Enhancing the FAIRness of Arctic Research Data Through Semantic Annotation

RESEARCH PAPER

STEVEN S. CHONG
MARK SCHILDHAUER
MARGARET O'BRIEN
BRYCE MECUM
MATTHEW B. JONES

*Author affiliations can be found in the back matter of this article

ubiquity press

ABSTRACT

The National Science Foundation's Arctic Data Center is the primary data repository for NSF-funded research conducted in the Arctic. There are major challenges in discovering and interpreting resources in a repository containing data as heterogeneous and interdisciplinary as those in the Arctic Data Center. This paper reports on advances in cyberinfrastructure at the Arctic Data Center that help address these issues by leveraging semantic technologies that enhance the repository's adherence to the FAIR data principles and improve the Findability, Accessibility, Interoperability, and Reusability of digital resources in the repository. We describe the Arctic Data Center's improvements. We use semantic annotation to bind metadata about Arctic data sets with concepts in web-accessible ontologies. The Arctic Data Center's implementation of a semantic annotation mechanism is accompanied by the development of an extended search interface that increases the findability of data by allowing users to search for specific, broader, and narrower meanings of measurement descriptions, as well as through their potential synonyms. Based on research carried out by the DataONE project, we evaluated the potential impact of this approach, regarding the accessibility, interoperability, and reusability of measurement data. Arctic research often benefits from having additional data, typically from multiple, heterogeneous sources, that complement and extend the bases – spatially, temporally, or thematically – for understanding Arctic phenomena. These relevant data resources must be 'found', and 'harmonized' prior to integration and analysis. The findings of a case study indicated that the semantic annotation of measurement data enhances the capabilities of researchers to accomplish these tasks.

CORRESPONDING AUTHOR:
Steven S. Chong
University of California Berkeley Library, University of California, Berkeley CA, USA
stevenchong@berkeley.edu

KEYWORDS:
Arctic research data; data discovery; FAIR; knowledge modeling; semantic annotation; data repository

TO CITE THIS ARTICLE:
Chong, SS, Schildhauer, M, O'Brien, M, Mecum, B and Jones, MB. 2024. Enhancing the FAIRness of Arctic Research Data Through Semantic Annotation. *Data Science Journal*, 23: 2, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2024-002

INTRODUCTION

The United States National Science Foundation's (NSF) Arctic Data Center (https://arcticdata.io) is the primary data repository for NSF-funded research conducted in the Arctic. Tied together by geography, the digital resources of diverse research communities are represented in the repository, including the natural sciences – such as earth science and biology, and social sciences – such as anthropology, archaeology, and economics. Each group defines and describes observational data according to the conventions of their respective disciplines, from ice core samples to atmospheric flux measurements to Alaskan Native food systems (Katz et al. 2019; Mantovani, Piana & Lombardo 2020), leading to numerous specialized vocabularies that vary both within and among scientific communities. Archiving data spanning across research domains also requires managing diverse file formats, ranging from PDF files to geospatial NetCDF files, along with accompanying metadata (Baumann et al. 2016; Kempler & Mathews 2017). Heterogeneity in data content, structure, and description has led to challenges in finding, discovering, interpreting, and analyzing data archived in the Arctic Data Center.

Initially, researchers need to find data of interest. Challenges in data discovery arise, however, because information systems traditionally rely on full-text search on the metadata to retrieve data, rather than searching by concepts, where the intended 'meaning' of the string is made clearer. The mismatch between the concepts that researchers use in search strings and how data are described can have detrimental effects on both search *precision* and *recall*. Other barriers to data discovery frequently arise from common linguistic issues that can lead to incomplete or incorrect search results (Aguado-de-Cea et al. 2015; Faria et al. 2018; Krovetz 1997; Krovetz & Croft 1992). Some linguistic features that confound typical data searches include (1) homonyms, (2) synonyms, and (3) hierarchically related concepts.

1. *Homonyms*. Data may be described using terms having distinct meanings in different disciplines, or even within the same discipline, such that identical homographs can yield false positive search results. For example, the term 'litter' has multiple meanings, including trash, a group of mammals born together, decomposing plant material on top of soil, and a wheel-less human-powered vehicle used for conveying people. A plant ecologist interested in the third meaning of litter would retrieve irrelevant data related to the other concepts when a query for the text string 'litter' is performed on an information system lacking some mechanism for disambiguation.
2. *Synonyms*. Data may also be described using different terms that have the same meaning, leading to missing data in search results. For example, 'carbon dioxide flux' may also be referred to as 'CO2 flux', i.e., substituting the compound name for its chemical formula. It is reasonable for an atmospheric scientist to compose their search using either term. However, depending on the search term used, the results retrieved can differ because of how the data were named and described. Figure 1 displays the result sets for queries on 'carbon dioxide flux' and 'co2 flux' in the Arctic Data Center's default search interface, which utilizes an enhanced string-matching approach. Because the search terms are synonyms and represent the same concepts, ideally the system should retrieve the same datasets for both queries. Figure 1 illustrates that the number and identity of datasets retrieved differ, with some datasets only appearing in one query but not the other.

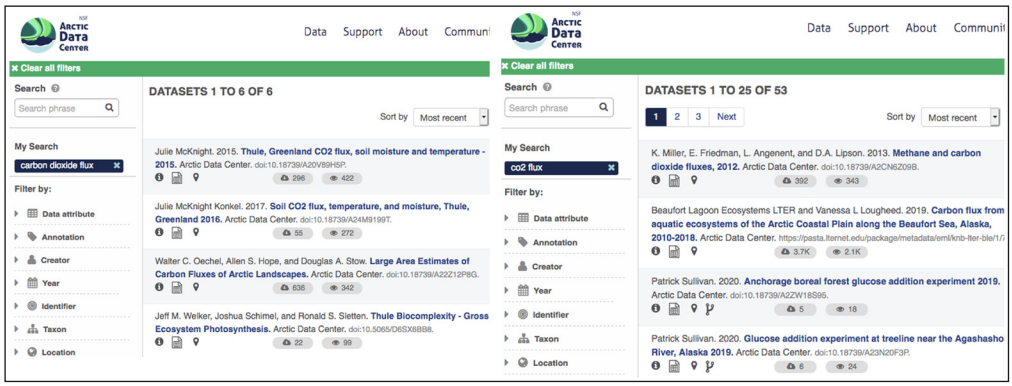

Figure 1 Search results for the synonyms 'carbon dioxide flux' (left) and 'co2 flux' (right) in the Arctic Data Center's default search interface, showing different counts and datasets.

3. *Hierarchically related concepts*. When a researcher performs a search, the information system may only retrieve results that match the searched text string and ignore related concepts, leading to incomplete search results. This problem is exemplified by concepts related through broader and narrower relationships. Ideally, when a concept is searched, all narrower concepts related to the broader term are also retrieved. For example, if a researcher searched on 'carbon flux' it is also desirable to retrieve data about 'carbon dioxide flux' and 'carbon monoxide flux' because both are types of carbon flux. Some, but not all information systems that use string matching might return the narrower concepts because these do contain the strings 'carbon' and 'flux'. However, a concept like 'methane flux', which is a type of 'carbon flux', might not be returned because it does not match any specific strings in the search terms.

The FAIR data principles (Findable, Accessible, Interoperable, and Reusable) describe several features and technologies to consider for generally increasing the utility of research data and metadata, for direct human interaction, as well as machine-assisted services (Wilkinson et al. 2016). Although repositories have had difficulty interpreting how to implement the FAIR principles (Dunning, De Smaele & Böhmer 2017), the principles provide practical guidance on improving information architecture, including recommendations to use languages like RDF/XML. Following these recommendations, the Arctic Data Center is leveraging Semantic Web technologies (Berners-Lee, Hendler & Lassila 2001) to better conform to the FAIR principles. By constructing an ontology using the World Wide Web Consortium (W3C) recommended RDF/OWL framework and language, the *Findability* and interpretability of dataset attributes in the Arctic Data Center are enhanced through well-described terms related in a hierarchy. Furthermore the *Interoperability* and *Reusability* of the ontology benefit from the use of RDF/OWL. We operationalize the ontology through semantic annotation, that links dataset attributes to terms in the ontology.

A semantic annotation approach is broadly adopted in some fields, most notably in genomics (Gene Ontology; Ashburner et al. 2000), and the biomedical sciences, e.g., the National Library of Medicine's continually updated Medical Subject Headings vocabulary, MeSH (Rogers 1963). Persistent identifiers corresponding to described resources in these vocabularies are used to annotate everything from journal articles to contributions to shared databases such as Genbank (https://www.ncbi.nlm.nih.gov/genbank/), providing clarity and interoperability, and facilitating synthetic insights. Identifiers are associated with persistent, dereferenceable HTTP IRI's, such as *http://id.nlm.nih.gov/mesh/D003920*, that can be used to annotate articles or other instances of the concept. Despite the clear advantages of this practice, it is not yet well-established in the environmental sciences, where entities and processes often have multiple, context-dependent labels and associations that detract from scientific clarity.

PRELIMINARY WORK

Based on extensive experience assisting researchers at the National Center for Ecological Analysis and Synthesis (NCEAS), a major ecological synthesis center, it was clear that to accomplish most integrative analyses or syntheses, scientists must spend a huge amount of time searching for data, which can be a frustratingly inefficient process. Moreover, we noted that researchers were often searching for specific measurements of interest, sometimes within a specific geographical area and time period of interest, but often much more broadly. Accordingly, our initial focus has been on deploying semantic methods to improve the efficiency and accuracy of search for scientific measurements, as an extension of the semantic web (Berners-Lee, Hendler & Lassila 2001).

DataONE (https://www.dataone.org) is a federation of institutions involved with the earth and environmental sciences that share data through common cyberinfrastructure. The DataONE project carried out a preliminary quantification of the utility of semantic query on the precision and recall of relevant data available through the DataONE catalog. *Precision* is here defined as the proportion of *relevant* data in the retrieved results, and *recall* is defined as the proportion of relevant data retrieved, compared to all relevant data present in the repository. To quantify precision and recall, a set of natural language queries was drafted (Table 1) based on interactions with NCEAS researchers and executed using various search mechanisms (e.g., specific areas of structured metadata or free text anywhere in metadata). When run against

Above-ground net primary production in a grassland, in dimensions of biomass of plant material (with or without area or duration)
Soil carbon content in dimensions of amount or mass of carbon per volume of soil or area of surface
Concentration of dissolved carbon dioxide, carbonate, or bicarbonate in the water of an aquatic system
Carbon dioxide flux in an enrichment experiment in dimensions of amount or mass of carbon per area or volume per time
Primary production of coastal macroalgae in dimensions of amount or mass of carbon per area per time

Table 1 Examples of queries used by scientists when searching for data about carbon-related processes were restructured (see text) and tested for precision and recall.

approximately 1000 datasets, results for the ten queries ranged from 0%–50% (precision) to 0%–100% (recall), indicating that traditional searches may sometimes be adequate to return all relevant data in a corpus, but results can be erratic and inconsistent, with potentially large returns of irrelevant data in the result set.

The measurement metadata of this same corpus of datasets was manually annotated with semantic concepts using an early version of the Ecosystem Ontology (ECSO; https://bioportal.bioontology.org/ontologies/ECSO) described below. When querying through semantic classes, precision and recall were much higher and more consistent (90%–100% and 75%–100%, respectively; Figure 2; O'Brien et al. 2023). This preliminary work demonstrated that when semantic annotation was applied and the search interface tailored to it, both dataset search precision and recall are enhanced.

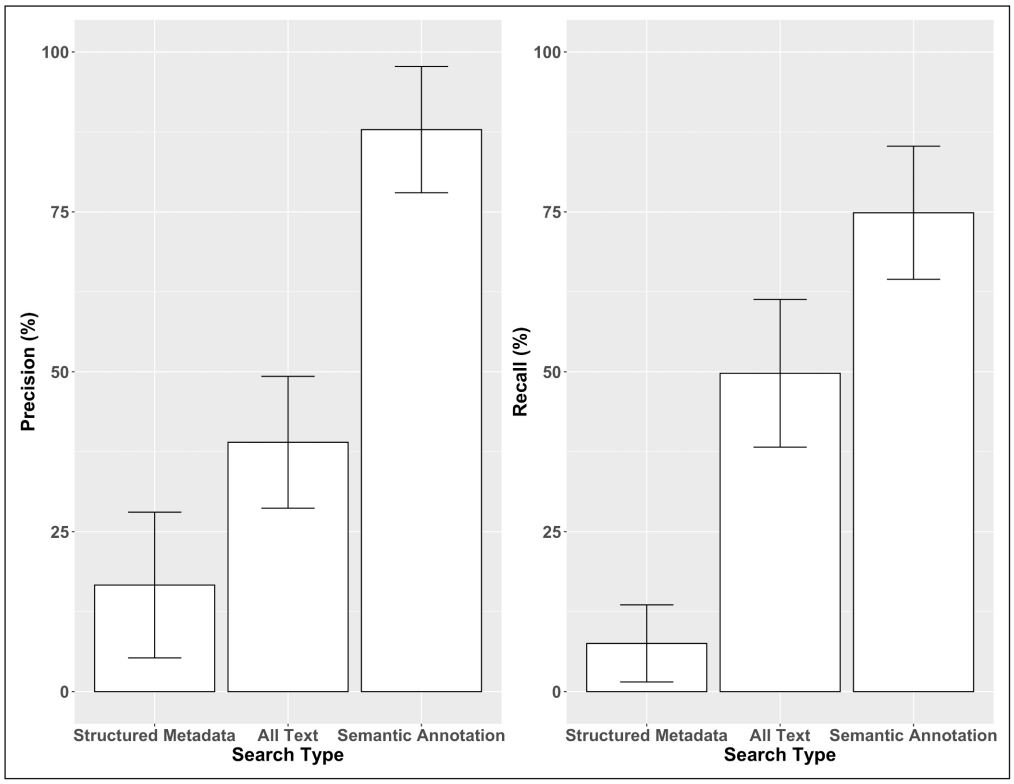

Figure 2 Results for ten queries against datasets in the DataONE corpus (left: precision, right: recall) Structured Metadata: searched fragments of the natural language query on a subset of metadata fields (title, abstract, and column description); All Text: searched fragments of the natural language query anywhere in metadata; Semantic Annotation: searched on ECSO IDs within semantic metadata only.

APPROACH

Following these encouraging results, the Arctic Data Center implemented the use of ontologies for enhancing search precision and recall. This effort entailed three basic tasks: (1) expanding the controlled vocabulary (Ecosystems Ontology, ECSO) to incorporate Arctic-relevant terms, (2) semantically annotating the data by binding metadata elements to terms in ECSO, and (3) enabling search on the semantically annotated data.

As a member node of the DataONE network, the Arctic Data Center chose the ECSO ontology (maintained by DataONE) for its annotations. Unlike many controlled vocabularies that are 'flat', ontologies can describe precise relationships among terms and are often stored as graph structures (Smith 2012). To further enhance open data sharing and discovery, ECSO is constructed according to the W3C recommendations of the Resource Description Framework

(RDF; Cyganiak, Wood & Lanthaler 2014) and OWL language (McGuinness & Van Harmelen 2004). In addition, we imported terms from other existing Web-accessible ontologies when these were relevant.

We chose an initial theme of carbon-related processes and measurements since these are critical components of ecosystem function in the Arctic. The extreme heterogeneity in carbon-related data, including how concepts, processes, and measurements are defined, leads to difficulties in interpreting measurements and inhibits ecological synthesis (Chapin et al. 2006; Harden et al. 2018). Synthesis typically requires access to data created by other researchers, to extend the thematic scope of an investigation (e.g., by including data on new parameters), as well as to expand the geospatial and temporal scale of the data. Improving researchers' ability to discover relevant carbon measurements would certainly improve scientists' understanding of the carbon cycle, a topic of critical importance for understanding the potential global implications of the rapidly changing climate of the Arctic region (Chapin et al. 2009).

To identify carbon-related datasets for annotation, we created a set of R scripts to query the Arctic Data Center with terms related to 'environmental carbon' gleaned from the published literature. The results of these queries revealed over 4000 datasets that potentially contained environmental carbon-related measurements or phenomena. From their metadata, we assembled descriptions of the carbon measurements into a single table, along with dataset identifiers. This table served as our key for manually annotating the dataset measurements with the URIs of relevant terms from our ontology. This process also informed efforts to improve ECSO, by adding terms or modifying existing terms in the ontology. Finally, the semantic annotations were inserted into the appropriate metadata records, ingested into the Arctic Data Center's Solr index (https://solr.apache.org), and a new user interface was developed to enable searching for data through the annotations. The R scripts used for the automated queries and insertion of the semantic annotations into the relevant records, along with instructions, are accessible in our GitHub repository (Chong et al. 2021).

ENHANCING THE ONTOLOGY (ECSO) AND KNOWLEDGE MODELING

ECSO contains terms that represent the types of measurements collected by ecosystem researchers and is an extension of the OBOE ontology (Madin et al. 2007). The description of measurements at the variable level is critical for understanding the contents of a data table. We used the Protégé ontology editor (https://protege.stanford.edu/) to expand the ontology and employed a bottom-up approach of analyzing the Arctic Data Center's holdings and adding new vocabulary terms to ECSO as needed to describe the measurement types present. Each term in ECSO uses RDFS (Resource Description Framework Schema; Brickley & Guha 2014) and SKOS annotation properties and ideally includes a label, a definition, alternative labels for any synonyms, and a preferred label that should be given priority for display, to align with the principles for ontology design promoted by the Open Biological and Biomedical Ontologies Technical Working Group (2020).

We currently use the ECSO ontology to annotate measurements in the repository's datasets. However, our annotation system can support the use of other ontologies, such as the Environment Ontology (ENVO; Buttigieg et al. 2013, 2016), Chemical Entities of Biological Interest (ChEBI; https://www.ebi.ac.uk/chebi/), and Phenotype and Trait Ontology (PATO; https://github.com/pato-ontology/pato/). For example, in future iterations, *permafrost depth* could be modeled using *permafrost* from ENVO (*http://purl.obolibrary.org/obo/ENVO_00000134*) as the entity, and *depth* from PATO (*http://purl.obolibrary.org/obo/PATO_0001595*) as the characteristic. Importing or referencing (via skos:exactMatch or similar properties) terms from other ontologies into ECSO is beneficial because it minimizes duplication of effort, and reduces confusion arising from the unnecessary proliferation of 'representations' of the same term by many different vocabularies.

SEMANTICALLY ANNOTATING THE DATA BY BINDING THE METADATA TO TERMS IN THE CONTROLLED VOCABULARY

The Annotation process consists of associating, or 'binding' an ontology term through its URI to a specific metadata element, e.g., the description of a column in a tabular dataset. A semantic annotation links a resource to a term in an ontology, enabling access, through the URI, to

descriptions of the type of variable measured, with the ontology clarifying that measurement types' relationships with other concepts in a machine-readable manner.

For metadata, the Arctic Data Center repository employs Ecological Metadata Language (EML, version 2.2.0; Jones et al. 2019), an XML schema widely adopted by environmental data repositories for describing the metadata for finding and interpreting the contents of products of scientific research, such as datasets, software, etc. The EML schema allows annotation on individual measurements, making it a good match for measurement types defined by ESCO. The semantic triples are serialized into EML annotations and inserted into existing metadata records to make them accessible to our optimized search interfaces. Additional details about the serialization of semantic annotation triples into EML are documented in the Semantic Annotation Primer section of the EML 2.2.0 specification (Jones et al. 2019).

The schema for annotations is a graph consisting of three parts (a 'triple'), the fundamental structure of the W3C's Resource Description Framework (http://w3.org/RDF), and a core component of the Semantic Web. Each triple consists of a subject, predicate, and object (Decker et al. 2000). A semantic triple's subject is the identifier (URI) for a specific variable in a dataset, the predicate (another unique URI) from OBOE describes a relationship of 'contains measurements of type', and the object is the appropriate measurement type class from ECSO, again indicated by its URI. While the annotation is serialized as three identifiers, it can be interpreted through its associated labels, as shown in the following statement for a carbon dioxide flux measurement.

Variable 'cflux' in dataset DOI:xx.yyy/zzz *contains measurement of type* carbon dioxide flux

In Figure 3, the data column 'CO2 exchange' is semantically annotated to the ECSO term labeled 'carbon dioxide flux'. The example EML snippet (Figure 3) shows that the subject of the semantic triple is implicitly the variable (described by the EML attribute element) containing the annotation node. The variable with the attribute name 'CO2 exchange' is the subject (accessed via its EML 'attribute id'). The EML propertyURI node describes the predicate in the semantic triple and contains the predicate's URI, along with an XML label attribute to present a more readable form of the predicate. In the example, the propertyURI node references the URI (*http://ecoinformatics.org/oboe/oboe.1.2/oboe-core.owl#containsMeasurementsOfType*) from the OBOE ontology (Madin et al. 2007), along with its 'contains measurements of type' label. The EML valueURI represents the object in the semantic triple, displaying the user-friendly label associated with this node. Here, the object is the 'Carbon Dioxide Flux' term from ECSO that is defined at the URI '*http://purl.dataone.org/odo/ECSO_00000536*'.

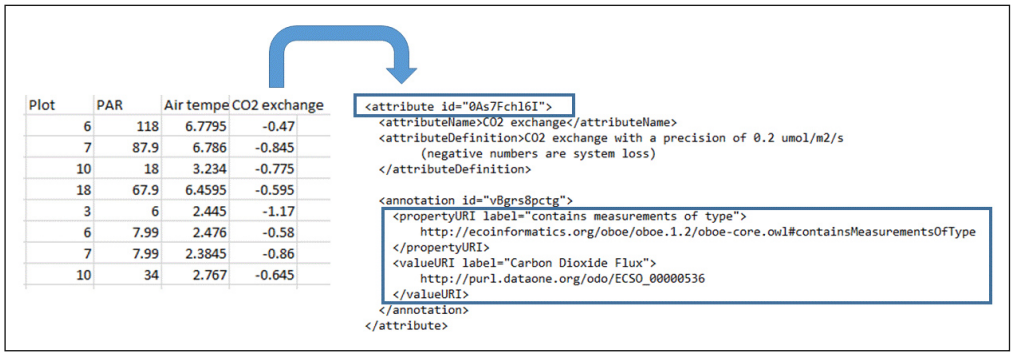

Figure 3 Semantic annotation of a dataset containing a carbon dioxide flux measurement, depicting how it is serialized into EML. The subject (accessed by the EML attribute id '0As7Fchl6I'), predicate (propertyURI), and object (valueURI) are contained within the boxes. Example taken from: https://arcticdata.io/catalog/view/doi%3A10.18739%2FA25M6275K.

IMPLEMENTING THE SEMANTIC SEARCH INTERFACE

An ontology-based semantic search interface was developed to enable enhanced finding of relevant data resources within the Arctic Data Center. The semantically annotated search feature is offered in addition to the default search method that utilizes traditional string matches on metadata contents. The ECSO ontology can be accessed through the National Center for Biomedical Ontology (NCBO) BioPortal repository (https://bioportal.bioontology.org/), for review and potential use by other systems. The BioPortal website also allows exploration of other Ontologies in its holdings.

When a user searches for concepts, text typed into the annotation search form leads to a list of suggested terms that 'match' the concept as defined in the ontology (Figure 4). After selecting one, the user is presented with a list of datasets that contain measurements annotated with that concept.

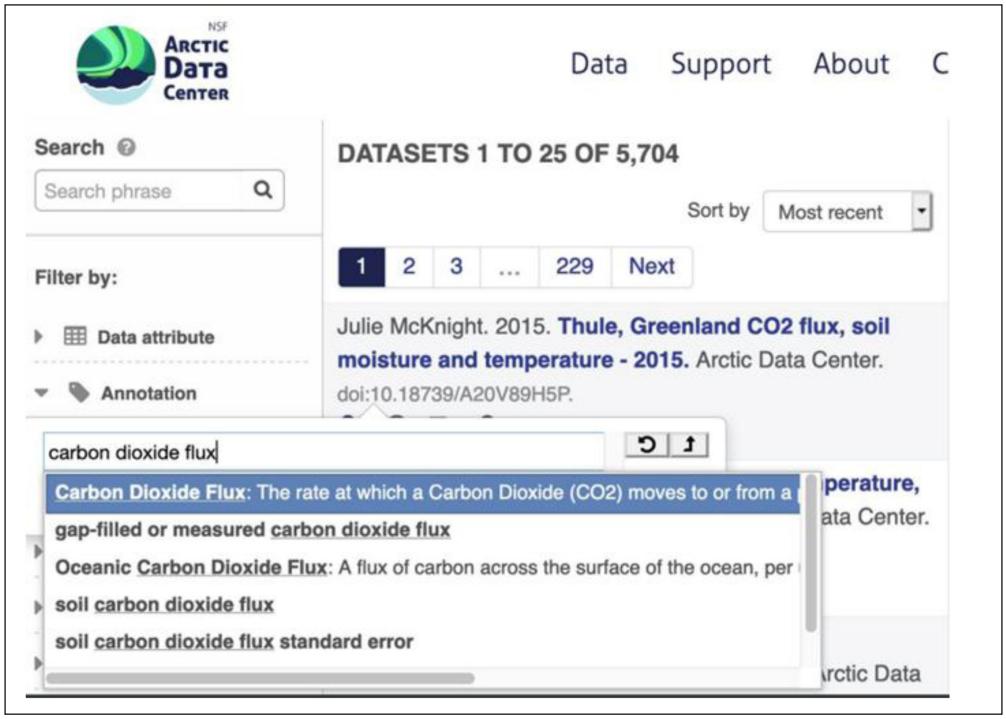

Figure 4 Semantic search interface displaying suggested concepts based on text typed in by the user.

The annotation form (Figure 5) permits a user to navigate the ontology's term hierarchy, including expanding and collapsing term classes, or 'drilling down' a class hierarchy to increase one's search precision. These are useful features if the user is unsure of what concept to search for and wants to explore related terms, or further refine their search to sub-topics. Once a concept is selected, the search returns all the datasets containing variables that have been

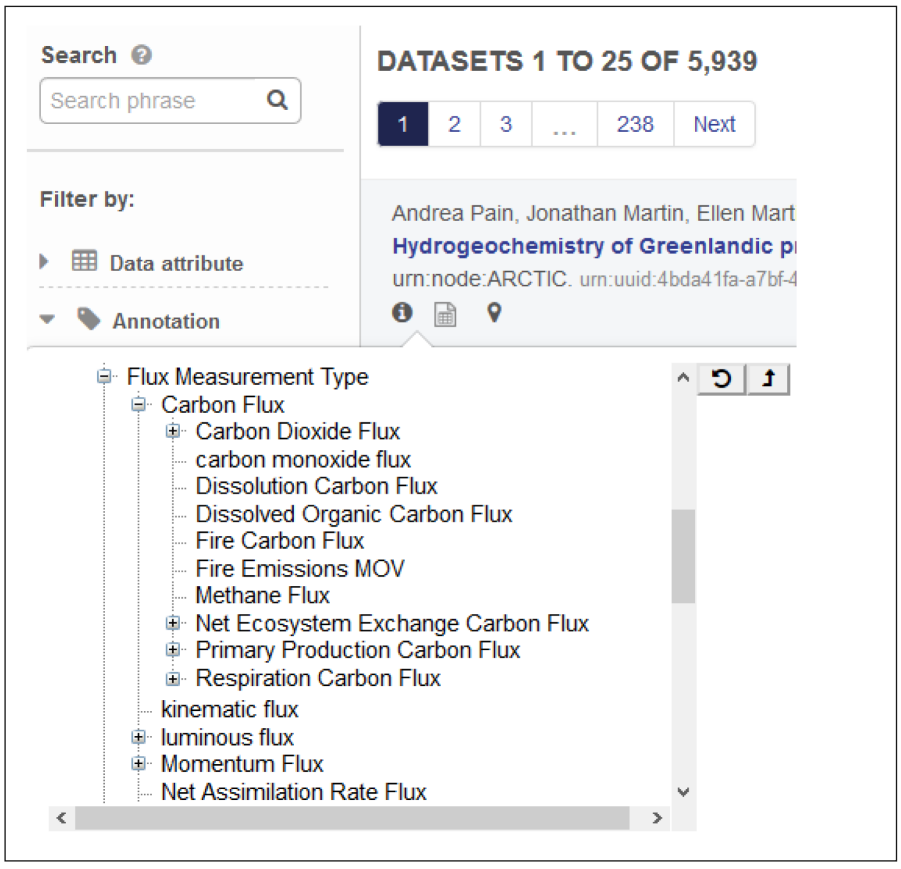

Figure 5 Browsing feature for viewing the annotation term hierarchy in the semantic search interface.

annotated with the selected concept plus datasets annotated with semantic children of that concept.

The hierarchical browsing feature enables a quick overview of potential measurements of interest in the repository and also provides clarification of these concepts, indicated by a definition (pulled from the ontology) appearing over a term when it is selected by a mouseover. In contrast, the natural language search provides no added clarification as to what a string actually 'means', and provides limited faceting along only a few topics – e.g., constraining a string search to specific metadata fields such as 'Creator', 'Data Attribute', 'Taxon', etc. The benefits of the class hierarchy provided by the (semantic) Annotation interface, along with the additional definitional descriptors, enable *greater precision* in selecting measurements of interest, and due to the binding of those parameters to the datasets holding those measurements, *higher recall* is also assured.

Synonyms are taken into account through the SKOS (Simple Knowledge Organization System; Miles & Bechhofer 2009) *alternative label* annotation property, such that, e.g., a user searching for 'carbon dioxide flux' will also retrieve datasets with variables described as 'CO2 flux'. Users searching for measurements will also retrieve all instances of measurements annotated to subclasses (narrower classes) of that term. Thus, searching for 'carbon flux' measurements will also return data annotated with 'carbon dioxide flux', 'carbon monoxide flux', and 'stomatal conductance', as these are all subclasses of 'carbon flux'.

Once the user finds an annotated measurement, the dataset landing page (Figure 6) includes interactive widgets that display additional information. Dataset variables (also referred to as dataset attributes) that contain semantic annotations are indicated with a badge displaying a checkmark in the Attribute Information box. The user has the ability to click on an annotation to gain further knowledge about the term, such as its definition, and initiate a new search for other datasets that are annotated with the same term. Figure 6 shows an example of a user selecting an annotated variable called 'CO2 Exchange' to reveal an interactive widget showing that the variable is actually about carbon dioxide flux. The term's definition, globally unique URI, and links to additional contextual information are provided, as well as the ability to find additional datasets annotated with the same term.

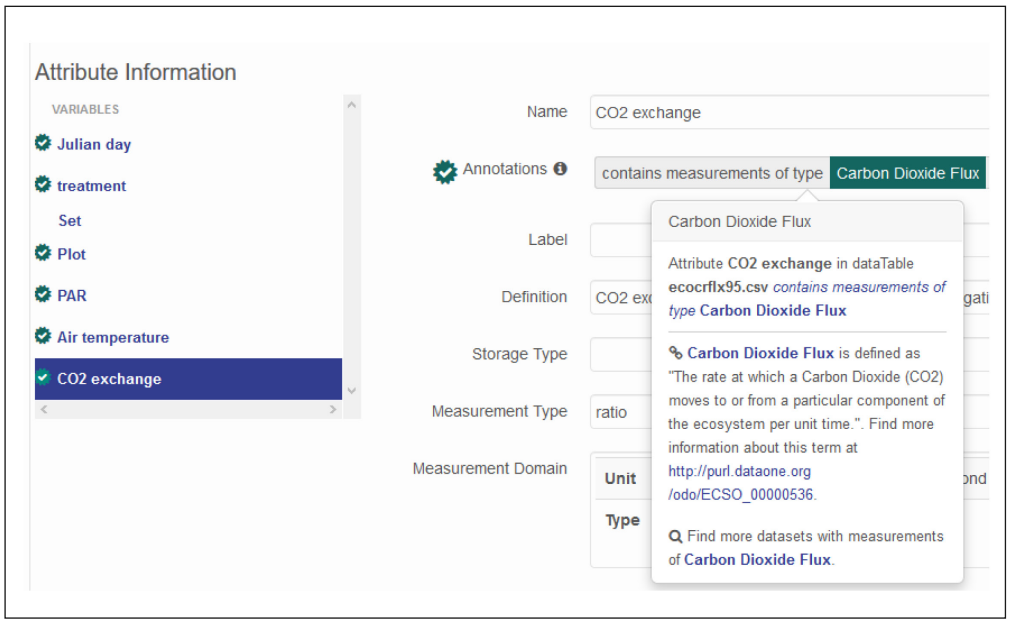

Figure 6 An informational interactive widget is displayed after clicking on an annotated variable (indicated by the 'check mark' to the left of the associated variable name). Example taken from: https://arcticdata.io/catalog/view/doi%3A10.18739%2FA25M6275K.

BENEFITS OF SEMANTIC ANNOTATION

The semantic annotation process links the Arctic Data Center's holdings to terms described in the formally constructed ECSO RDF/OWL ontology, thereby clarifying concepts and relationships among concepts in a machine-accessible manner. This enables the repository's semantic search interface to improve the utility of its holdings, in accordance with the FAIR data principles, that data are: Findable, Accessible, Interoperable, and Reusable.

FINDABLE

Attaching standardized descriptions to carbon measurements, as opposed to simple 'string searches', improves the ability to find data of interest. With the Arctic Data Center's semantic search interface, a user searching for data about *carbon dioxide flux* will automatically retrieve data annotated with subclasses of that term, even though the dataset descriptions do not explicitly contain the string 'carbon dioxide flux'. Thus, a user searching for *carbon dioxide flux* in the semantic search interface would also retrieve data about *stomatal conductance*.

Because measurement concepts are organized as a hierarchy in ECSO, these can be displayed in a manner that allows users to more precisely select the type of data they are looking for (Figure 5). The ontology also enables users to efficiently find data that may be described differently depending on discipline, as in the case with synonyms, or to differentiate among datasets represented by the same 'term' but with different meanings, as in the case with homographs. As a result these semantically enabled features available improve search precision and recall.

ACCESSIBLE

Data accessibility is promoted through the use of commonly accepted data transfer protocols. ECSO conforms to W3C-recommended semantic web standards, including RDF, OWL, and SKOS. Each term in ECSO contains a web-accessible URI that can be dereferenced using the HTTP protocol, allowing users to easily look up additional information about each term over the Web, how that term is described, and how it is related to other terms in the ontology. The terms imported into ECSO from other ontologies also conform to these same standards.

INTEROPERABLE AND REUSABLE

Adequate understanding of any measurement is essential to data reuse. Ontologies provide a standard way to describe and inter-relate measurements. Data interoperability and reusability are strengthened through the use of ontologies built according to common standards, such as those recommended by the W3C. Incorporating terms from existing ontologies provides the opportunity to build upon the work of others and prevents duplication of effort. With community involvement, vocabularies can be expanded and refined over time. In addition to vocabulary content, usage of RDF and OWL standards makes ontologies accessible over the Web, enhancing opportunities for work to be interoperable and reusable by others.

DataONE has deployed improvements to its search interface similar to those in the Arctic Data Center (Schildhauer et al. 2019). Other member repositories, such as the Environmental Data Initiative, have started on semantic annotations within their own organization (Gries et al. 2023; Vanderbilt, Gries & O'Brien 2020). As more members adopt semantic annotation of their data using shared vocabularies, users will be able to perform more precise searches for data across repositories, regardless of how the data are natively described. Agreement upon re-using well-defined terms in structured ontologies thus represents a major step forward in the data harmonization process, much as the adoption of standard units (e.g., meter, kilogram) facilitated comparability and interpretability of scientific measurements.

Semantic annotation helps clarify the interpretation of the data, promoting interoperability and reuse. Conventions in naming datasets and their attributes can differ according to discipline, leading to potential confusion and misinterpretation. The Arctic Data Center's semantic search interface helps resolve these issues through features that provide greater context for the data, including widgets that display additional information to the user, such as definitions and relationships of terms to other terms, and linkages to other datasets of potential interest. These features promote the reuse of related resources for potential synthesis. Further description of the Arctic Data Center's semantic search approach and justification are described on its website (Schildhauer, n.d.).

LESSONS LEARNED FROM ANNOTATING REAL-WORLD DATA

Developing an ontology, a semantic annotation process, and an enhanced search interface for the Arctic Data Center revealed several issues that are pertinent to data repositories considering implementing semantic annotations.

First, the semantic annotation process can be time-consuming and labor-intensive when done manually, especially when dealing with archived data that are only minimally described with existing metadata. In some cases, the original data creator may need to be contacted because the metadata are insufficient to determine the meaning of the intended measurement. Although our workflow for creating annotations made use of R scripts to generate EML from spreadsheets, scaling up semantic annotation to variables aside from carbon measurements and to higher levels (e.g., entire datasets) will require the development of more comprehensive ontologies, new software tools to assist researchers in creating annotations, and techniques such as machine learning, to help classify data and measurements. The Arctic Data Center has plans to extend its evaluation suite to cover checking for the presence of external annotations, but confirming the correctness of those annotations will be more complex. There are efforts underway to advance each of these aspects, as indicated by the number of ongoing semantically focused collaboration areas at, e.g., the Earth Science Information Partners (ESIP; https://esipfed.org), and Research Data Alliance (RDA; https://rd-alliance.org, e.g., Magagna et al. 2021); and sustained grass-roots efforts such as the OBO Foundry (https://obofoundry.org/). While advances in machine learning are enabling more efficient semantic annotation, these efforts rely on an established underlying knowledge base or ontology (Jovanović & Bagheri 2017; Liao & Zhao 2019). Thus, engagement with scientists knowledgeable in specific domains, to explicate concepts and their logical relationships, is a necessary precursor for more automated machine learning approaches.

Second, it is not always straightforward to decide on which controlled vocabularies or knowledge bases to use for annotations. We focused here on the annotation of carbon-related measurements, and so dealt with only one ontology (ECSO). We believe that existing vocabularies should be reused when possible, with the caveat that those existing vocabularies are constructed according to established knowledge-modeling principles, and adhere to Semantic Web principles and W3C recommendations. If a need for additional terms arises, these should be contributed to well-established vocabularies rather than minting entirely new vocabularies. Alternatively, if there is a need to develop a new vocabulary, it should reference existing terms in other vocabularies wherever possible. This can be done by referencing those terms' URIs using, e.g., SKOS 'exactMatch' (*https://www.w3.org/2009/08/skos-reference/skos.html*) or OWL 'sameAs' (*https://www.w3.org/TR/2012/REC-owl2-quick-reference-20121211/*) properties.

Coordinated efforts should be made *within disciplines* to converge on specific vocabularies tailored to meet community needs, but these should also attend to their vocabularies' interoperability *across disciplines*. Detailed criteria are only now being developed that identify the essential qualities of vocabularies so that they adhere to FAIR principles (Cox et al. 2021; Hugo et al. 2020). The development of such guidelines should improve the quality of existing vocabularies by guiding community vocabulary-building practices, minimizing duplication of work, and promoting the interoperability and reuse of fewer, high-quality vocabularies.

Third, there are major challenges in making highly usable search interfaces for exploring multiple vocabularies and annotations. Currently, the Arctic Data Center's Annotation search only displays ECSO's terms for measurement types. If additional vocabularies are needed for semantic annotation of measurements, these will require careful consideration as to how to apply and display 'mixed' hierarchies, so that finding a search term is not cumbersome or confusing to users. Another usability issue arises from the large number of terms in some vocabularies. For example, ENVO contains over 6000 terms, many of them relevant to the Arctic Data Center, but many that are not. Displaying every ENVO term in the browsing feature is likely to overwhelm and confuse users. One potential solution is to create thematic subsets of vocabularies so that irrelevant terms can be excluded from view. The display of search results for annotations made at different levels, e.g., at the dataset and variable levels, will also require the fine-tuning of user interfaces, where the client needs and expectations might vary depending on the discipline or anticipated level of technical expertise of the audience.

FUTURE PLANS

A long-term goal of the Arctic Data Center is to semantically annotate all of its data holdings. This includes annotating data at levels aside from the variable level (e.g., at the table, dataset,

and project levels) and indeed, a number of such semantic annotations have already been applied to a large subset of the data, clarifying the disciplinary themes of datasets, as well as the methodologies and specific instruments used in acquiring measurements. A User Interface has been developed to permit annotation of a dataset's measurements after uploading to the repository, but we are still in need of disciplinary-specific ontologies that express the full suite of measurement types and related concepts used across the Arctic.

We are also working to improve the specification of contextual information among variables in a data file. In a relational model, all of the attributes are properties of a common entity and are functionally linked. The nature of these linkages is typically not explicit; rather these are simply indications of some 'association'. Measurements of variables are similarly implicitly connected by virtue of being in the same file, as well as in the same 'record'. These connections are often more complicated than simply sharing some common theme, location, or spatial context, e.g., where one column is a ratio of two others. Accordingly, the Arctic Data Center retrieves the entire data package in which a semantically annotated variable appears, potentially providing such additional clarifying context. In addition, in contrast to the relational model, ontologies can explicitly describe relationships among variables. We are currently exploring the potential of the OBOÉ ontology to further explicate the contextual relationship between, e.g., two columns in a dataset (Madin et al. 2007, 2008).

CONCLUSION

Results from the DataONE project indicated that by linking dataset elements to terms defined in broadly accessible standards-based ontologies, semantic annotation makes data more FAIR, compared with simple text string searches across dataset contents, or across structured metadata corpora. Accordingly, the Arctic Data Center focused development on an ontology for carbon measurements, and expanded its metadata-based framework to enable semantic annotation of its dataset holdings. A semantic search interface based on this work was unveiled in Fall 2019. The interface improves the findability of carbon measurement data, making the Arctic Data Center a more useful knowledge resource to environmental carbon researchers. By binding carbon measurements to terms in a controlled vocabulary, we promoted standardization for how scientists describe their data, potentially providing greater clarity and precision in measurement interpretation. Our long-term goal is to have all data in the repository semantically annotated for improved discovery, interpretation, and reuse.

Although our implementation stored the semantic annotations in EML metadata, the process itself is generalizable to other serializations, e.g., JSON-LD or RDF. While our initial use case only incorporated terms from the ECSO ontology into the semantic search, the data model and process we created follow a design pattern that enables the inclusion of other vocabularies and annotations at additional levels, such as at the dataset level.

Since the Arctic Data Center released its semantic search capabilities, other repositories in the DataONE network have followed suit and begun annotating their data. Semantic annotation is a collaborative process and as more research organizations adopt semantic web technologies, broader disciplinary communities should work together and coordinate on best practices for vocabulary usage and annotation. Careful consideration should be made to evaluate existing controlled vocabularies before creating new ones.

The advantages of collecting data, e.g., measurements, that are semantically described or defined at the outset, rather than 'custom labeled' in some spreadsheet or database table, will expand the corpus of information that is amenable to 'FAIRer' semantic search in the future. By following W3C semantic web recommendations and adopting well-established controlled vocabularies, we have made the Arctic Data Center's data more findable, accessible, and interoperable with other repositories; and reusable to other researchers by providing additional context for interpreting the data.

DATA ACCESSIBILITY STATEMENT

Data from the DataONE case study are archived in the Environmental Data Initiative Data Portal and are freely available at: https://doi.org/10.6073/pasta/c93d87c2000715eaa2f70d079965c6a5.

The R scripts, along with the instructions used to create the semantic annotations and the output EML files, may be freely accessed at: https://archive.softwareheritage.org/swh:1:rel:b615ca4601ae230cdefa8035b708384d6fce3d06.

FUNDING INFORMATION

This work was supported as part of the Arctic Data Center funded by the U.S. National Science Foundation, Office of Polar Programs under Award Number 1546024 and 2042102. Earlier support was provided by the National Science Foundation through the DataONE project under Award Numbers 0830944 and 1430508. M.O. support was provided by US National Science Foundation (NSF) grant number 1931174.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

MS and MJ conceived the project. MS, MJ, and MO provided supervision and technical guidance for the execution of the project. MS and MO led the pilot work and the interpretation of its results. SSC, MS, and MO contributed terms to the ECSO ontology and added semantic annotations to the metadata. MJ and BM designed, and BM implemented, the repository user interface enhancements. SSC wrote the initial manuscript, with all authors subsequently contributing to editing the manuscript.

AUTHOR AFFILIATIONS

Steven S. Chong orcid.org/0000-0003-1264-1166
University of California Berkeley Library, University of California, Berkeley CA, USA

Mark Schildhauer orcid.org/0000-0003-0632-7576
National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara CA, USA

Margaret O'Brien orcid.org/0000-0002-1693-8322
Marine Science Institute, University of California, Santa Barbara CA, USA

Bryce Mecum orcid.org/0000-0002-0381-3766
National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara CA, USA

Matthew B. Jones orcid.org/0000-0003-0077-4738
National Center for Ecological Analysis and Synthesis, University of California, Santa Barbara CA, USA

REFERENCES

Aguado-de-Cea, G, Montiel-Ponsada, E, Poveda-Villalón, M and **Giraldo-Pasmin, O.** 2015. Lexicalizing ontologies: The issues behind the labels. *Procedia: Social and Behavioral Sciences*, 212(2015): 151–158. DOI: https://doi.org/10.1080/17538947.2014.1003106

Ashburner, M, Ball, CA, Blake, JA, Botstein, D, Butler, H, Cherry, JM, Davis, AP, Dolinski, K, Dwight, SS, Eppig, JT and **Harris, MA.** 2000. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1): 25–29. DOI: https://doi.org/10.1038/75556

Baumann, P, Mazzetti, P, Ungar, J, Barbera, R, Barboni, D, Beccati, A, Bigagli, L, Boldrini, E, Bruno, R, Calanducci, A and **others.** 2016. Big data analytics for earth sciences: The EarthServer approach. *International Journal of Digital Earth*, 9(1): 3–29. DOI: https://doi.org/10.1080/17538947.2014.1003106

Berners-Lee, T, Hendler, J and **Lassila, O.** 2001. The semantic web. *Scientific American*, 284(5): 34–43. DOI: https://doi.org/10.1038/scientificamerican0501-34

Brickley, D and **Guha, RV.** 2014. RDF Schema 1.1. Available at https://www.w3.org/TR/rdf-schema/ [Last accessed 11 January 2021].

Buttigieg, PL, Morrison, N, Smith, B, Mungall, CJ, Lewis, SE and **Envo Consortium.** 2013. The environment ontology: Contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, 4(1): 43. DOI: https://doi.org/10.1186/2041-1480-4-43

Buttigieg, PL, Pafilis, E, Lewis, SE, Schildhauer, MP, Walls, RL and **Mungall, CJ.** 2016. The environment ontology in 2016: Bridging domains with increased scope, semantic density, and interoperation. *Journal of Biomedical Semantics*, 7(1): 57. DOI: https://doi.org/10.1186/s13326-016-0097-6

Chapin, FS, McFarland, J, McGuire, AD, Euskirchen, ES, Ruess, RW and **Kielland, K.** 2009. The changing global carbon cycle: Linking plant–soil carbon dynamics to global consequences. *Journal of Ecology*, 97: 840–850. DOI: https://doi.org/10.1111/j.1365-2745.2009.01529.x

Chapin, FS, Woodwell, GM, Randerson, JT, Rastetter, EB, Lovett, GM, Baldocchi, DD, Clark, DA, Harmon, ME, Schimel, DS, Valentini, R and **others.** 2006. Reconciling carbon-cycle concepts, terminology, and methods. *Ecosystems*, 9(7): 1041–1050. DOI: https://doi.org/10.1007/s10021-005-0105-7

Chong, S, Jones, M, Mecum, B and **O'Brien, M.** 2021. Arctic semantics GitHub repository (Supplement to "Enhancing the FAIRness of Arctic research data" in IJDC release) [software]. Available from. https://archive.softwareheritage.org/swh:1:rel:b615ca4601ae230cdefa8035b708384d6fce3d06.

Cox, SJ, Gonzalez-Beltran, AN, Magagna, B and **Marinescu, MC.** 2021. Ten simple rules for making a vocabulary FAIR. *PLoS Computational Biology*, 17(6). DOI: https://doi.org/10.1371/journal.pcbi.1009041

Cyganiak, R, Wood, D and **Lanthaler, M.** 2014. RDF 1.1 Concepts and Abstract Syntax. Available at https://www.w3.org/TR/rdf11-concepts/ [Last accessed January 13, 2021].

Decker, S, Melnik, S, Van Harmelen, F, Fensel, D, Klein, M, Broekstra, J, Erdmann, M and **Horrocks, I.** 2000. The semantic web: The roles of XML and RDF. *IEEE Internet Computing*, 4(5): 63–73. DOI: https://doi.org/10.1109/4236.877487

Dunning, A, De Smaele, M and **Böhmer, J.** 2017. Are the fair data principles fair? *International Journal of Digital Curation*, 12(2): 177–195. DOI: https://doi.org/10.2218/ijdc.v12i2.567

Faria, D, Pesquita, C, Mott, I, Martins, C, Couto, F and **Cruz, I.** 2018. Tackling the challenges of matching biomedical ontologies. *Journal of Biomedical Semantics*, 9: 4. DOI: https://doi.org/10.1002/ece3.9592

Gries, C, Hanson, PC, O'Brien, M, Servilla, M, Vanderbilt, K and **Waide, R.** 2023. The environmental data initiative: Connecting the past to the future through data reuse. *Ecology and Evolution*, 13(1): e9592. DOI: https://doi.org/10.1002/ece3.9592

Harden, JW, Hugelius, G, Ahlström, A, Blankinship, JC, Bond-Lamberty, B, Lawrence, CR, Loisel, J, Malhotra, A, Jackson, RB, Ogle, S and **others.** 2018. Networking our science to characterize the state, vulnerabilities, and management opportunities of soil organic matter. *Global Change Biology*, 24(2): e705–e718. DOI: https://doi.org/10.1111/gcb.13896

Hugo, W, Le Franc, Y, Coen, G, Parland-von Essen, J and **Bonino, L.** 2020. D2.5 FAIR Semantics Recommendations Second Iteration (1.0). *Zenodo*. DOI: https://doi.org/10.5281/zenodo.5362010

Jones, MB, O'Brien, M, Mecum, B, Boettiger, C, Schildhauer, M, Maier, M, Whiteaker, T, Earl, S and **Chong, S.** 2019. *Ecological Metadata Language* (version 2.2.0). KNB Data Repository. DOI: https://doi.org/10.5063/F11834T2

Jovanović, J and **Bagheri, E.** 2017. Semantic annotation in biomedicine: the current landscape. *Journal of Biomedical Semantics*, 8(1): 1–18. DOI: https://doi.org/10.1186/s13326-017-0153-x

Katz, SL, Barnas, KA, Diaz, M and **Hampton, SE.** 2019. Data system design alters meaning in ecological data: Salmon habitat restoration across the US Pacific Northwest. *Ecosphere*, 10(11): e02920. DOI: https://doi.org/10.1002/ecs2.2920

Kempler, S and **Mathews, T.** 2017. Earth science data analytics: Definitions, techniques and skills. *Data Science Journal*, 16: Article 6. DOI: https://doi.org/10.5334/dsj-2017-006

Krovetz, R. 1997. Homonymy and polysemy in information retrieval. In: *35th Annual Meeting of the Association for Computational Linguistics and 8th Conference of the European Chapter of the Association for Computational Linguistics*. Madrid, Spain on 7–12 July 1997, pp. 72–79. DOI: https://doi.org/10.3115/976909.979627

Krovetz, R and **Croft, WB.** 1992. Lexical ambiguity and information retrieval. *ACM Transactions on Information Systems*, 10(2): 115–141. DOI: https://doi.org/10.1145/146802.146810

Liao, X and **Zhao, Z.** 2019. Unsupervised approaches for textual semantic annotation, a survey. *ACM Computing Surveys (CSUR)*, 52(4): 1–45. DOI: https://doi.org/10.1145/3324473

Madin, JS, Bowers, S, Schildhauer, MP and **Jones, MB.** 2008. Advancing ecological research with ontologies. *Trends in Ecology & Evolution*, 23(3): 159–168. DOI: https://doi.org/10.1016/j.tree.2007.11.007

Madin, J, Bowers, S, Schildhauer, M, Krivov, S, Pennington, D and **Villa, F.** 2007. An ontology for describing and synthesizing ecological observation data. *Ecological Informatics*, 2(3): 279–296. DOI: https://doi.org/10.1016/j.ecoinf.2007.05.004

Magagna, B, Schindler, S, Stoica, M, Moncoiffe, G, Devaraju, A and **Pamment, A.** 2021. I-ADOPT Framework ontology. Retrieved from https://w3id.org/iadopt/ont/0.9.1.

Mantovani, A, Piana, F and **Lombardo, V.** 2020. Ontology-driven representation of knowledge for geological maps. *Computers & Geosciences*, 139: 104446. DOI: https://doi.org/10.1016/j.cageo.2020.104446

McGuinness, DL and **Van Harmelen, F.** 2004. OWL web ontology language overview. *W3C Recommendation*, 10(10): 2004.

Miles, A and **Bechhofer, S.** 2009. SKOS simple knowledge organization system reference. Available at https://www.w3.org/TR/2009/REC-skos-reference-20090818/ [Last accessed 11 January 2021].

O'Brien, M, Jones, M, Schildhauer, M, Hou, S, Mecum, B, McCusker, J and **McGuinness, D.** 2023. (Dataset) Results of semantic queries for "carbon cycling" for datasets in the DataONE catalog ver 2. *Environmental Data Initiative*. DOI: https://doi.org/10.6073/pasta/c93d87c2000715eaa2f70d079965c6a5 (Accessed 2023-03-29).

Open Biological and Biomedical Ontologies Technical Working Group. 2020. Principles: Overview. Available at http://www.obofoundry.org/principles/fp-000-summary.html [Last accessed 23 September 2020].

Rogers, FB. 1963. Medical subject headings. *Bulletin of the Medical Library Association*, 51: 114–116.

Schildhauer, M. (n.d.) *Semantic annotations*. Available at https://arcticdata.io/semantic-annotations [Last accessed 19 October 2023].

Schildhauer, M, Chong, S, O'Brien, M, Mecum, B and **Jones, MB.** 2019. Semantic approaches to enhancing data findability and interoperability in the NSF DataONE and Arctic Data Center Data Repositories. *American Geophysical Union Fall Meeting*, 2019, IN22C-19. Available at https://ui.adsabs.harvard.edu/abs/2019AGUFMIN22C..19S/abstract [Last accessed 31 August 2020].

Smith, B. 2012. Ontology. In: Hurtado, G and Nudler, O (eds.), *The Furniture of the World: Essays in Ontology and Metaphysics*. New York: Brill Rodopi. pp. 47–68. DOI: https://doi.org/10.1163/9789401207799_005

Vanderbilt, KL, Gries, C and **O'Brien, MC.** 2020. Annotating metadata to improve data discovery and reuse. In: *2020 ESA Annual Meeting*, 3–6 August. Ecological Society of America. Available at https://eco.confex.com/eco/2020/meetingapp.cgi/Paper/85318 [Last accessed 31 August 2020].

Wilkinson, MD, Dumontier, M, Aalbersberg, IJ, Appleton, G, Axton, M, Baak, A, Blomberg, N, Boiten, J-W, da Silva Santos, LB, Bourne, PE and **others.** 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3(1): 1–9. DOI: https://doi.org/10.1038/sdata.2016.18

TO CITE THIS ARTICLE:
Chong, SS, Schildhauer, M, O'Brien, M, Mecum, B and Jones, MB. 2024. Enhancing the FAIRness of Arctic Research Data Through Semantic Annotation. *Data Science Journal*, 23: 2, pp. 1–14. DOI: https://doi.org/10.5334/dsj-2024-002

Submitted: 05 April 2023
Accepted: 28 November 2023
Published: 17 January 2024

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.