Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

Article

Metadata Life Cycles, Use Cases and Hierarchies

Ted Habermann[1,*]
[1,*] The HDF Group, thabermann@hdfgroup.org
* Correspondence: thabermann@hdfgroup.org; Tel.: +01-217-531-4202

1 Abstract

The historic view of metadata as "data about data" is expanding to include data about other items that must be created, used and understood throughout the data and project life cycles. In this context, metadata might better be defined as the structured and standard part of documentation and the metadata life cycle can be described as the metadata content that is required for documentation in each phase of the project and data life cycles. This incremental approach to metadata creation is similar to the spiral model used in software development. Each phase also has distinct users and specific questions they need answers to. In many cases, the metadata life cycle involves hierarchies where latter phases have increased numbers of items. The relationships between metadata in different phases can be captured through structure in the metadata standard or through conventions for identifiers. Metadata creation and management can be streamlined and simplified by re-using metadata across many records. Many of these ideas are being used in metadata for documenting the life cycle of research projects in the Arctic.

2 Introduction

The Data Life Cycle is a well know high-level description of typical steps or phases in scientific projects. There are many descriptions of this life cycle that vary in detail, but Figure 1 shows a general framework that includes planning, data collection, analysis, archiving, sharing, and reuse. The first three phases of this life cycle are well known in the scientific community as scientists have been planning experiments and observational campaigns for centuries in the context of the scientific method [1]. The later phases (sharing, archiving, and reuse) have received considerable attention during the last several decades as data collection and processing become more complex and expensive and many scientific problems require large multi-disciplinary teams. Maximizing the value of data, both expected and unexpected, is increasingly important. In fact, many would agree that the path connecting Data Sharing to Data Archive and Re-Use should not go through the End of Project, as sharing metadata (and data) before a project is over is now considered a best practice in many discussions of open science.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

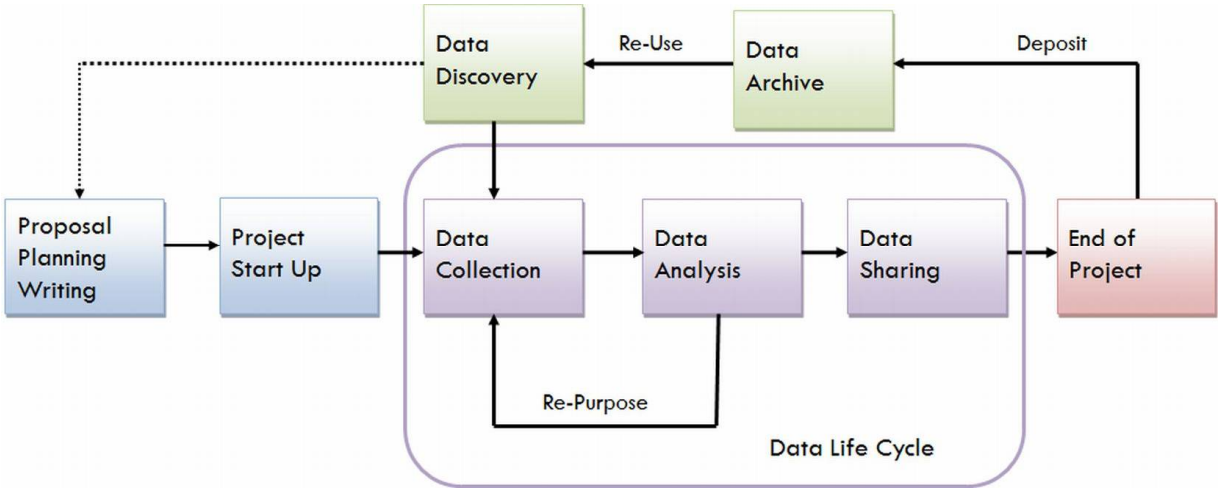

Figure 1. Overview of the Research Data Life Cycle [2].

The Open Archival Information System Reference Model [3] describes the processes and requirements for preserving scientific data and sharing it with users (designated communities). The OAIS-RM lists six mandatory requirements for open archival systems. One of these is clearly related to documentation of the data: "Ensure that the information to be preserved is independently understandable to the designated community. In other words, the community should be able to understand the information without needing the assistance of the experts who produced the information." Addressing this requirement clearly requires documentation that includes an unstructured component (papers, reports, presentations, …) as well as a structured component. We term this structured, and typically standardized, component of the documentation *metadata.*

The important role of metadata in supporting data discovery, access, use, and understanding has been clearly described many times and many of these discussions have focused on the discovery and re-use phases of the data life-cycle as exemplified by the OAIS-RM requirement. The archive is responsible for ensuring that the data are independently understandable. Clearly this requirement is difficult to satisfy without a significant collaboration between the archive and the information provider and the OAIS-RM acknowledges this (requirement 1 is "Negotiate for and accept appropriate information from the Information Producers"). Note that information includes data and documentation.

Several authors have discussed the concept of a metadata life cycle (see [4, 5] for example) and have emphasized the importance of on-going metadata creation, either automated or manual, during the entire data life cycle. We explore this idea along with associated metadata properties and use cases, some of which are within the data life cycle and some of which extend it. We also discuss a framework of metadata management that extends beyond specific datasets to include projects and collection sites as well as a mechanism for linking metadata across elements of the extended metadata life cycle.

3 Metadata Life Cycle and Spirals

The metadata life cycle can be described in terms of the phases in the data life cycle (i.e. collection/creation, sharing, discovery), but these phases apply equally to each type of metadata rather than occurring just once in a project. All metadata must be created, managed, and shared. We choose to focus on the metadata content that might be created at different phases of the data life cycle rather than the process(es) used to create that content.

The software engineering community has recently been very successful by envisioning and implementing the software development process as a series of spirals, each of which addresses a relatively small set of user requirements [6]. Each spiral involves several phases: requirements collection and prioritization, implementation, testing, and, most importantly, on-going interaction with users. Each spiral builds on previous work and requirements are addressed through a series of on-going iterations, each of which results in a more capable system.

Like a multi-spiral software development process, the creation of complete, high-quality documentation is an on-going interaction between several groups. It is an end-to-end process that occurs many times during the complete data life cycle. This idea was described in the NOAA Geo-IDE Wiki [7, 8] with spirals for identification, connection, extent, distribution, text searches, acquisition information, content information, and quality/lineage (Figure 2) and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

these spirals were used to guide the development of metadata evaluation tools and related guidance for metadata producers.

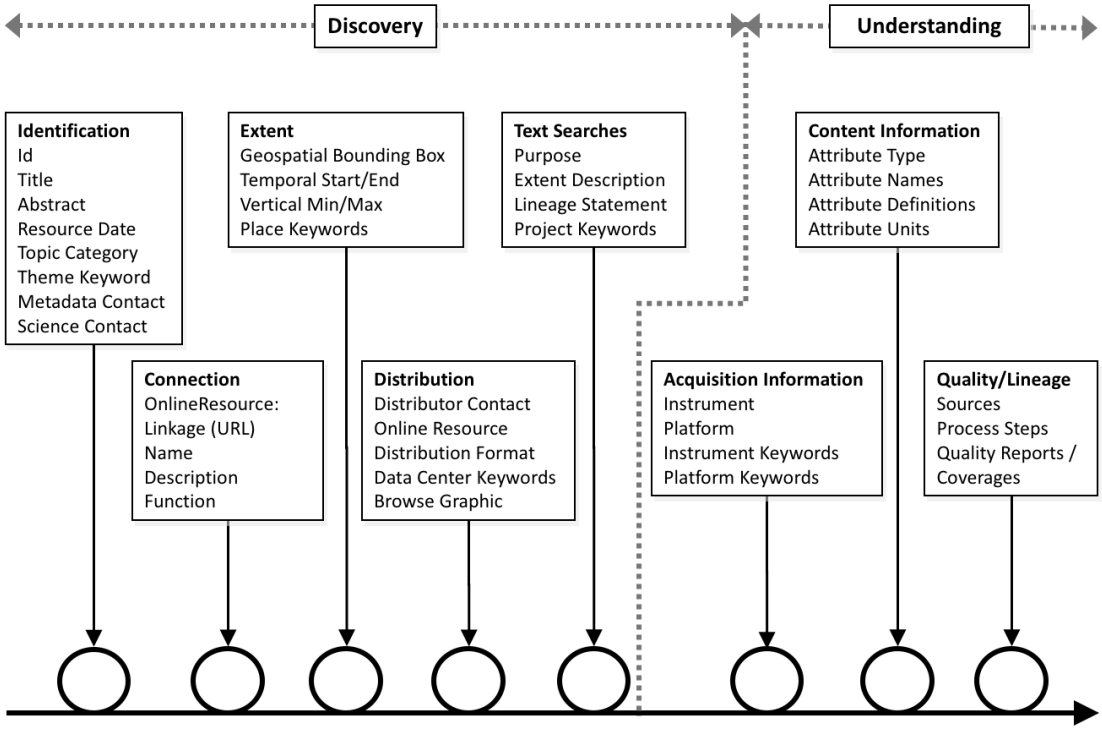

Figure 2. Potential metadata improvement spirals (Documentation Spirals).

The NOAA spirals were developed with metadata for archived datasets and the OAIS-RM in mind, but the idea can easily be extended beyond the data life cycle. NOAA's Satellite Products and Services Review Board [9] developed a system for accepting suggestions for new satellite products from users. The idea was that users would provide a fairly detailed description of a product that they would like NOAA to produce from operational satellite data and submit that as a request to NOAA who would then evaluate the request and respond.

The SPSRB user request form included over fifty elements, many of which are included in the metadata standards used by NOAA's Center for Environmental Information to describe datasets, e.g. technical and operational points of contact, geographic coverage, horizontal and vertical resolutions, data format, horizontal accuracy, archive requirements, instruments for acquisition, … This suggests that a useful metadata record could be created when the product was initially suggested, i.e. before it was produced or archived, and developed through time, along with the product.

The idea that metadata could describe planned data acquisitions is actually much older than this NOAA project. Appendix C of OMB Circular A-16 Revised [9] describes the history of the U.S. Federal Government guidance on spatial data and notes that the purpose of the original circular written during 1953 was "to insure (sic) that surveying and mapping activities may be directed toward meeting the needs of federal and state agencies and the general public, and will be performed expeditiously, without duplication of effort." This duplication of effort was to be avoided by U.S. agencies sharing information about where they would be collecting future data and what data they would be collecting, so that other agencies could take advantage of shared data rather than (re-)collecting it themselves. This is spelled out later in the Circular: "Federal agencies will promote and fully utilize partnerships that promote cost-effective data collection…". The mechanism for sharing this data described in this circular was the FGDC Content Standard for Geospatial Metadata [11] that became the foundation for U.S. metadata systems and continues to be used in many U.S. Agencies.

Habermann [12] integrated this idea with concepts from the ISO 19115 Geospatial Data Metadata Standard [13] to create a proposed set of spirals covering the data life cycle from initial user request through OAIS-RM negotiations (submission agreements) to feedback from real time and archive users (Figure 3). The metadata generated along this

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

path would cover the typical basic metadata use cases (discovery and access, i.e. Figure 2) and also support a significant step towards the more demanding data use and understanding use cases. This approach would also distribute the metadata production process across time and, if the metadata were managed as one growing information resource, conserve valuable metadata produced early in the project for later use.

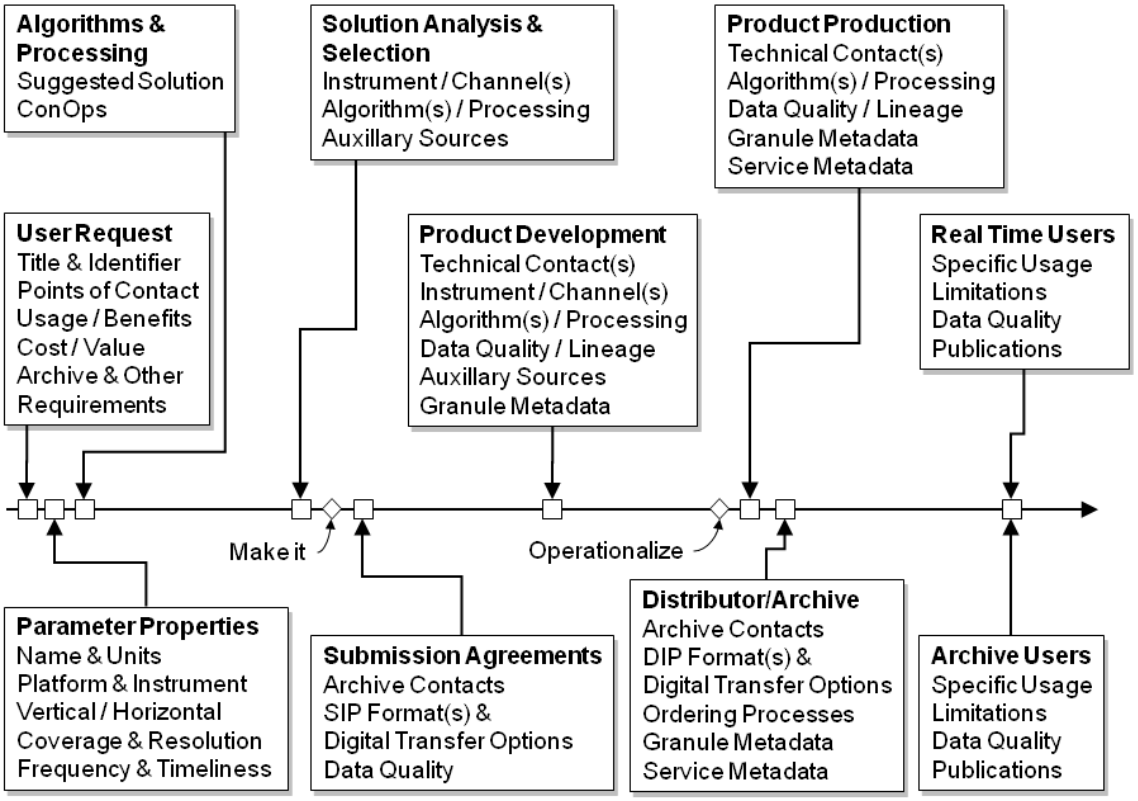

Figure 3. Metadata content through the extended data life cycle [12].

4 Metadata Scope

The discussion above focused on metadata for datasets, as do many metadata standards and management systems built for data discovery. Historically, many satellite remote sensing metadata systems have described data using two kinds of metadata: metadata for collections (or directory information interchange, see [14]) and metadata for granules [15].

The ISO Metadata Standard [13] formalized this idea into a concept called metadata scope. The standard values for scope are defined in a shared vocabulary (codelist) that contains the values listed in Table 1. The default value of this codelist is dataset, but scopes, and therefore metadata records, can refer to subsets of a dataset, i.e. an attribute, attribute type, coverage, dimension group, collectionSession, tile, etc. or to supersets, i.e. collection, aggregate, initiative, productionSeries, etc. ISO Metadata records can also describe resources that are not data, e.g. document, repository, sensor, or other metadata.

Table 1. Standard values for metadata scope from ISO 19115-1 [13].

aggregate	application	attribute	attributeType	**collection**
collectionHardware	collectionSession	coverage	**dataset***	dimensionGroup
document	feature	featureType	fieldSession	initiative
metadata	**model**	nonGeographicDataset	otherAggregate	platformSeries
product	productionSeries	propertyType	repository	sample

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

sensor	sensorSeries	series	**service**	**software**
tile		transferAggregate		

* - default value

The DataCite metadata schema [16] implements the metadata scope idea using the resourceTypeGeneral attribute on the resourceType metadata elements. The values that attribute can have are listed in Table 2 and there is some overlap between these two sets as one would expect (data types that are included in both lists are **bold** in Tables 1 and 2).

Table 2. Standard values for resourceTypeGeneral from DataCite [16].

Audiovisual	**Dataset**	InteractiveResource	**Service**	Text
Collection	Event	**Model**	**Software**	Workflow
DataPaper	Image	PhysicalObject	Sound	Other

There are many other examples of defining metadata types for different data types and use cases. In the W3C Dataset Profile for Health Care and Life Sciences, three types of metadata were identified (Summary, Distribution, and Version), again each with specific requirements and targets [17]. Different metadata for different types has also been demonstrated for data from the humanities [18,19]. Of course, definitions of some of these types may differ in detail, but the important point is that metadata for different things has different requirements, elements, and conventions. The inclusion of these different types as elements in metadata standards reflects the fact that many of these types share common documentation needs and understanding these shared needs is an important standardization element that can facilitate interoperability.

Another aspect of metadata scope that is important is that sub-sections of metadata records can have scope as well as entire records. This capability is critical if, for example, a dataset has multiple attributes (e.g. variables or parameters), each of which has different quality assessment procedures or different processing software or tools. Similar differences in quality could occur across different tiles or collection sessions so the metadata that describes the data quality needs to be connected to the part of the dataset that it is relevant to. These kinds of connections between different kinds of metadata within a single record are very important as datasets become more complex and detailed understanding becomes more critical during reuse.

This case is illustrated schematically in Figure 4 which shows a dataset (box) that includes multiple parameters of different types (e.g. physical measurements, model results, reference data, …). Each parameter (P_1, P_2, P_3) has specific information about its content and structure (contentInformation), some parameters have elements that pertain only to them (data quality and lineage for P_1), and some metadata are shared between several parameters (spatialRepresentation for P_2 and P_3). This approach essentially links multiple metadata sections in a single record and supports queries like "give me the quality information related to parameter P_1".

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

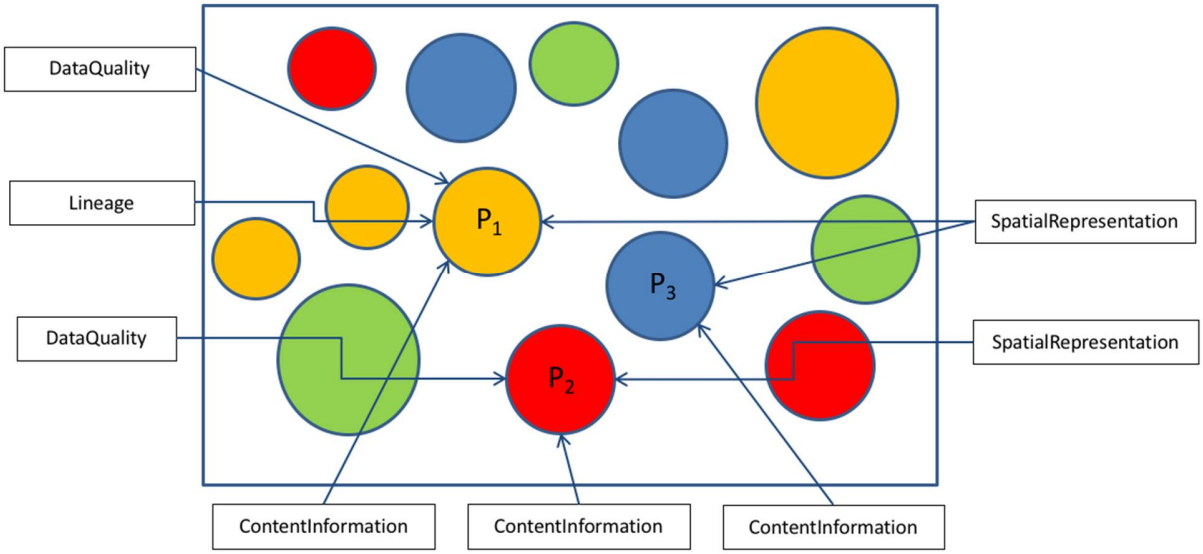

Figure 4. Schematic diagram of dataset with multiple parameters (circles), each of which has metadata.

5 Metadata Use Cases

The ultimate purpose of metadata is to answer questions about the resource(s) that they describe. The historical focus on metadata for datasets is reflected in the commonly given definition of metadata: "data about data". The plethora of resources that can be described with metadata (listed in Tables 1 and 2) clearly suggest that this definition is hopelessly out-of-date. A commonly used set of metadata use cases: discovery, access, use, and understanding, also needs updating.

Several on-going projects combine to provide a good example of different types of metadata that address different questions across the extended life-cycle illustrated in Figure 3. The Arctic Research Mapping Project [20] is designed to help funding agencies, logistics planners, research investigators, students, and others explore information about science being conducted across the Arctic and includes metadata about over 2700 projects funded by more than 18 different agencies. The metadata are created and maintained to help program managers at the National Science Foundation (NSF), and other users, answer planning questions like:

- Who has and is doing what, when and where?
- How do we plan for logistics?
- Where are medical facilities, field research stations, ship tracks, airports, etc.?

Requirements for these project metadata were initially developed by NSF and the Alaska Data Integration Working Group [21] during 2010 and they were implemented in the Content Standard for Digital Geospatial Metadata (CSDGM) dialect developed by the U.S. Federal Geospatial Data Committee [11]. The implementation included standard content as well as conventional uses of some elements, e.g. the bounding box gives the location of the logistics site(s) for the project; and some extensions, e.g. an ids section was added to hold identifiers for several kinds of organizations related to each project.

The project metadata requirements were re-implemented during 2014 using the ISO 191* dialects primarily because of the multi-national nature of research in the Arctic and because the required capabilities could be implemented in a standard way (i.e. without extensions) in those dialects [21]. For example, the scope of the project metadata could be unambiguously identified as "project" using a community-specific shared vocabulary (codelist) and the structure of the metadata could be expanded to accommodate hierarchical structures associated with the extended project life-cycle described in detail below and summarized here.

The projects described by the ARMAP metadata involve observations made in some region of the Arctic. The details of the observation locations are not known during the earliest phases of project planning. Once scientists are on the ground, they select collection sites based on local conditions. Projects typically encompass some number of collections sites and these sites make up the second layer of the metadata, i.e. Project/Collection Site.

The collections site metadata addresses several questions that come up during the collection site selection phase. These questions are particularly important in the Arctic, where physical access may be difficult or limited. These metadata are designed to help the Principal Investigators answer questions like:

- Where are existing data collection sites and what is being collected there?
- Where are more sites needed?
- Who operates and manages existing sites?
- Which sites can I use?

These questions reemphasize the importance of metadata during the initial project planning phases before any data have been collected or analyzed. In this case they are designed to facilitate re-use of collection sites and simplification of logistics and support.

As data are collected and analyzed, the dataset metadata can be written to document the specifics of the data and the analysis. These metadata are generally more familiar than project or collection site metadata and are intended to help scientists and other users trying to discover, access, use, understand, and trust the data:

- Is this dataset suitable for my research?
- Does it cover my area for the right time period?
- How were the data collected and processed?
- How was the quality of the data measured and documented?
- Are there papers or reports that used these data?
- How do I access the data and who do I contact with questions?

These metadata questions and use cases provide background and motivation for describing resources that are not just datasets, like those listed in Tables 1 and 2. These metadata can include structured and standard information that is useful throughout the extended data life cycle shown in Figure 3, raising the possibility that metadata "workflows" may connect multiple metadata records. The next section discusses an approach to organizing and connecting these records.

6 Data and Metadata Hierarchies

Items that support these use cases can have many connections and kinds of relationships. Hierarchical relationships and organization systems are ubiquitous in all scientific disciplines and familiar to all scientists and computer users as hierarchical file systems. Files on our computers (and email messages) are organized into folders within folders that we use regularly without thinking. Figure 5 shows a general structure for these directories along with three sample instances that hold information about a scientific project.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

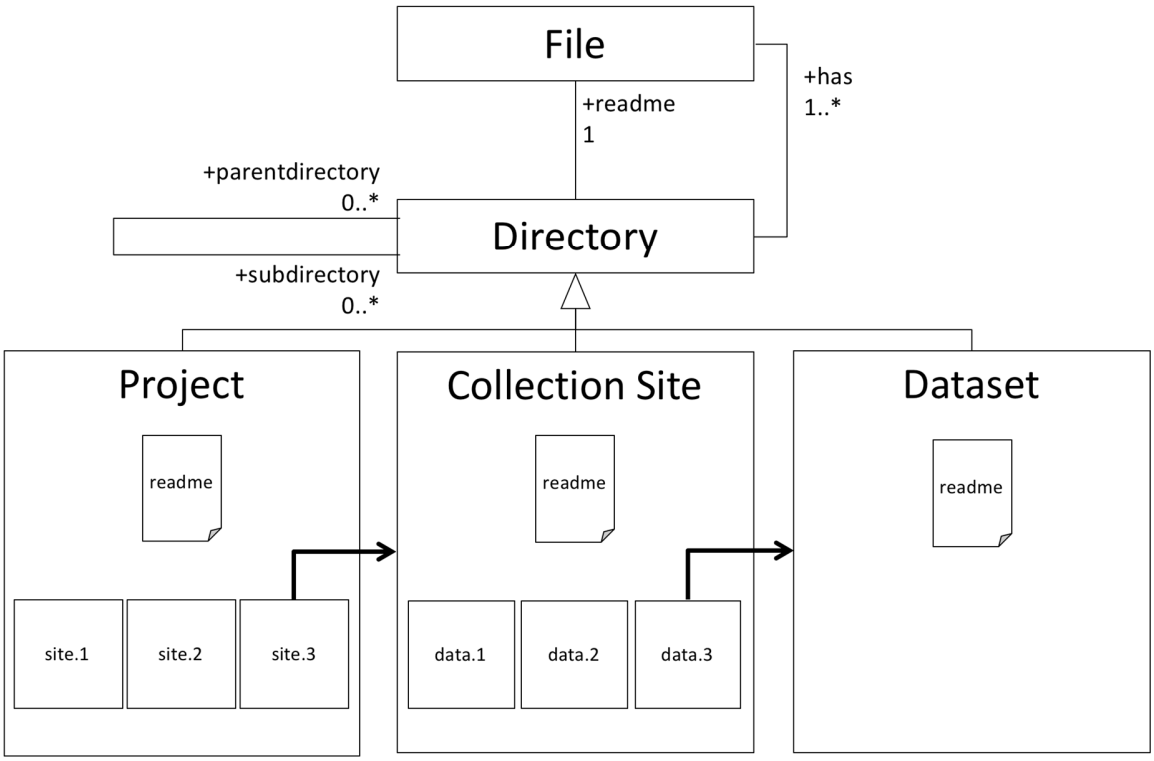

Figure 5. Typical hierarchical file system for observational data.

The top half of Figure 5 shows a simple UML class diagram [23] for a hierarchical file system. Each directory can have any number (0..*) of parent directories or subdirectories, as well as one or more (1..*) files (ignoring empty directories at this point) and only one file called readme. The readme file holds the documentation for the directory, typically in a text description written when the directory is created (hopefully), and other files hold the data.

The lower half of Figure 5 shows three directories that hold data associated with the metadata use cases described in the last section. The Project directory holds documentation for the project in a readme file and a subdirectory for each collection site. The Collection Site directories hold a readme file that holds documentation for the collection site and subdirectories that hold the datasets collected at the collection site. The dataset directories hold documentation for the data in a readme file and files that contain the actual data.

This directory layout is simplified to illustrate the general case described above. In real situations there may be layers for instruments, different collection times, quality information, and many others. In practice, the same information might be organized differently given the preferences of the person doing the organization. Systems for defining hierarchies always need to be flexible and extensible.

6.1 Hierarchies in Metadata

As described above, hierarchies are ubiquitous in scientific data communities. Can these same concepts and structures be used to organize related metadata? Historically this requirement has been approached using the concept of *parent metadata* [11, 13, 14] typically implemented as a reference to a collection included in metadata for datasets (sometimes granules) that make up the collection. For example, a collection of bathymetry surveys can include thousands of surveys from all over the world and the metadata for each of those surveys includes an identifier for the entire collection, the parent. This approach provides the equivalent of a back button, leading from discovered resources up one level, but it does not support an efficient mechanism for identifying the members of a collection from the parent, i.e. all of the collections sites in a project.

The directory structure schematic in Figure 5 outlines the requirements for a generic hierarchical metadata structure: it must include containers that can hold metadata that pertains to themselves, i.e. readme files, as well as items or references to other containers. These requirements are different than the historic requirements for dataset metadata records just as requirements for directories are different than those for files.

6.2 Metadata Hierarchies in ISO Geospatial Metadata

The ISO TC211 geospatial metadata standards [13] contain structures designed to support metadata hierarchies. Standard aggregation types (directories) can contain information related to an initiative (DS_Initiative), a series related to a particular platform (DS_Platform), or a sensor (DS_Sensor), a series produced by the same workflow (DS_ProductionSeries) or it can contain a set of items aggregated for some other reason (DS_OtherAggregate). Each of these aggregates has a metadata record (MD_Metadata) associated directly with it and is composed of any number of other records or containers.

The ISO standards are written as conceptual models, implemented in UML, that can be represented using whatever approaches communities choose. Standard implementations have been created in XML [24] and OWL [25], but others are possible. The XML framework for a single directory with two items is:

```
<DS_OtherAggregate>
  <has>
    <MD_Metadata> <!-- Metadata for aggregate --> </MD_Metadata>
  </has>
  <composedOf>
    <DS_DataSet>
      <has>
        <MD_Metadata> <!-- Metadata for item 1 --> </MD_Metadata>
      </has>
      <has>
        <MD_Metadata> <!-- Metadata for item 2 --> </MD_Metadata>
      </has>
    </DS_DataSet>
  </composedOf>
</DS_OtherAggregate>
```

The DS_OtherAggregate is the most general root element for the hierarchy framework. It could be any of the DS_* elements listed above. It "has" metadata (MD_Metadata) that describes the aggregate (the readme file) and is "composedOf" a DS_DataSet that contains the metadata record (MD_Metadata) for the items in the aggregate (files in the directory).

6.3 Metadata Hierarchies with Identifiers

A second approach to metadata hierarchies has been implemented for humanities data at the DARIAH Repository using generic handle [27] URLs as aggregation identifiers and extensions to the handle local identifier to indicate members of the aggregation [18, 19]. Figure 6 shows an example for the handle 21.1113/0000-000B-CA4C-D which identifies a container related to a book and extensions to that URL, e.g. /metadata, /index, link to various related metadata with various types.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

Handle.Net®

Handle Values for: 21.11113/0000-000B-CA4D-C

Index	Type	Timestamp	Data
1	CREATOR	2017-12-07 20:59:10Z	PID Service pid-webapp-4.22.0.201711102014
2	ADM_MD	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/adm
3	FILESIZE	2017-12-07 20:59:27Z	29740496
4	RESPONSIBLE	2017-12-07 20:59:27Z	BeataMache@dariah.eu
5	CHECKSUM	2017-12-07 20:59:27Z	md5:11992dc328bde12af85656964f2f28b4
6	BAG	2017-12-07 20:59:27Z	https://cdstar.de.dariah.eu/public/EAEA0-67B7-F5DC-4A4C-0
7	PUBDATE	2017-12-07 20:59:27Z	2017-12-07 21:59:19 +0100
8	PROV_MD	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/prov
9	URL	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C
10	DATA	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/data
11	LANDING	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/landing
12	INDEX	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/index
13	METADATA	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/metadata
14	TECH_MD	2017-12-07 20:59:27Z	https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-CA4D-C/tech
15	DOI	2017-12-07 20:59:27Z	http://dx.doi.org/10.20375/0000-000B-CA4D-C
16	INST	2017-12-07 20:59:27Z	2000
100	HS_ADMIN	2017-12-07 20:59:10Z	handle=21.11113/USER02; index=1; [create hdl,delete hdl,read val,modify val,del val,add val,modify admin,del admin,add admin]

Handle Proxy Server Documentation
Handle.net Web Site

Please contact hdladmin@cnri.reston.va.us for your handle questions and comments.

Figure 6. Example of metadata hierarchy using handles [19].

7 Metadata Components

The example shown in Figure 6 has several essential elements, identifiers for metadata associated with particular kinds of things (projects, collection sites, datasets, landing pages, dates, etc.) and URLs that resolve those identifiers into metadata (i.e. structured information) relevant to specific instances of those things. These elements allow metadata organization and access to take advantage of the same benefits that we take for granted on the World Wide Web: distributed resources linked together to provide information at various levels of detail. They also take advantage of persistent identifiers to encourage re-use of metadata which has the potential of greatly simplifying metadata creation and management.

The NOAA National Centers for Environmental Information (NCEI) have taken advantage of a similar approach in the Docucomp Component Management System [28] that supports creation and access to thousands of re-usable metadata components for many kinds of metadata. The system is designed around the ISO TC211 Standards which include the capability to attach identifiers to any section of a metadata record and to then include that section in any other metadata record with a reference implemented using xLink as shown below. Docucomp includes metadata components for over fifty kinds of metadata objects including: citations, responsible parties, platforms, instruments, quality measures, spatial/temporal extents, coordinate systems, algorithms, process steps, license information, distribution information, and many more.

This implementation differs from the DARIAH implementation shown above in that local identifiers (UUID's) that do not share a common root are used, rather than handles, and that metadata records are built by combining resource specific information with components that may be used in multiple records (thus no shared identifier root). In this case, the metadata record is actually the collection of all relevant components and specific use cases are addressed by specific subsets of that collection. In the DARIAH case, the metadata are thought of more as separate items that support specific use cases (tech-MD, prov_md, metadata, landing) and may not generally be viewed all together except as a list of related links (Figure 6).

Metadata creators and managers take advantage of Docucomp by creating components for metadata fragments that they expect to re-use in metadata records. Once those components are created, they can be referenced many times. For example, the link <gmd:referenceSystemInfo xlink:href="https://www.ngdc.noaa.gov/docucomp/895cc120-

95ed-11e0-aa80-0800200c9a66" xlink:title="WGS 84 / World Mercator"/> resolves to metadata for the WGS84 coordinate reference system that is used in many geospatial datasets. Using this component in those metadata records simplifies the inclusion of coordinate reference system metadata and also ensures consistent information about this reference system in all of the records that reference the component. The NCEI system monitors the integrity of the component links over time using standard link checkers and provides a standard report of component usage to metadata managers. It is not unusual for single components to be used in 10-100 metadata records.

This approach directly addresses a common obstacle to creating metadata: complexity of metadata dialects (perceived or real) and time required to learn how to create metadata in these dialects. A system built around re-usable components allows metadata creators to pick, for example, the WGS84 coordinate reference system to include in a metadata record without any knowledge of the details of how it will be represented in the metadata. The component, created by a metadata expert, ensures correct, consistent and complete content. This approach is also implemented in the open source catalog application and metadata management tool GeoNetwork [29] where components are termed "fragments".

Another example of this approach comes from the Biological and Chemical Oceanography Data Management Office [30] at Woods Hole Oceanographic Institute. BCO-DMO is a data repository that curates data from many scientific programs. They provide metadata for these datasets in many formats on top of a linked-data management system based on the Resource Definition Framework (RDF). In contrast to the NCEI system, therefore, the native format for BCO-DMO components is RDF. Several of these references are used in this citation from a BCO-DMO metadata record with the familiar identifier/URL pattern:

```
<citation>
  <CI_Citation>
    <title>
      <Anchor xlink:href=http://lod.bco-dmo.org/id/dataset/3673.rdf
       xlink:actuate="onRequest">GT11 - CTD - GT-C Sample Logs from the U.S. GEOTRACES NAT project of the
       U.S. GEOTRACES program</Anchor>
    </title>
    <citedResponsibleParty>
      <CI_ResponsibleParty>
        <individualName>
          <Anchor xlink:href=http://lod.bco-dmo.org/id/person/50984.rdf
              xlink:actuate="onRequest">Dr Edward  A. Boyle</Anchor>
        </individualName>
        <organisationName>
          <Anchor xlink:href=http://lod.bco-dmo.org/id/affiliation/222.rdf
          xlink:actuate="onRequest">Massachusetts Institute of Technology</Anchor>
        </organisationName>
    </citedResponsibleParty>
  </CI_Citation>
</citation>
```

This example uses components in RDF for the resource, the author, and the institution and the creator of the metadata can pick appropriate instances of these types during the metadata creation process. The standard ISO element Anchor is a substitution for a generic character string used to alert clients that the link given in the href attribute is important in this case.

8 Real-World Example

The hierarchy introduced above with three levels Project, Collection Site, and Dataset) has been implemented by the ARMAP project [20] along with a set of web services that connect the hierarchy levels (shaded arrow in Figure 7). The project metadata are created and stewarded by Polar Services Inc. in Denver, Colorado, as part of logistical support for the NSF Polar Program and each Project contains several collection sites. Five collection sites are shown for one project in Figure 7. The collections site metadata are created and stewarded at the University of Texas El Paso, and the dataset metadata is created and stewarded at Universities and Data Centers all over the world (DM in Figure

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 April 2018 doi:10.20944/preprints201804.0054.v1

Peer-reviewed version available at Geosciences 2018, 8, 179; doi:10.3390/geosciences8050179

7). The long-term goal of this project is to connect the collection sites to datasets that are collected at those sites (open arrows in Figure 7). Four datasets are shown for one collection site in Figure 7.

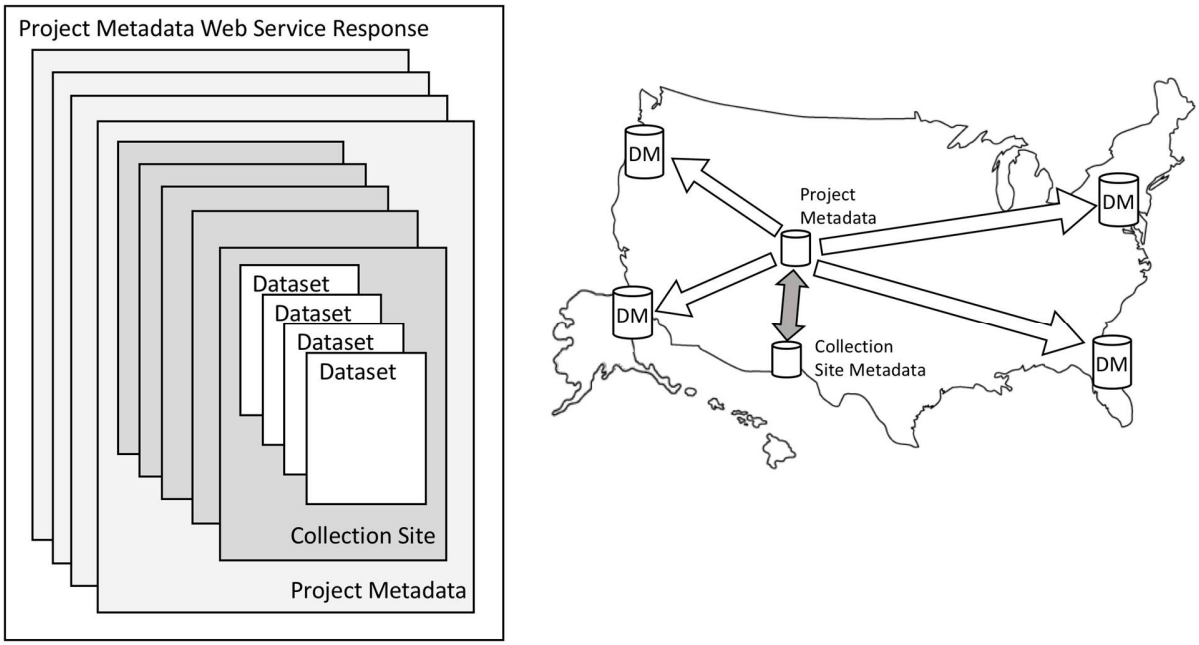

Figure 7. Hierarchical metadata distributed and connected by web services.

This distributed system introduces a benefit of hierarchical metadata that is not readily apparent from the discussion above. It facilitates spreading metadata creation and stewardship responsibilities across organizations with vested interests in different users of the hierarchy. In this case, Polar Services Inc. is primarily interested in serving project managers, planners, and logistics support groups that think about projects, so they handle the project metadata. The collection site metadata are created and stewarded in Texas as part of the Arctic Observing Viewer Project [31, 32] that focuses on locations, activities, and resources and includes information about well over 10,000 collection sites. This metadata is useful for detailed planning and for users interested in Arctic observational contexts. Finally, the dataset metadata are created and stewarded by scientists and data repositories that support scientific reusers during later stages of the project life cycle. No single organization alone can steward all of the metadata, but together they support the entire project life-cycle.

The project metadata are available in several formats (FGDC, ISO 19115-2, ISO 19115-3) through a web query interface (ARMAP). Both ISO formats provide project metadata using the hierarchical structure described above. The query interface provides metadata for multiple projects in a single response, effectively adding another layer to the hierarchy (left side of Figure 7). This schematic demonstrates that collections across complete hierarchies could be quite large, e.g. 4 projects X 5 collection sites / project X 4 datasets/collection site = 80 metadata records.

In order to avoid transferring large amounts of metadata, the web services support progressive discovery through the hierarchy using references (i.e. links) to metadata components and records rather than the records themselves. In the ISO Standards these links are implemented using the xLink standard (XLink) and the collection site element of the project metadata is a link to a list of collection site metadata records. For example, the list of collection sites for project 0084858 is given by:

<composedOf xlink:href="http://arcticobserving.utep.edu/DataCollectionSitesComponents/0084858CollectionSiteList.xml " xlink:title="Collection Sites for Project 0084858"/>

This collection site list is itself a metadata component like those discussed above. These collection site lists change infrequently, so they are stored in XML files that are updated when needed rather than in a database as in the example from NCEI discussed above. This demonstrates a well-known benefit that web services can be implemented in front of back ends that are appropriate for the organization serving the data and the specific requirements.

The collection site list also takes advantage of references to components to minimize network traffic and allow users to select the collection sites that they are interested in. For example, the list of collect sites looks like:

```
<DS_DataSet>
    <has xlink:href=http://arcticobserving.utep.edu/DataCollectionSitesComponents/AutonomousOcean-
    FluxBuoy(AOFB)_gbe8j269v2de_AOFB4_270.xml xlink:title="AutonomousOcean-
    FluxBuoy(AOFB)_gbe8j269v2de_AOFB4_270"/>
    <has xlink:href=http://arcticobserving.utep.edu/DataCollectionSitesComponents/AutonomousOcean-
    FluxBuoy(AOFB)_gbe8j269v2de_AOFB3_271.xml
    xlink:title="AutonomousOceanFluxBuoy(AOFB)_gbe8j269v2de_AOFB3_271"/>
    …
</DS_DataSet>
```

and each of these is a link to a single collection site metadata record. In this case, entire collection site metadata records are treated as components with the DS_DataSet providing a standard wrapper. Note that the hierarchy implementation currently ends at the collection site level. Future AOV plans will connect the collections sites to datasets using the same mechanism.

9 Conclusions

The classic definition of metadata, i.e. "data about data", is ubiquitous (Wikipedia), but covers only a fragment of the important roles that structured and standard documentation plays throughout the extended data life-cycle from project inception and planning through long-term use and re-use of data and related results. As projects move through this cycle, many different users, i.e. planners, project managers, principal investigators, and scientific communities, benefit from metadata about many different kinds of things. Examples of metadata about projects, collection sites, and datasets were described here and others have described metadata about many other kinds of things (people, institutions, provenance, algorithms, dates, software, ...).

This more comprehensive definition of metadata brings with it a broadened idea of who metadata management systems might serve and how they might operate. There are a growing number of examples of metadata systems that are built with the concepts of linking related metadata using approaches that have become commonplace in the World Wide Web and in the linked data world. These systems can be linked using a variety of persistent identifiers, URIs, and URLs.

Managing metadata as sets of linked resources, like we manage web pages and complex datasets, brings benefits of create once and re-use many times to metadata management. This greatly simplifies metadata creation while improving completeness and consistency, goals that are difficult to achieve in more typical record management systems. When combined with metadata hierarchies, and other generic link structures, this approach also facilitates separation of metadata creation and maintenance concerns.

10 Acknowledgements

Many colleagues at NOAA, NASA, ESIP, OGC, ISO TC211, NSF and elsewhere have contributed ideas described here and systems that implement many of them. Initial versions of the NOAA Enterprise Metadata Management Architecture provided early tests and proofs-of-concept. Working to evolve the ISO Geospatial Metadata Standards provided many long discussions and working sessions with experienced international practitioners. ARMAP and AOV provided colleagues committed to international collaboration and connected metadata repositories. Parts of this material are based upon work supported by the National Science Foundation under Grant No. NSFDACS11C1675. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

11 References

1. Scientific method, https://en.wikipedia.org/w/index.php?title=Scientific_method&oldid=833493426 (accessed 16 January 2018).
2. University of Virginia, Steps in the Data Life Cycle, http://data.library.virginia.edu/data-management/lifecycle/ (accessed 16 January 2018).

3. Consultative Committee for Space Data Systems (CCSDS), 2012-06, Reference Model for an Open Archival Information System (OAIS). Recommended Practice CCSDS 650.0-M-2, Issue 2, https://public.ccsds.org/pubs/650x0m2.pdf 2018-03-12 (identical to ISO 14721:2012).

4. Chen, Y.-N., Chen, S.-J., & S.C. Lin. (2003). A metadata lifecycle model for digital libraries: Methodology and application. Paper presented at The World Library and Information Congress: 69th IFLA General Conference and Council, 1-9 Aug. 2003, https://www.researchgate.net/publication/238680785_A_metadata_lifecycle_model_for_digital_libraries_methodology_and_application_for_an_evidence-based_approach_to_library_research (accessed 17 January 2018).

5. Dekkers, M., M. De Keyzer, N. Loutas, and S. Goedertier, 2013, Introduction to metadata management, https://www.slideshare.net/OpenDataSupport/introduction-to-metadata-management (accessed 2 April 2018).

6. Spiral Model, https://en.wikipedia.org/wiki/Spiral_model (accessed 17 January 2018).

7. Documentation Spirals, https://geo-ide.noaa.gov/wiki/index.php?title=Documentation_Spirals (accessed 17 January 2018).

8. Creating Good Documentation, https://geo-ide.noaa.gov/wiki/index.php?title=Creating_Good_Documentation (accessed 17 January 2018).

9. NOAA Satellite Products and Services Review Board, http://projects.osd.noaa.gov/SPSRB/index.htm (accessed 17 January 2018).

10. Office of Management and Budget (OMB), Circular A-16 Revised, https://www.whitehouse.gov/wp-content/uploads/2017/11/Circular-016.pdf (accessed 17 January 2018).

11. FGDC Content Standard for Geospatial Metadata, https://www.fgdc.gov/metadata/csdgm-standard (accessed 17 January 2018).

12. Habermann, T., 2011, Presentation at GeoData Forum, https://tw.rpi.edu/web/Workshop/Community/GeoData2011/MetadataLifeCycle (accessed 17 January 2018).

13. ISO 19115-1:2014, Geographic information -- Metadata -- Part 1: Fundamentals, https://www.iso.org/standard/53798.html (accessed 17 January 2018).

14. GCMD, A Short History of the Directory Interchange Format (DIF), https://gcmd.gsfc.nasa.gov/add/difguide/whatisadif.html (accessed 17 January 2018).

15. NASA ECHO Project (2010), ECHO Data Partner's Guide, https://cdn.earthdata.nasa.gov/conduit/upload/1475/ECHO_Data_Partner_User_Guide_0.pdf (accessed 15 January 2018).

16. DataCite Metadata Working Group. (2017). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V., http://doi.org/10.5438/0014 (accessed 31 January 2018).

17. Gray et al. (2015), Dataset Descriptions: HCLS Community Profile, https://www.w3.org/TR/hcls-dataset/#datasetdescriptionlevels (accessed 15 January 2018).

18. Schwardmann, U., 2016, PID Information Types Will Leverage Interoperability Between PID Systems, https://figshare.com/articles/PID_Information_Types_will_leverage_Interoperability_between_PID_systems/4224314 (accessed 1 February 2018).

19. Schwardmann, U. (2018): PIDs, Information Types and Collections - a Research Data Framework, https://doi.org/10.6084/m9.figshare.5849640.v1 (accessed 1 February 2018).

20. Gaylord, A.G., A. Kassin, W.F. Manley, M. Barba, R. Cody, M. Dover, S. Escarzaga, T. Habermann, J. Kozimor, R. Score, and C.E. Tweedie, 2016. Arctic Research Mapping Application (ARMAP). Englewood, Colorado USA: CH2M HILL Polar Services. Digital Media, http://www.armap.org (accessed 15 January 2018).

21. Alaska Data Integration Working Group (ADIWG), Projects, https://adiwg.github.io/projects/, (.

22. ARMAP Field Research Projects REST, http://armap.org/web-services/rest/ (accessed 15 January 2018).

23. Unified Modeling Language, https://en.wikipedia.org/wiki/Unified_Modeling_Language (accessed 17 January 2018).
24. ISO TC211 XML Management Group (XMG) Git Repository, https://github.com/ISO-TC211/XML (accessed 11 February 2018).
25. ISO TC211Group for Ontology Management (GOM) Git Repository, https://github.com/ISO-TC211/GOM (accessed 11 February 2018).
26. Digital Research Infrastructure for the Arts and Humanities (DARIAH), https://www.dariah.eu/ (accessed 11 February 2018).
27. The Handle Network, http://handle.net/ (accessed 17 January 2018).
28. Docucomp Component Management System, https://geo-ide.noaa.gov/wiki/index.php?title=Docucomp_Component_Management_System (accessed 7 February 2018).
29. GeoNetwork, Harvesting Fragments of Metadata to support re-use, https://geonetwork-opensource.org/manuals/3.4.x/is/user-guide/harvesting/index.html#harvesting-fragments-of-metadata-to-support-re-use (accessed 11 February 2018).
30. Biological and Chemical Oceanography Data Management Office (BCO-DMO), https://www.bco-dmo.org/ (accessed 7 February 2018).
31. Manley, William, F., Allison G. Gaylord, Ari Kassin, Ryan Cody, Walter A. Copenhaver, Mike Dover, Stephen M. Escarzaga, Ryan Font, Alan E. Garcia, Ted Habermann, David H. Lin, Roberta Score, Sandra Villarreal, Craig E. Tweedie, The U.S. Arctic Observing Viewer: A Web-Mapping Application for Enhancing Environmental Observation of the Changing Arctic, **ARCTIC**, [S.l.], v. 68, n. 5, p. 100-110, may 2015. ISSN 1923-1245. Available at: http://arctic.journalhosting.ucalgary.ca/arctic/index.php/arctic/article/view/4477 (accessed 7 February 2018), doi:http://dx.doi.org/10.14430/arctic4477.
32. Manley, W.F., Gaylord, A.G., Kassin, A., Villarreal, S., Cody, R., Barba, M., Dover, M., Escarzaga, S., Habermann, T., Kozimor, J., Score, R., and Tweedie, C.E., 2016, Arctic Observing Viewer (AOV): Englewood, Colorado USA, CH2M HILL Polar Services. http://arcticobservingviewer.org (accessed 7 February 2018).
33. XLink, https://en.wikipedia.org/wiki/XLink (accessed 2018-02-11).