, Accessible, Interoperable, and Reusable (GO FAIR 2021). These principles were developed by an international collaboration dedicated to reducing data discovery and reuse challenges, with the goal of facilitating scientific discovery based on publicly accessible data collections (Wilkinson et al. 2016). The principles have been adopted by scholarly publishers, academic institutions, and information and data repositories.

Journal publishers are increasingly expecting data to be deposited into repositories that support the FAIR principles (Stall et al. 2019). Data repositories thus need to determine what implications the FAIR principles have on their existing services and systems. Determining if a repository is FAIR-compliant is not always straightforward. “[T]he FAIR guiding principles on their own are unlikely to lead to responsible forms of data sharing. Although they provide a much-needed step forward for furthering the cause of data stewardship, they do not provide a complete set of guiding principles for improving data-driven science” (Boeckhout, Zielhuis, & Bredenoord 2018, pg. 935)

Data repositories need to interpret the FAIR principles as appropriate for their local circumstances and their “designated community,” to use the terminology of the Open Archival Information Systems (OAIS) reference model (Donaldson, Zegler-Poleska, & Yarmey 2020). Local interpretation of the FAIR principles leads to diverging implementations across communities, as well as between repositories within the same community (Jacobsen et al. 2020). For example, scientific data repositories and library-based data archives may interpret the principles differently (Mannheimer, Sterman, & Borda 2016).

This paper investigates how the FAIR principles focused on metadata standards and controlled vocabularies manifest within Earth system science data repositories. As described by Srinivasan et al. (2009, pg. 268): “Two of the most important kinds of decisions taken by object record creators in today’s [information systems] are (i) the choice of a metadata schema—i.e., the selection of the categories, facets, metadata elements, attribute-types, or fields ... that collectively make up a record, and (ii) the choice of vocabularies—i.e., the selection of the term sets or value sets from which are drawn the values that are assigned to given fields in given records.”

These topics are both directly referenced by the interoperability-related FAIR principles, specifically in sub-principle I1, “(meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation,” and I2, “(meta)data use vocabularies that follow FAIR principles.” A prior cross-disciplinary study of 40 data facilities found that the interoperability-related FAIR principles were inconsistently implemented, with slightly more than half of the repositories (59%) being compliant with Principle I1, but no repositories being compliant with Principle I2 (Dunning, De Smaele, & Böhmer 2017). It is well known that metadata and vocabulary use vary widely when looking across disciplines (Willis, Greenberg, & White 2012). The goal of this paper is to examine consistency of metadata schema and subject vocabulary use within

a specific community of facilities. We would expect that metadata and subject vocabularies will manifest differently across disciplines, but how will they manifest within a fairly narrowly defined community of like repositories?

Our study was organized around the following guiding questions:

- How do metadata standards, vocabularies, and data access modes differ between data facilities within a specific community of data facilities?
- What are the implications of those differences for access, interoperability, and reuse of data?

Background

The FAIR Principles emerged from discussions over the past few decades about the potential benefits of comprehensive data and metadata stewardship for scholarly research. The FAIR Principles are not like the Open Archival Information System (OAIS) reference model, which has been used as a technical architecture and a common set of terminology for data repositories since the 1990s (Lee 2009). FAIR is also not a repository certification such as the CoreTrustSeal (<https://www.coretrustseal.org>), which involve formal assessment by outside reviewers according to specific actionable criteria. The FAIR Principles are instead intended to be guidelines that enable data and associated research materials to be used and reused by others.

Metadata are central to many of the FAIR principles. Metadata can be evaluated with regards to its correctness, completeness, and consistency (Bugbee et al. 2021). Metadata errors and interoperability failures can occur across all three of these categories (Arms et al. 2002; Kervin, Michener, & Cook 2013).

Metadata standards cannot fully specify all details of their interpretation. Every repository must make interpretive decisions when implementing particular metadata standards (Feinberg 2017). Community-vetted recommendations can facilitate more consistent interpretations of metadata standards (Gordon & Habermann 2018), but they can never fully eliminate the interpretive flexibility inherent in standards implementation.

There are different kinds of metadata standards, including structure, value, content, and interchange standards (Gilliland 2008). Standards are also often used in combination, and are released, revised, and re-released in iterative fashion over years or decades. For example, many libraries, though not all, have migrated their content standard from the Anglo-American Cataloging Rules 2nd edition (AACR) to the Resource Description and Access (RDA) standard (MacLennan & Walicka 2020). Most, however, still use Machine Readable Cataloging (MARC) format created in the 1960s to store and exchange catalog records. Finally, “metadata” itself can encompass information of varying kinds, including formal documents structured according to particular structure or interchange standards, like MARC, XML, or JSON, or informal and unstructured descriptive text posted on data facility web sites (Mayernik 2020). Metadata may

be downloadable as discrete files, displayed in html tables, or served up via APIs or web services. Our study is thus targeted broadly at understanding the Earth system science data facility metadata landscape.

Methods

To understand the data facility landscape within Earth system science, we examined data facilities from two groups: the Council of Data Facilities (CDF) and the Earth Science Information Partners (ESIP). The CDF was created in 2014 under the auspices of the US National Science Foundation's (NSF) EarthCube program to facilitate coordination and collaboration among Earth system science data facilities (EarthCube Council of Data Facilities 2018). ESIP was created in 1998 by the National Aeronautics and Space Administration (NASA) to facilitate more effective production and distribution of Earth science data. Since then, it has expanded into a community backbone organization that enables collaboration on Earth science data and information issues (Robinson et al. 2019).

Data facilities must apply to become members of the CDF and ESIP. As such, they together encompass self-identified data facilities within the Earth system sciences in the US, and provide a good population with which to gain insight into the community trends of metadata standard and vocabulary use.

Repository selection

Council of Data Facilities Members

The CDF members are grouped into four categories (EarthCube Council of Data Facilities 2018). We included CDF categories A, B, and C in our study. Category A consists of NSF-funded not-for-profit or academic data facilities. Category B includes Federally Funded Research and Development Centers (FFRDCs) and other federal, state, and local data facilities. The data facilities in category C are international, private, and other not-for-profit or academic data facilities.

We excluded CDF category D organizations, which are called "Associate Members." This category includes professional societies, publishers, and other organizations that do not do data collection or distribution. We used the list of CDF members found at <https://www.earthcube.org/cdf-members>.

ESIP Members

ESIP also contains members of multiple categories, called Types I-V. Categories included in our study were Type I (distributors of satellite and ground-based data sets), Type II (providers of data and information products, technologies or services aimed at the Earth science communities), and Type IV (ESIP financial sponsors). We excluded the Type III category because it contains commercial and non-commercial organizations that develop tools, not datasets or metadata. We also excluded Type V organizations, non-voting financial supporters of ESIP, because these organizations are also unlikely to have data or metadata that is of interest for this study. We used the list of ESIP members found at <https://www.esipfed.org/partners>.

Data Collection Process

The total population of data facilities that were initially examined across CDF and ESIP categories was 149: 23 CDF members, 110 ESIP members, and 16 that were members of both. Although all CDF and ESIP members examined can be called “data facilities,” their scope, target audiences, and collections varied considerably. Many are hosts of distinct data collections, but some are geoscience research organizations, or data/metadata aggregators, such as the Data Observation Network for Earth (DataONE) facility. Not all facilities expose Application Programming Interfaces (APIs) or metadata catalog web services. API and web service information was very difficult to find and to understand for many facilities, and thus were not used for this study.

For each facility, we followed links from the CDF and ESIP lists, or if no link was included, we attempted to find a web site using Google. Once a web site was found, it was checked for the presence of data sets. Upon conducting this initial examination, many CDF and ESIP members were excluded from this study for one or more of the following reasons: 1) we were unable to find a working website for the facility, 2) the facilities did not host data or metadata on their website or did not provide a way to discover their data, and 3) it was unclear how the facilities hosted data, as many of the facilities provided links to external repositories or did not enable searching on their website directly. Ninety-four data facilities were excluded for these reasons.

The main analysis focused on 55 facilities. We examined a couple of these facilities in more detail, as discussed below, to gain some insight into metadata schema and controlled vocabulary usage within a single organization. For each facility, we performed a high-level examination of several data sets to determine the degree of consistency of the information that was provided for each data set. This typically involved reviewing 5-10 data sets in a comparative fashion, to ensure that the same metadata options and keyword displays existed. Once we determined that we were looking at a consistent collection, we examined in detail one representative data set. The data sets that were chosen were generally recent data sets (when it was possible to determine this) because we were interested in current repository practices. Otherwise, they were chosen randomly to represent the larger collections from which they came.

This analysis was based on the information shown on the data set “landing pages,” meaning the most comprehensive representation of that data set. In some cases, it was difficult to clearly identify a “landing page” for a data set.

For each data facility, once a representative data set was chosen, the link for that dataset was noted in our data collection sheet. We then examined the data set’s landing page for links, files, and display of metadata. The available metadata was then checked to determine which metadata schemas were being used and the presence of any subject/keyword terms, along with any vocabulary used. We also tracked whether DOIs were present.

The data set that underlies this project is available at (Liapich & Mayernik 2021).

Results

We first outline some noteworthy characteristics of the data facilities, namely, their extent, web site structures, and data set landing pages. We then present the results of the investigation of the usage of metadata schemas, controlled vocabularies, and DOIs.

Data Facility Characteristics

The data facilities we examined vary considerably in the size and complexity of their data holdings, metadata catalogs, and web sites. Figure 1 provides a breakdown of their relative extents, based on the number of data sets that were presented on their web sites. While defining a “data set” in a consistent way is difficult (Renear, Sacchi, & Wickett 2010), we attempted to determine discrete data sets based on their logical and structural connectedness. Individual data sets often contained multiple files, up to millions of files for large satellite data sets. The facility sizes ranged from three data sets held by the Crustal Dynamics Data Information System, to almost 13 million records listed in the USGS ScienceBase catalog. The median number was 2,072 data sets. All but 12 of the 55 data facilities had more than 500 data sets available.

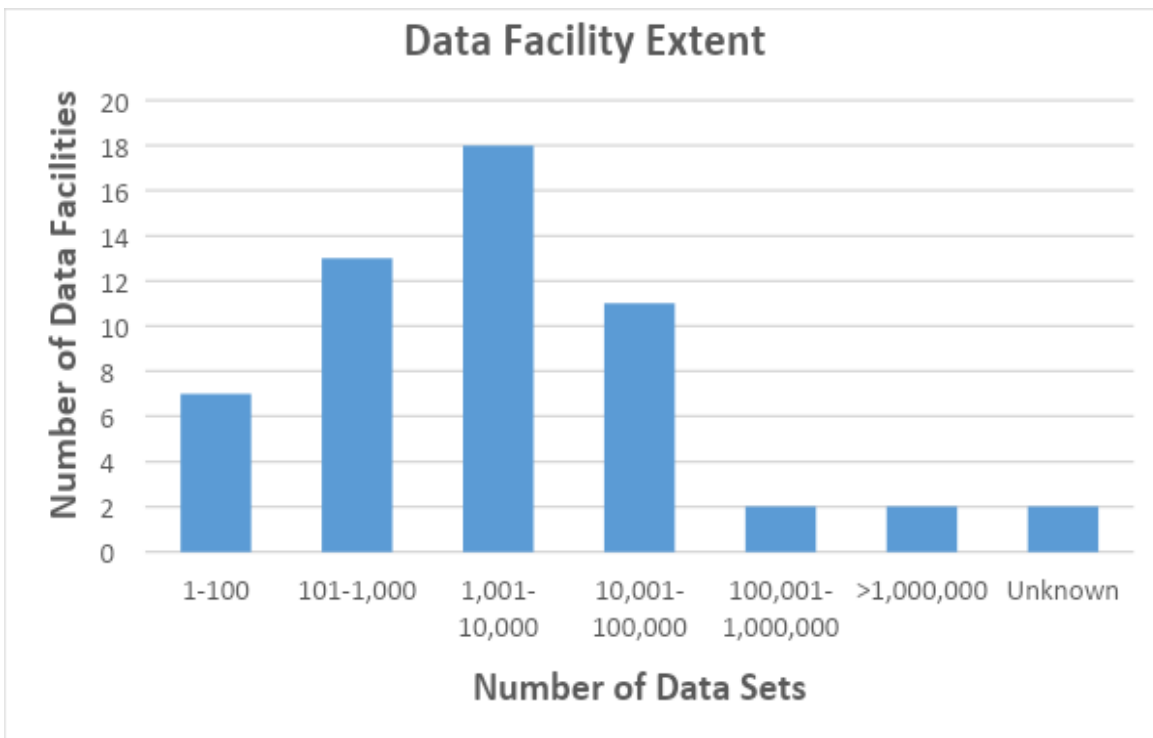


Figure 1: Number of data sets per data facility

As we began our investigation, we quickly encountered considerable variation between the web interfaces that host and enable discovery of data. Some facilities present simple Google-like search systems. Other facilities first present a map interface, and individual data sets are selected from filters. Other facilities require the user to select a particular project or collection, and then enable searching or browsing within individual collections. Likewise, the “landing pages” for individual data sets varied considerably. Facilities also varied in if/how they presented any structured metadata via XML or JSON, either as downloads or linked pages, and in whether they presented metadata via a single standard or via multiple standards.

Metadata Standard Usage

Table 1 shows the high-level results of the metadata standards examination. The International Standards Organization (ISO) 19115:2003 standard is the most widely used metadata standard among these data facilities, being used by 23 of 55 repositories (42%). Five metadata standards were used by 7-8 repositories (12-15%). Several these were used by the same group of NASA-associated data facilities, such as the Atom specification, the Directory Interchange Format (DIF), and the native NASA Common Metadata Repository (CMR) format. Seven other metadata standards were only used by one or two facilities.

For 20 of the 55 (36%) data sets we examined, no standardized metadata schema was mentioned at all. In Table 1, these are listed as “unknown.” These facilities may display extensive metadata on the data set landing pages, or as separate downloadable documents, but do not present any indication of standards. We labeled these as “unknown” rather than “none,” since it is possible that some of these facilities may in fact be structuring their metadata according to some standard that was not readily apparent for our examination. Thus, the “unknown” label in Table 1 should be interpreted as “it is unknown and unclear if any schema is being used,” and not as “this facility is using a schema but we cannot determine what it is.”

The data facilities that use the Ecological Metadata Language (EML) come in two types. Five data facilities in this group have a primary focus on ecologically-related research, including DataONE. Two other organizations have deployed their own versions of the DataONE web interface, specifically the Arctic Data Center and the Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) facility. Both have some ecology-related data, as well as data related to other sciences.

Notably, while ISO 19115 standard was seen the most commonly among all facilities, in nearly all cases, the version that is being used by data facilities is the 2003 version of the standard. ISO 19115:2003 has since been superseded by an updated version, ISO 19115-1:2014 (Brodeur et al. 2019). The newer version of the ISO 19115 standard (ISO 19115:2014) was only observed once, on the site for the Geoscience Australia data facility. They also link to a Geoscience Australia recommendation for how to implement the ISO 19115-1:2014 standard (Bastrakova 2018).

Table 1: Results of the metadata schema investigation

Metadata Standard Name	Number of facilities that used the standard
ISO 19115:2003	24
Unknown	20
Native NASA Common Metadata Repository	8
Atom	8
Directory Interchange Format	8
Ecological Metadata Language	7
FGDC Content Standards for Digital Geospatial Metadata	7
Data Catalog Vocabulary	2
Metadata Object Description Schema	2
ECHO 10	2
EPA Metadata Technical Specification	1
ISO 19115-1:2014	1
Schema.org	1
UMM-G	1

Keyword Usage

Table 2 shows the results of the examination of keyword / subject term vocabulary usage within geoscience data facilities. The NASA Global Change Master Directory (GCMD) keyword vocabularies are the most widely used, with 18 data facilities using GCMD (33%). No other controlled vocabulary or thesaurus saw significant usage within this group of facilities. 18 vocabularies were used by just a single facility. For the 11 data facilities listed as “None” in Table 2, it was not possible to find any keywords or subject terms associated with the data.

Seventeen facilities were coded as using “custom” vocabularies, meaning that we were unable to determine if the keywords or subject terms were drawn from a particular vocabulary. For example, the data set we examined from the Arctic Data Center dataset present terms from a custom controlled vocabulary, including topical terms or geographic names, such as “bathymetry” and Alaska. But some were vague, such as “T6” and “lake data.” The particularity of these keywords gives them limited usefulness for individuals who are not familiar with the area of study. Unique vocabularies can have benefits with regards to specificity and

accuracy, but they have limited reusability, as they are not transferrable to the other settings, environments, and facilities.

Table 2: Results of the subject term / keyword-controlled vocabulary investigation

Controlled Keyword Vocabulary Name	Number of facilities that used the vocabulary
Global Change Master Directory (GCMD)	18
Custom	17
None	11
Long Term Ecological Research (LTER)	2
ISO 19115 Topic Category	2
Library of Congress Subject Headings (LCSH)	1
Andrews Experimental Forest	1
Environmental Protection Agency (EPA) place names	1
Geonames	1
Government of Canada Core Subject Thesaurus	1
ANZRC Fields of Research	1
Metadata Service Keyword Thesaurus	1
IGS Metadata Keyword Thesaurus	1
IEDA data type categories	1
Marine-Geo Digital Library (MGDL)	1
MGDL Data Type vocabulary	1
MGDL Device Type vocabulary	1
National Agricultural Library Thesaurus	1
Ag Data Commons Keywords	1
NSIDC DAAC	1
Marine Realms Information Bank	1
USGS Thesaurus	1
USGS Scientific Topic Keyword	1

DOI Assignment/Usage

The presence of the DOIs is important to our research because they highlight the theme of Findability in the context of the FAIR principles. DOIs for data sets were displayed on landing pages of 31 data facilities (56%). These DOIs may have been created by the data facilities themselves, or in some cases the DOIs that were presented were created by other organizations. This was the case, for example, for DataONE, which serves primarily as a data discovery interface for data sets distributed by other organizations. Previous study has shown that most DOIs created by Earth system science data facilities are created through the DataCite DOI registration agency (Goldstein, Mayernik, & Ramapriyan 2017). All DOIs created through DataCite are required to be associated with a metadata record structured according to the DataCite metadata standard. Thus, DOIs provide another source of metadata.

Using the DataCite Search service (<https://search.datacite.org>), we examined the DataCite metadata records for all 31 of the data sets that had associated DOIs. We examined whether the same subject terms / keywords were present as in our investigation of the facilities' web sites, and whether there was any indication in the DataCite records of which subject / keyword vocabularies those terms were drawn from.

Table 3 shows that for data facilities where keywords were present in the original metadata (25), a small number of the DataCite metadata records (7) contained the same keywords as the original metadata, and in one case, there were more keywords listed in the DataCite metadata record than in the original record. The rest of the DataCite records had less keywords / subject terms than the original metadata source. For one data facility that lacked keywords in the original metadata, a single subject term was listed in the DataCite metadata record.

Table 3: Presence of keywords / subject terms in DataCite metadata records

	More	Same	Less
Data facilities with keywords in original metadata	1	7	17
Data facilities without keywords in original metadata	1	5	N/A

Finally, although the DataCite schema allows a subject schema to be designated for subject terms, this was only present for four of the data facilities' DataCite records. No subject schema was seen in more than one DataCite record.

Intra-Organizational Variation

We noted earlier that some data facilities present users with multiple data discovery and access interfaces, and in some cases multiple data facilities exist under one organizational umbrella. Previous research has also indicated that intra-organizational variation should be an important consideration for data curation-related studies (Mayernik 2016). Here, we illustrate variation in metadata standard and controlled vocabulary use within a single organization, the US Geological Survey (USGS).

USGS is a large and diverse federal agency that collects and provides access to data of various kinds. The USGS data catalog presents metadata via the FGDC CSDGM standard for all data sets. A consistent set of four subject/keyword vocabularies are used within this catalog: “USGS Thesaurus Keywords,” “ISO 19115 Topic Category” keywords, and two categories simply labeled “Other Subject Keywords” and “Place Keywords.” But when leaving the data catalog to get to the data landing page, different metadata and keywords are displayed. For example, the USGS ScienceBase system presents users with the option to download metadata structured according to many different standards: 19115:2003, MODS, FGDC CSDGM, and Atom. Within any individual ScienceBase metadata record, subject terms drawn from different controlled vocabularies can be found, such as “Data Categories for Marine Planning” and the “Geographic Names Information System”.

Some records in the USGS Data Catalog, however, take the user to web sites for other USGS units. These web sites are more variable in terms of their display and metadata access. All appear to present metadata via the FGDC CSDGM standard, but the use of controlled vocabularies is much more variable. Some present keywords drawn from specific standards such as the NASA GCMD. Some other metadata records only use one or two USGS-specific keyword vocabularies.

Discussion

We start this discussion section with some reflections on the FAIR principles based on the findings of this study. We then provide some discussion of the implications of the study for the scientific data repositories broadly, and then for individual repositories.

Implications related to the FAIR Principles

Standardized metadata and controlled vocabularies impact data findability via increased ability to be aggregated and discovered through multiple systems. Further, for the data sets that lacked DOIs, the future findability is in question given known challenges with link rot in internet-based resources that do not have persistent identifiers. Regarding data accessibility, simply finding data sets and associated metadata to analyze was frequently a challenge due to complex interfaces and, in some cases, login requirements. Recommendations for data set “landing pages,” such as those presented by Starr et al. (2015), have not yet gained wide adoption within this community. On interoperability, the widespread use of both metadata

standards and controlled vocabularies within this community suggests that these facilities are individually meeting the FAIR Interoperability principles. Looking at the results cumulatively, however, the variation in the standards and vocabularies presents significant interoperability challenges. Finally, reusability of data is implicated in any study of metadata. Metadata supports both individuals and machines in assessing if particular data sets are useful for their distinct contexts. Lack of standardization hinders attempts to assess data in this way.

Implications for the Earth system science data facility community

From a metadata standard point of view, the Earth science data facility ecosystem shows a few partially overlapping segments. Many facilities use 19115:2003. NASA-related data facilities use the ATOM, DIF and Native NASA Common Metadata Repository (CMR) formats. Several facilities use the FGDC CSDGM standard, which is a predecessor to the ISO standard. An additional segment of the facility community uses the Ecological Metadata Language (EML) which emphasizes the databases that are concerned with ecology.

The prevalence of DOIs for data sets suggests that the usage of DataCite metadata schema is more common than any other schema that we observed. Prior research, however, has shown that large proportions of the DataCite metadata collection contain incomplete, non-standardized, or otherwise ambiguous information (Habermann 2020; Strecker 2021). Our study of the presence of subject terms in DataCite metadata records had a similar finding. Thus, questions remain about the use of the DataCite metadata schema as a path toward metadata consistency.

Older metadata standards still hold power for the data facilities in this community. The GCMD vocabulary has existed for over 20 years, even if their management and maintenance has been inconsistent over the years (Parsons, Duerr, & Godøy 2022). Likewise, the FGDC CSDGM standard was developed in the 1990s, and ISO 19115 was released in 2003. Neither standard was designed to manage persistent web-based identifiers such as DOIs for data sets or Open Research and Contributor IDs (ORCID) for people. FAIR compliance is a difficult bar for facilities that have metadata based in these older standards that lack support for managing persistent identifiers.

One salient reason for the persistence of older standards is simply the work required to complete a migration to a newer standard. As Willis, Greenberg, and White (2012, pg. 1513) state: “Metadata is part of a larger information ecology that includes systems and software.” Migrating thus essentially involves moving from one information ecology to another. Implementation guides need to be developed for the standard, and tools must be built and tested to migrate the metadata. Other tools also need to be developed or updated, such as metadata editors, data and metadata landing page displays, and search and discovery interfaces.

Data facilities may deliberately avoid migrating to new standards due to the work involved (Thomer, Weber, & Twidale 2018).

Another example of inertia relates to the changes and versioning of vocabulary standards over time. Organizations might apply vocabulary terms from a particular standard at a given point in time, but not update them later, even if the originating vocabulary standard changes. Many of the keyword lists shown in Table 2 have changed over time, including the GCMD. It is unknown if data facilities are keeping in sync with changes to the GCMD.

Thus, there may be a sort of early adopter penalty for data facilities that are the first to migrate to a new standard. Not only is the workload high, the early adopters become community outliers, since they are now using a different standard than the others who use older standards. This is a network effect of standards – the value of standards is limited unless others are using them (Ghosh 2011). As Cargill (2011) noted: “the choice usually left to the standardizers is either to standardize in anticipation of the market (anticipatory standardization) or to standardize after the specification has been implemented (standardize current practice).” The results of our study suggest that few data facilities have been willing to standardize to newer schemas, such as ISO 19115-1:2014, in an “anticipatory” fashion.

Transforms are another path to metadata consistency. But given the broad range of schemas that are being used, the number of transforms that would be needed to enable a move even to a single schema would be considerable. Metadata transforms can also lead to information loss if one schema is more complex than another. Thus, while transforms are important, they are not a cure-all for challenges related to metadata schema diversity.

Limitations

While our study conducted a broad investigation of the Earth system science data facility community, it was limited in a few ways. One limitation of our study is that we examined a small number of data set per facility, and directly collected data about one data set per facility. This methodology was based on the initial examination of 5-10 data sets per data facility to determine whether each facility practices were internally consistent, but we may have missed additional variation within the individual data facilities. Second, we did not analyze unstructured and non-standardized metadata, beyond looking for explicitly labeled keywords or subject terms. Many facilities included significant amounts of metadata or documentation such as documents or other text on data set landing pages. These kinds of unstructured metadata and documentation are critical to enabling users to understand data, while structured metadata and keywords are more conducive to supporting data discovery (Habermann 2018). Third, our examination simply noted when data facilities demonstrated usage of particular standards and controlled vocabularies. We did not examine their implementations of those standards or which version they might be using. Finally, we only

examined metadata presented on the web interfaces, and not APIs or other metadata formats such as JSON-LD or Schema.org that may not be viewable by web users but may be accessible by web crawlers or automated metadata aggregators. Many of these decisions were made to keep the study tractable, and merit further study.

Conclusion

This paper presents the results of a study of metadata schema and controlled vocabulary use by Earth system science data facilities. It investigates the community consistency in the use of such standards, and discusses implications for the FAIR principles. Some consistency is seen in the use of metadata standards, with a few specific standards being used more commonly than others, specifically ISO 19115:2003, Ecological Metadata Language, and a few NASA metadata standards. But none of these were used by more than 42% of the observed data facilities. DOIs are displayed for data sets at more than 50% of the data facilities, indicating that the DataCite metadata schema is potentially the mostly widely used schema among this group of facilities, but DataCite records from these facilities are less likely to include subject terms and indications of subject schema than the original metadata sources.

This study raises questions about the extent to which metadata standards and keyword vocabularies can facilitate interoperability beyond fairly narrow sub-sections of the Earth system science data community. It also raises questions about the paths to migration to new metadata standards. For example, this research shows that investing in migration from ISO 19115:2003 to ISO 19115-1:2014 may be a relatively futile consideration given the low adoption rate of the newer standard. On the flip side, it may present a leadership opportunity for one or more data facilities to invest in tool and community building around standards such as the newer ISO 19115-1:2014, since newer standards have capabilities that are more amenable to the FAIR-centered data ecosystem. Additionally, increased participation in the standards-setting processes by broader range of organizations may be helpful as new standards are developed. If organizations contribute to new standards as they are being written, they may be more likely to adopt those standards, and see them as enablers of interoperability or other beneficial outcomes, rather than as constraints on data management infrastructures.

Regarding the controlled vocabularies, the NASA Global Change Master Directory (GCMD) is the only vocabulary with any critical mass of use. Other vocabularies may be more appropriate for specific scientific research topics, but the added findability and interoperability of data due to common keywords within the broader Earth System science data ecosystem appears dependent on broader use of GCMD.

This research complements prior work that has demonstrated that applying the FAIR Principles to data facility operations is difficult (Dunning, De Smaele, & Böhmer 2017; Boeckhout, Zielhuis, & Bredenoord 2018). In addition, being compliant with the individual FAIR principles, such as I1, “(meta)data use a formal,

accessible, shared, and broadly applicable language for knowledge representation,” does not necessarily directly lead to data/metadata being interoperable. This does not reduce FAIR’s value as aspirational principles, but calls into question policies (whether publisher, funder, or organizational) that explicitly direct data facilities to be FAIR-compliant. Further investigation is needed to clarify the practical role of the FAIR principles on the operations of data facilities.

Acknowledgements

The authors thank Ted Habermann, the members of the NCAR Data Stewardship Engineering Team, and the anonymous reviewers for comments on earlier versions of this paper. This material is based upon work supported by the National Center for Atmospheric Research (NCAR), which is a major facility sponsored by the National Science Foundation under Cooperative Agreement No. 1852977, and managed by the University Corporation for Atmospheric Research (UCAR). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author, and do not necessarily reflect the views of NCAR, UCAR, or the National Science Foundation.

Data Availability

All data underlying this study are available at: Liapich, Y. & Mayernik, M.S. (2021). Investigation of metadata standard use by geoscience data repositories [data set]. UCAR/NCAR - DASH Repository. Version 2.0. <https://doi.org/10.5065/z9ch-wk24>.

Competing Interests

The authors declare that they have no competing interests.

References

- Arms, William Y., Diane Hillmann, Carl Lagoze, Dean Krafft, Richard Marisa, John Saylor, Carol Terrizzi, and Herbert Van de Sompel. 2002. “A Spectrum of Interoperability.” *D-Lib Magazine* 8(1). <https://doi.org/10.1045/january2002-arms>.
- Bastrakova, I.V. 2018. “Geoscience Australia Community Metadata Profile of ISO 19115-1:2014.” Edited by A. Sedgmen and M.F.M. Smith. Geoscience Australia. <https://doi.org/10.11636/record.2018.026>.
- Boeckhout, Martin, Gerhard A. Zielhuis, and Annelien L. Bredenoord. 2018. “The FAIR Guiding Principles for Data Stewardship: Fair Enough?” *European Journal of Human Genetics* 26(7): 931–936. <https://doi.org/10.1038/s41431-018-0160-0>.
- Brodeur, Jean, Serena Coetzee, David Danko, Stéphane Garcia, and Jan Hjelmager. 2019. “Geographic Information Metadata—An Outlook from the International Standardization Perspective.” *ISPRS International Journal of Geo-Information* 8(6): 280. <https://doi.org/10.3390/ijgi8060280>.
- Bugbee, Kaylin, Jeanné le Roux, Adam Sisco, Aaron Kaulfus, Patrick Staton, Camille Woods, Valerie Dixon, Christopher Lynnes, and Rahul Ramachandran. 2021. “Improving Discovery and Use of NASA’s Earth Observation Data Through Metadata Quality Assessments.” *Data Science Journal* 20. <https://doi.org/10.5334/dsj-2021-017>.

- Cargill, Carl F. 2011. "Why Standardization Efforts Fail." *The Journal of Electronic Publishing* 14(1). <https://doi.org/10.3998/3336451.0014.103>.
- Donaldson, Devan Ray, Ewa Zegler-Poleska, and Lynn Yarmey. 2020. "Data Managers' Perspectives on OAIS Designated Communities and the FAIR Principles: Mediation, Tools and Conceptual Models." *Journal of Documentation* 76(6): 1261–1277. <https://doi.org/10.1108/jd-10-2019-0204>.
- Dunning, Alastair, Madeleine De Smaele, and Jasmin Böhmer. 1970. "Are the FAIR Data Principles Fair?" *International Journal of Digital Curation* 12(2): 177–195. <https://doi.org/10.2218/ijdc.v12i2.567>.
- EarthCube. 2019. "Council of Data Facilities Founding Charter." University Corporation for Atmospheric Research (UCAR). <https://doi.org/10.5065/Y82R-3431>.
- Feinberg, Melanie. 2017. "The Value of Discernment: Making Use of Interpretive Flexibility in Metadata Generation and Aggregation." *Information Research* 22(1): CoLIS paper 1649. <http://informationr.net/ir/22-1/colis/colis1649.html>.
- Ghosh, Rishab. 2011. "An Economic Basis for Open Standards." In *Opening Standards: The Global Politics of Interoperability*, edited by Laura DeNardis, 75–96. Cambridge, MA: MIT Press.
- Gilliland, Anne J. 2008. "Setting the Stage." In *Introduction to Metadata: Pathways to Digital Information*, Online Edition, Version 3.0, edited by Murtha Baca. Los Angeles, CA: Getty Information Institute. http://www.getty.edu/research/publications/electronic_publications/intrometadata/setting.html.
- GO FAIR. 2021. FAIR Principles. <https://www.go-fair.org/fair-principles>.
- Goldstein, Justin C., Matthew S. Mayernik, and Hampapuram K. Ramapriyan, 2017. "Identifiers for Earth Science Data Sets: Where We Have Been and Where We Need To Go." *Data Science Journal* 16: Article 23. <http://doi.org/10.5334/dsj-2017-023>.
- Gordon, Sean, and Ted Habermann. 2018. "The Influence of Community Recommendations on Metadata Completeness." *Ecological Informatics* 43: 38–51. <https://doi.org/10.1016/j.ecoinf.2017.09.005>.
- Habermann, Ted. 2018. "Metadata Life Cycles, Use Cases and Hierarchies." *Geosciences* 8(5): Paper #179. <https://doi.org/10.3390/geosciences8050179>.
- Habermann, Ted. 2020. "DataCite Subject Metadata." Metadata Gamechangers Blog, August 13, 2020. <https://metadatagamechangers.com/blog/2020/7/13/datacite-subject-metadata>.
- Jacobsen, Annika, Ricardo de Miranda Azevedo, Nick Juty, Dominique Batista, Simon Coles, Ronald Cornet, Mélanie Courtot, et al. 2020. "FAIR Principles: Interpretations and Implementation Considerations." *Data Intelligence* 2(1–2): 10–29. https://doi.org/10.1162/dint_r_00024.
- Kervin, Karina, William Michener, and Robert Cook. 2013. "Common Errors in Ecological Data Sharing." *Journal of eScience Librarianship* 2(2): e1024. <https://doi.org/10.7191/jeslib.2013.1024>.
- Liapich, Yauheniya, and Matthew S. Mayernik. 2021. Investigation of Metadata Standard Use by Geoscience Data Repositories [data set]. UCAR/NCAR - DASH Repository. Version 2.0. <https://doi.org/10.5065/z9ch-wk24>.
- MacLennan, Alan, and Agnieszka Walicka. 2019. "An Investigation into Cataloguers' Experiences with RDA." *Journal of Librarianship and Information Science* 52(2): 464–475. <https://doi.org/10.1177/0961000618820655>.

Mannheimer, Sara, Leila Sterman, and Susan Borda. 2016. "Discovery and Reuse of Open Datasets: An Exploratory Study." *Journal of eScience Librarianship* 5(1): e1091.

<https://doi.org/10.7191/jeslib.2016.1091>.

Mayernik, Matthew S. 2016. "Research Data and Metadata Curation as Institutional Issues." *Journal of the Association for Information Science and Technology* 67(4): 973–993.

<https://doi.org/10.1002/asi.23425>.

Mayernik, Matthew S. 2020. "Metadata." *Knowledge Organization* 47(8): 696–713.

<https://doi.org/10.5771/0943-7444-2020-8-696>.

Parsons, Mark A., Ruth Duerr, and Øystein Godøy. 2022. "The Evolution of a Geoscience Standard: An Instructive Tale of Science Keyword Development and Adoption." *Geoscience Frontiers* 101400.

<https://doi.org/10.1016/j.gsf.2022.101400>.

Renear, Allen H., Simone Sacchi, and Karen M. Wickett. 2010. "Definitions of Dataset in the Scientific and Technical Literature." *Proceedings of the American Society for Information Science and Technology* 47(1): 1–4. <https://doi.org/10.1002/meet.14504701240>.

Robinson, Erin, Lesley Wyborn, Ben Evans, Adrian Burton, Simon Cox, and Tim Rawlings. 2019. "Earth and Environment Science Information Partners: ESIP & E2SIP Parallel Pathways on Opposite Sides of the Globe." *Earth and Space Science Open Archive (ESSOAr)*. Wiley.

<https://doi.org/10.1002/essoar.10500447.1>.

Srinivasan, Ramesh, Robin Boast, Jonathan Furner, and Katherine M. Becvar. 2009. "Digital Museums and Diverse Cultural Knowledges: Moving Past the Traditional Catalog." *The Information Society* 25(4): 265–278. <https://doi.org/10.1080/01972240903028714>.

Stall, Shelley, Lynn Yarmey, Joel Cutcher-Gershenfeld, Brooks Hanson, Kerstin Lehnert, Brian Nosek, Mark Parsons, Erin Robinson, and Lesley Wyborn. 2019. "Make Scientific Data FAIR." *Nature* 570(7759): 27–29. <https://doi.org/10.1038/d41586-019-01720-7>.

Starr, Joan, Eleni Castro, Mercè Crosas, Michel Dumontier, Robert R. Downs, Ruth Duerr, Laurel L. Haak, et al. 2015. "Achieving Human and Machine Accessibility of Cited Data in Scholarly Publications." *PeerJ Computer Science* 1: e1. <https://doi.org/10.7717/peerj-cs.1>.

Strecker, Dorothea. 2021. "Quantitative Assessment of Metadata Collections of Research Data Repositories." Humboldt-Universität Zu Berlin, May. <https://doi.org/10.18452/22916>.

Thomer, Andrea K., Nicholas M. Weber, and Michael B. Twidale. 2018. "Supporting the Long-term Curation and Migration of Natural History Museum Collections Databases." *Proceedings of the Association for Information Science and Technology* 55(1): 504–513.

<https://doi.org/10.1002/pra2.2018.14505501055>.

Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data* 3(1). <https://doi.org/10.1038/sdata.2016.18>.

Willis, Craig, Jane Greenberg, and Hollie White. 2012. "Analysis and Synthesis of Metadata Goals for Scientific Data." *Journal of the American Society for Information Science and Technology* 63(8): 1505–1520. <https://doi.org/10.1002/asi.22683>.