RESEARCH PAPER

An Analysis of Crosswalks from Research Data Schemas to Schema.org

Mingfang Wu[1†], Stephen M. Richard[2], Chantelle Verhey[3], Leyla Jael Castro[4], Baptiste Cecconi[5], Nick Juty[6]

[1]Australian Research Data Commons, Melbourne, Victoria 3145, Australia

[2]US Geoscience Information Network, Neward DE 19716-7501, USA

[3]International Science Council, World Data System, Victoria BC V8N 1V8, Canada

[4]ZB MED – Information Centre for Life Sciences, Cologne 50931, Germany

[5]Observatoire de Paris-PSL, Paris Astronomical Data Center, Paris 75001, France

[6]Department of Computer Science, The University of Manchester, Oxford Road, Manchester M13 9PL, UK

Keywords: metadata schema; Schema.org; metadata interoperability; FAIR (meta)data; metadata schemas crosswalk; research data schemas

Citation: Wu, M.F., Richard, S.M., Verhey, C., et al.: An analysis of crosswalks from research data schemas to Schema.org. Data Intelligence 5(1), 100-121 (2023). doi:10.1162/dint_a_00186
Received: November 12, 2021; Revised: April 7, 2022; Accepted: May 7, 2022

ABSTRACT

The increased number of data repositories has greatly increased the availability of open data. To enable broad discovery and access to research dataset, some data repositories have begun leveraging the web architecture by embedding structured metadata markup in dataset web landing pages using vocabularies from Schema.org and extensions. This paper aims to examine metadata interoperability for supporting global data discovery. Specifically, the paper reports a survey on which metadata schema has been adopted by participating data repositories, and presents an analysis of crosswalks from fourteen research data schemas to Schema.org. The analysis indicates most descriptive metadata are interoperable among the schemas, the most inconsistent mapping is the rights metadata, and a large gap exists in the structural metadata and controlled vocabularies to specify various property values. The analysis and collated crosswalks can serve as a reference for data repositories when they develop crosswalks from their own schemas to Schema.org, and provide the research data community a benchmark of structured metadata implementation.

† Corresponding author: Mingfang Wu (Email: Mingfang.Wu@ardc.edu.au; ORCID: 0000-0003-1206-3431).

1. INTRODUCTION

In recent years, it has become more and more common to share research data together with its corresponding description through metadata, thanks to initiatives such as Open Science and the FAIR (Findable, Accessible, Interoperable and Reusable) data principles [26]. To make data publicly accessible, researchers and data collectors deposit their datasets into a data repository and provide metadata that conforms to the repository's metadata schema[①]; data repositories or metadata aggregators provide data discovery capabilities to make dataset discoverable through indexed metadata. With the increase of datasets managed in data repositories, some challenges arise including exchanging metadata, discovering relevant datasets, and supporting (semi)automatic metadata processing [29].

Data repositories typically host metadata, embed metadata in a web page and publish the web page on the Web to make the dataset discoverable; such a web page, as shown in Figure 1a, is referenced as a metadata landing page. Like any other web pages, a web landing page is encoded with HTML tags, optimised for human readability. Before the recent explosion in commercial web index and search technology, repositories also offered access to structured, machine-readable metadata for their holdings using various metadata content and serialization schemes such as Dublin Core XML, Ecology Markup Language (EML), the U.S. Content Standard for Digital Geospatial Metadata (CSDGM), ISO 19115/19139, and so on. This metadata was accessed through a standard API like Open Archives Initiative Protocol for Metadata Harvest (OAI-PMH) or the Open Geospatial Consortium Catalogue Service for the Web (OGC-CSW).

Around 2004, developers started introducing semantic markup in HTML documents to add information about the web page subject and content to improve the display of search results, making it easier for people to find the right web pages. In 2011, a consortium of search engines including Bing, Google, Yahoo! and Yandex began developing a vocabulary of entities and properties that could be used in this semantic mark-up to make it interoperable across browser systems [11]. The Schema.org vocabulary is the outcome of this effort, with version 1 released in 2013. This initial release included an Entity for describing datasets (https://www.w3.org/wiki/WebSchemas/Datasets), which was significantly revised in 2016 (https://github.com/schemaorg/schemaorg/pull/1247).

This approach of publishing machine-readable metadata, i.e., structured metadata as shown in Figure 1b, brings new opportunities for making research data FAIRer. For instance, the use of these common vocabularies makes it easier for commercial web search engines like Google dataset search[②], or any metadata aggregators, to crawl and index metadata across data repositories globally in a more useful, consistent and robust way. The interoperability of metadata sharing the same schema allows metadata from different sources to be harvested and indexed without any intermediate mapping between schemas. Furthermore, it makes it easier to create federated queries across resources from different sources relevant to a research need. Metadata aggregators are exploring new methods for metadata syndication via the web architecture. The NSF

[①] We use the term 'schema' instead of 'scheme' throughout the paper, as this study focuses on the semantic meaning of data properties.

[②] https://datasetsearch.research.google.com/

EarthCube GeoCODES platform[③] is indexing schema.org metadata in landing pages from 12 US NSF data facilities. DataCite has already offered to crawl metadata through its embedded web page [10], DataOne[④] and ARDC's catalogue service Research Data Australia[⑤] are planning to offer a similar service.

However, these opportunities also come with new challenges. Schema.org provides a domain agnostic vocabulary to describe common data entities. By design, Schema.org expects and has enabled domains of practice to extend this core vocabulary [11]. Similar to other domains of practice, research data communities have their own needs for extending Schema.org core to describe research data and its relationships to other resources. These extensions include, for instance, specific data types and their corresponding properties pertaining to a particular domain as well as support for persistent identifiers to meet needs for a specific community: for example, bioschemas.org [12] for life sciences, science-on-schema.org for earth and environmental sciences [14] and CodeMeta[⑥] for research software.

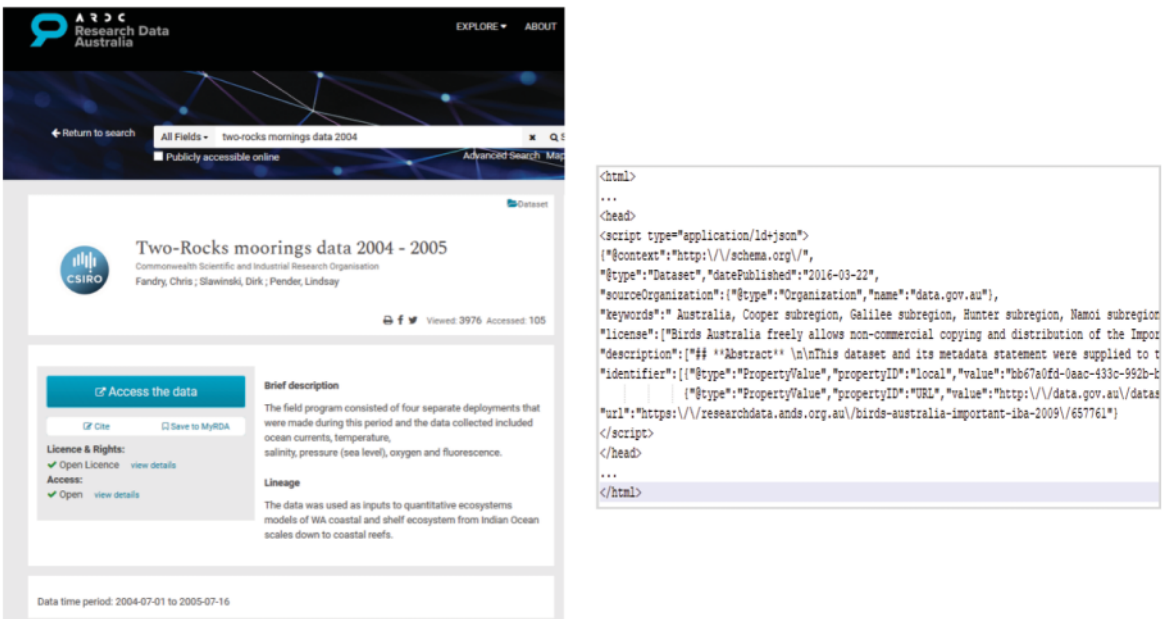


Figure 1. a) Left: an example of metadata landing page—metadata is published and embedded in a webpage for human users to read, b) Right: Some metadata as shown in the left html page is marked up and embedded in the source html for machine to access and parse.

To investigate interoperability and usability of Schema.org for describing research data, we collected 14 crosswalks from research data schemas to Schema.org [28], this crosswalk is a crucial step for repositories to publish structured metadata [27]. A schema crosswalk is commonly expressed as a table showing

③ https://geocodes.earthcube.org/
④ https://www.dataone.org/
⑤ https://researchdata.edu.au/
⑥ https://codemeta.github.io/

equivalent terms across one or more data schemas. To source research data schemas, we used a survey asking participating data repositories to share any crosswalk they had, as well as gaps and challenges that they identified while creating the crosswalk. For schema providers, we used openly published crosswalks available on the Internet; in particular, we found crosswalks corresponding to DCAT, Dublin Core and ISO19115 to Schema.org[7]. This collection of crosswalks helps us to identify and bridge gaps in research data communities when they mapped their metadata schemas to Schema.org.

This paper covers a report on the survey and an analysis of the crosswalks. The sections below are organised as follows: we review the type of metadata schemas for research data in Section 2, present the analysis of a survey and crosswalks in Section 3 and conclude the paper with a discussion of findings in Section 4.

2. METADATA SCHEMAS FOR RESEARCH DATA

2.1 General and Discipline-specific Metadata

There are many metadata standards for documenting research datasets; Wallis et al. [25] analysed 9 metadata schemas for describing scientific data and synthesised 22 metadata-related goals. In general, a metadata schema should address the seven requirements for metadata schemas of all resources—abstraction, extensibility, flexibility, modularity, comprehensiveness, sufficiency, and simplicity; and four requirements for any schema to support data interchange, retrieval, achieving and publication.

The metadata directory implemented by the RDA Metadata Standard Directory Working Group includes about 65 standards[8], ranging from general to extremely discipline specific [1]. General metadata schemas, for example, Data Catalogue Vocabulary (DCAT) and Dublin Core include data properties that are common to almost all types of dataset. This general metadata can be widely adopted and easily used by metadata providers, and supports broad data discovery use cases from data seekers, regardless of their research areas.

Discipline specific metadata, for example, the Data Documentation Initiative (DDI for Social and Behavioral Science data) and the Space Physics Archive Search and Extract (SPASE for heliophysics data), usually include properties from general metadata standards, and provide additional properties and richer vocabularies to allow detailed and more granular contextual information. This enriched information increases data discovery efficiency and effectiveness for those with domain knowledge, and assists the assessment of data reusability.

It is common practice for data repositories to publish metadata for their holdings, allowing it to be harvested by aggregating metadata catalogs that offer indexing and user interfaces to support data search. Such aggregation typically involves a mapping or crosswalk between metadata schemes or profiles used by the various contributing repositories if there isn't a schema agreed by all repositories for exchanging

[7] ISO19115—DCAT—Schema.org mapping: https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping
[8] http://rd-alliance.github.io/metadata-directory/standards/

metadata. This landscape may change due to major search engines starting to harvest structured metadata using the standardized, schema.org vocabulary embedded in metadata landing pages that can be parsed and interpreted by machine, to provide more accurate results and richer presentation of results [4].

2.2 Schema.org Vocabulary and Structured Metadata

Schema.org is among the most visible metadata vocabularies on the open Web, according to NISO [19]. The driving factor in the design of Schema.org was to make it easy for webmasters to publish information with a single schema for a wide range of topics that included people, places, events, products and so on [11]. Schema.org is a general schema or a set of vocabularies, the current version (V13.0, 2021-07-07) consists of about 792 types (as RDF classes) and 1447 properties. The W3C Schema.org Community Group, that is governed by a steering group[9], is the main forum for the schema collaboration and the development new types and properties can be added if there is community need and supporting use case, for example, the new type 'LearningResource' was added as a subtype of 'CreativeWork' in 2020 July release (9.0)[10][11]. As another example, Bioschemas[12], focusing on life science, have successfully incorporated many biomedical terms into the schema.org vocabulary. The CodeMeta project[13] has developed the CodeMeta vocabulary for the description of software; 58 out of 68 Codemeta properties are from existing Schema.org vocabulary, 10 proposed new properties are based on the analysis of crosswalk from 23 software metadata, vocabulary and ontology to Schema.org. There are also a steering group and communities who support developing conventions for usage of the data model and guidelines for consistently implementing the data model. For example, the Schema.org Cluster of the Earth Science Information Partners (ESIP) working to develop best practices and to provide education and outreach to the Earth science community for web accessible structured data[14] [14], The Ocean InfoHub Project[15] provides an architecture solution for providing a Schema.org based interoperability layer and supporting technology to allow existing and emerging ocean data and information systems to interoperate with one another.

In order to make data widely discoverable, many research data repositories have started to implement structured metadata markup in their metadata landing page. As of March 26, 2020, Google dataset search has indexed 31M datasets from 4,600 domains, where the top 10 domains include data.gov, figshare.com, datacite.org. Geosciences and social sciences together accounted for 45% of the datasets, followed by biology (~15%) and other research topics [21]. Search results include those from NASA, NOAA, and many research repositories such as Harvard's Dataverse repository [20]. This approach allows for broader dissemination of metadata throughout the community to promote discoverability of datasets.

[9] https://schema.org/docs/about.html
[10] Schema.org Releases: https://schema.org/docs/releases.html
[11] Learning Resource Metadata is go for Schema: https://blogs.pjjk.net/phil/lrmi-in-schema/
[12] https://bioschemas.org
[13] https://codemeta.github.io/
[14] https://github.com/ESIPFed/science-on-schema.org/blob/master/guides/Dataset.md
[15] The Ocean InfoHub Project: https://book.oceaninfohub.org/index.html

2.3 Metadata Interoperability

The 'I' in 'FAIR' represents "interoperable" and is one of the four FAIR data principles [26], which apply to both data and metadata. According to this principle, metadata should use community agreed standards and vocabularies, and contain links to related information using persistent identifiers. Because there exist a number of community agreed metadata schemas for meeting specific community needs, mapping between schemas is necessary to make it possible for repositories to exchange and share metadata records [24].

There are different types of metadata interoperability, for example, Nilsson et al. [18] proposed four interoperability levels for Dublin Core Metadata. For a data repository to implementing interoperable metadata, we adopt the three levels of metadata interoperability proposed by Chan and Zeng [6]:

- Schema level—efforts are focused on the elements of the schemas, common results may include crosswalks, application profiles, derived element sets, et al.;
- Record level—efforts are intended to integrate the metadata records through the crosswalk of elements, common results include converted records, new records resulting from combined values of existing records; and
- Repository level—efforts are focused mapping values associated with particular elements, the results enable cross-collection searching.

We focus our analysis of crosswalks at the schema level: the elements of the schemas, being independent of any applications. In particular, we will apply crosswalk to analyse the interoperability among studied schemas. A crosswalk (or a mapping) is a chart or table (visual or virtual) that represents the semantic or technical mappings of data elements from one schema (source schema) to data elements in another schema (target schema) that has a similar function or meaning. The crosswalks guide record level interoperability, which enables repository level interoperability so that heterogeneous repositories can be searched simultaneously with a single query as if there were a single repository [2].

3. ANALYSIS OF MAPPINGS FROM RESEARCH DATA SCHEMAS TO SCHEMA.ORG

As discussed above a crosswalk attempts to map equivalent or comparable metadata elements from two schemas. We acknowledge that a crosswalk developed by a specific repository or a schema development community would better reflect a proper and realistic mapping, as those repositories and communities can provide a better interpretation of their implemented metadata terms. For this reason, we launched the survey "Current practices in using schemas to describe research datasets"[®] on 27th Feb. 2019 to gather information on how Schema.org is applied by data repositories to describe research data and related resources. We envisaged the gathered information would help repositories and the proposed Research Metadata Schema WG understand current practices, identify commonalities, gaps and barriers in using schemas for describing and discovering research datasets.

[®] Survey on current practices in using schemas describing datasets: https://docs.google.com/spreadsheets/d/19cuspUioXp1QgxGFph6tjjvNB6JHSOzIFxh8UCc3aVM

In Section 3.1, we highlight relevant parts of the survey and indicate which schemas are adopted or implemented by participating respondents, followed by our analysis of crosswalks from the available mappings to Schema.org.

3.1 Survey on Repository's Metadata Schema and the Implementation of Schema.org

Twenty-two organisations/data repository representatives participated in the survey. One respondent failed to answer the survey questions, so that submission has been excluded from this summary. As shown in Table 1, six of 21 responses are from the general repositories covering all domains: four of them are either based on or direct adoption of the DataCite schema; one is an application profile of DCAT—DCAT-AP, while the other follows the Registry Interchange Format—Collections and Services (RIF-CS) schema, which is a profile of ISO 2146, originally developed for library registry services now used as a data interchange format.

Among the 13 disciplinary repositories or projects, five are from the domain of Geoscience and Arctic Research and have adopted the ISO19115 schema or ISO19115 compatible schema (EML). ISO19115 is an internationally adopted schema for describing geographic information and services. ISO19115 provides information about the identification, the extent, the quality, the spatial and temporal schema, spatial reference, and distribution of digital geographic data[17]. One Social and Behaviour Science repository adopted the international standard 'Data Documentation Initiative' (DDI), for describing the data produced by surveys and other observational methods in the social, behavioural, economic, and health sciences[18]. The remaining nine disciplinary repositories and the two "other" repositories adopted community developed profiles or schemas. Most of them are compatible or interoperable with international standards, for example, the cultural heritage datasets in the 'Other' category defines a metadata profile based on Schema.org, DCAT and VoID[19], while the European Clinical Research Infrastructure Network (ECRIN) schema is an extension of DataCite [5], and GigaDB from the Life Sciences and Biomedical domain can export metadata in general purpose metadata such as DataCite and Schema.org.

We observe the following two trends from the survey responses:

1) Newer schemas tend to adopt existing commonly used elements. For example, the Data Catalogue Vocabulary (DCAT) makes extensive use of elements from Dublin Core: 20 out of 29 terms for describing a dataset are from Dublin Core[20]. The Bioschemas profiles adopt 5 mandatory properties and 8 recommended properties from Schema.org[21].

[17] https://www.dcc.ac.uk/resources/metadata-standards/iso-19115
[18] Document, Discover and Interoperate (DDI): https://ddialliance.org/
[19] http://data.europeana.eu
[20] Data Catalog Vocabulary (DCAT)—Version 3: https://www.w3.org/TR/vocab-dcat-3/
[21] Bioschemas: https://bioschemas.org/

2) Repositories, regardless of discipline, general, or specific, tend to use a general-purpose schema but also support domain specific standards or vocabularies. For example, the Dataverse project[20] supports general citation metadata compatible with the DataCite metadata schema [8] and DCMI metadata terms but also a suite of domain specific metadata for Geoscience, Social Science and Humanities. RIF-CS supports subject vocabularies from a range of disciplines for satisfying a range of data discovery needs. This observation also applies to discipline specific repositories, for example, the DAta Tag Suite (DATS), a data description model adopted by DataMed[21], has both core elements and additional elements: the core elements are generic and applicable to any type of dataset, while the additional elements are specific for life, environmental and biomedical science domains [22].

The observed trend is that general repositories adopt general purpose standards that support data discovery use cases at a high level for data searches across domains providing. Domain repositories adopt schemes that are compatible with general metadata profiles for metadata interoperability, but add elements to support a range of more granular disciplinary queries for more precise data discovery within a domain.

3.2 Analysis of the Mappings

We collected the 14 crosswalks from the following schemas to Schema.org through the survey and other publicly available crosswalks: B2FIND, DCAT-AP, DCAT, RIF-CS, Core DATS, Dataverse, DDI Codebook 2.5, DC & DCTerms, BioSchema, SPASE, DataCite, ISO-19115-1:2014, EOSC/EDMI, ECRIN Clinical Research Metadata Schema. We aligned the crosswalks with the mapped Schema.org properties. In total, there were 232 terms from the 14 crosswalks being mapped to 34 Schema.org properties.

Since the survey results were collected, some crosswalks may have been updated (e.g DCAT to Schema.org alignment) and some schemas (including Schema.org) may have been revised with additional properties. In October 2021, the first author cross checked all crosswalks, as well as referencing publicly available crosswalks. These included, for example, ISO-19115 (from this W3C group[23] and Habermann [13]), DCAT alignment with Schema.org[24], DataCite Schema to Dublin Core mapping[25], the CodeMeta crosswalks[26]. During the writing of this paper, the second author also added a mapping from ISO19115-1 to Schema.org. For the purposes of this analysis we used this subsequent mapping as it covers more elements than the original ISO-19115-1:2014 to schema.org mapping we collected from this website[27]. This resulted in 385 properties from the 14 crosswalks being mapped to the 40 Schema.org properties.

[20] https://guides.dataverse.org/en/latest/user/appendix.html
[21] https://datamed.org/
[23] ISO19115—DCAT—Schema.org mapping: https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping
[24] https://www.w3.org/TR/vocab-dcat-3/#dcat-sdo
[25] https://schema.datacite.org/meta/kernel-4.4/doc/DataCite_DublinCore_Mapping.pdf
[26] https://github.com/codemeta/codemeta/blob/master/crosswalk.csv
[27] https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping

Table 1. Supported schemas and crosswalks to Schema.org.

Domain (The number of responses)	Schema(s) supported by repository (direct survey responses)	Reference links to the crosswalk(s) or to any documentation about the crosswalk(s), where provided
All domains (6)	B2FIND established a generic, non-hierarchical metadata schema. This schema is based on the DataCite Metadata Schema. Additional elements of the B2FIND schema include "Discipline", "Instrument" and "TemporalCoverage".	Mappings: http://b2find.eudat.eu/guidelines/mapping.html
	Registry Interchange Format—Collections and Services (RIF-CS) https://www.ands.org.au/online-services/rif-cs-schema, supporting (published) vocabularies (e.g. subject headings) from all disciplines.	RIF-CS to Schema.org crosswalk: https://documentation.ands.org.au/display/DOC/RIF-CS+to+Schema.org+crosswalk
	Dataverse exports dataset metadata in general purpose standards (Schema.org, DC element and terms, DataCite, OAI-ORE) and domain specific standards (DDI, VO Resource, ISA-Tab)	Dataverse schemas to Schema.org crosswalk: https://docs.google.com/spreadsheets/d/10Luzti7svVTVKTA-px27oq3RxCUM-QbiTkm8iMd5C54/edit#gid=0&range=P1)
	We use DCAT-AP (an application profile of DCAT, used at the European level as a cross-domain metadata interchange format), which we extended in order to address cross-domain requirements of research data—e.g., data citation. A description of the DCAT-AP profile: https://doi.org/10.1145/3151759.3151810 https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_27	https://ec-jrc.github.io/dcat-ap-to-schema-org/
	DataCite and Schema.org	Not provided
	DataCite	Not provided
Geoscience and arctic research (5)	P418 is focused on schema.org (type Dataset) but is already looking to extend this in collaboration with ESIP. Note we are also looking at DCAT and expect to leverage its schema.org mapping. Also, through the above we want to work in OWL time approaches for time and also leverage CSIRO work on Geologic Time ontology.	Guide: https://github.com/ESIPFed/science-on-schema.org
	schema.org, DCAT, (NOAA) ISO19115-2: https://service.ncddc.noaa.gov/rdn/www/metadata-standards/documents/MI-Metadata.pdf	Guide: https://github.com/ESIPFed/science-on-schema.org
	We export ISO 19115 (conforming to NASA MENDS profile, https://git.earthdata.nasa.gov/projects/EMFD/repos/iso-schemas/browse), DIF (https://earthdata.nasa.gov/user-resources/standards-and-references/directory-interchange-format-dif-standard), and custom JSON.	Not provided

Table 1. *Continued*

Domain (The number of responses)	Schema(s) supported by repository (direct survey responses)	Reference links to the crosswalk(s) or to any documentation about the crosswalk(s), where provided
	Ecological Metadata Language (EML), which is closely compatible with ISO 19115. Element contents are currently only loosely constrained if at all (e.g. when describing variables/parameters, blank fields are presented for name, label, definition, etc.) https://search.dataone.org, https://arcticdata.io	We have mapped many of our EML metadata fields, as well as other environmental metadata schemas, to Schema.org.
	The Interdisciplinary Earth Data Alliance at Lamont Doherty Earth Observatory developed a transform to map ISO19115-1 metadata to schema.org for inclusion in metadata landing pages from the EarthChem Library, Marine Geoscience Data System (MGDS), and US Antarctic Program (USAP) data repository. The Mapping is embedded in an XSLT transform.	XSLT transform: https://github.com/usgin/metadataTransforms/blob/master/iso-19139-to-HTMLwSDO/ISO19139ToSDODatasetStandalone1.0.xslt
Medicine (2)	CEDAR has a general-purpose model for all its metadata templates. that is described at https://metadatacenter.org/tools-training/outreach/cedar-template-model. In general, CEDAR doesn't endorse a specific schema for the data it can collect but instead we give our users the freedom to choose and/or create their own schema through building a form. However, we are now experimenting with providing some pre-built forms for our users and we are testing with respect to schema.org.	Core DATS and HCLS to Schema.org crosswalks: https://docs.google.com/spreadsheets/d/16HNJVKUdueVIPEdcp3x2HXI0RJ4zrlpQWrTIkAf-lB4/edit?usp=sharing
	European Clinical Research Infrastructure Network (ECRIN): Metadata has been created for a catalogue of data objects from clinical research—many of which will be under managed access. Schema is an extension of DataCite, with additional data points to describe a) access arrangements, b) associated consent / de-identification, pseudonymisation, and c) basic characteristics of their source study. Most recent published version of the proposed schema is at https://zenodo.org/record/1312539#.XHfWdOj7Suc, but further changes are likely as the project proceeds.	Not provided
Biology (2)	We are the group behind the DATS metadata model. Full description of its development, the rational, and alignment (crosswalks) with other metadata models at https://doi.org/10.1038/sdata.2017.59	Not provided
	All specifications, JSON-LD serializations and examples are freely available from this Github repository: https://github.com/datatagsuite plans to adopt schema.org and Bioschemas profiles http://bioschemas.org/	https://github.com/BioSchemas

Table 1. *Continued*

Domain (The number of responses)	Schema(s) supported by repository (direct survey responses)	Reference links to the crosswalk(s) or to any documentation about the crosswalk(s), where provided
Material Sciences and Engineering (1)	schema.org + an internal json schema that we plan to align with the B2SHARE schema (+community extensions). Example of such a schema: https://b2share.eudat.eu/api/communities/e1800bc8-780e-4617-a7b6-2312cb6190c4/schemas/0#/json_schema Repository portal: https://archive.materialscloud.org/	See e.g. embedded JSON-LD in https://archive.materialscloud.org/2019.0007/v3
Social and behaviour science (1)	DDI: https://datacatalogue.cessda.eu/	Not provided
Agriculture;Forestry;Horticulture;Veterinary Medicine (1)	Data Inrae is based on Dataverse (with regular update), Data Inrae exports dataset metadata in Dublin Core, DDI, JSON, Schema.org JSON-LD https://data.inra.fr	Not provided
Life Sciences and Biomedical (1)	GigaDB exports dataset metadata in general purpose standards (Schema.org and DataCite), it is also possible to export in the domain specific standard ISA-Tab, or as our own complete XML metadata which is far more extensive than any of the standards.	Documentation of example of schema.org markup can be seen on any metadata landing page, e.g. http://dx.doi.org/10.5524/100552
Others (2)	Defines a metadata profile allowing cultural heritage datasets to be described with Schema.org, DCAT and VoID. http://data.europeana.eu	https://github.com/nfreire/Open-Data-Acquisition-Framework/blob/master/opaf-documentation/SpecifyingLodDatasetForEuropeana.md DCAT and Schema.org https://webgate.ec.europa.eu/CITnet/stash/projects/ODCKAN/repos/dcat-ap-to-schema.org/browse
	NASA's Unified Metadata Model (UMM; https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm) and schema.org. UMM goes to the NASA Common Metadata Repository for search across NASA data centers. We use the schema.org data on the landing pages to enhance other forms of search. Portal: ORNL Distributed active archive center for biogeochemical dynamics; https://daac.ornl.gov	Not provided

Categories of mapped terms

We classified the 40 mapped Schema.org properties/terms into 6 categories from the NISO (2004) metadata classification model. As shown in Table 2, we use three top level categories: descriptive metadata, administrative metadata and structural metadata; administrative metadata is further classified into technical metadata, rights metadata and preservation metadata. We summarise the analysis of mapped terms as follow:

- Descriptive metadata: Most of the mapped terms (17 out of 40) fall into the descriptive metadata category. The mapped descriptive terms cover six of seven recommended citation metadata from the DataCite guide[2]:
  Creator (PublicationYear): Title. Version. Publisher. (resourceTypeGeneral). Identifier
  The citation term "*resourceTypeGeneral*" (recommended) is the only term not explicitly included in the mapping, and we infer it to be of the type: *dataset,* since we asked for and collected all mappings from schemas for descripting data. All 14 source schemas include the 6 mapped citation metadata, except for the term "version" and "publisher" that occurred in the 13 out 14 source schemas.
- Administrative metadata:
  - Technical metadata: '*encodingFormat*' and '*contentSize*' are the two mapped technical metadata terms by the majority of the source schemas. The mappings are consistent: the term '*format*' is used by 9 out of 13 source schemas, the exact term '*encodingFormat*' by one source schema, and the alternate terms '*resource file type*', '*mediaType*', '*distributionFormat*' each by a source schema.
  - Rights metadata: There are three mapped terms in rights metadata. The property "*license*" has a mapping from 12 source schemas, however, five of them have the original term "*rights*". The term "*rights*" is the only one from the 15 Dublin Core terms (http://purl.org/dc/elements/1.1) that doesn't have an exact mapping in Schema.org. In Dublin Core, "*rights*" is defined as "information about rights held in and over the resource", "*license*" is subproperty of "*rights*" and has the definition "a legal document giving official permission to do something with the resource". According to this definition, the closest semantically matched term in Schema.org is "*copyrightHolder*" (https://schema.org/copyrightHolder): The party holding the legal copyright to the CreativeWork.
  - Preservation Metadata: There are 11 mapped preservation metadata terms: five of them are dates about data creation, modification, availability and copyright; another five about data access method or location; and one about data (observation/process/reprocess) frequency. The mappings of the dates and the access methods are consistent, except that the term '*expectedArriveFrom/expectedArriveUntil*' is mapped from four different terms: '*distribution date*', '*released date*', '*available*', and '*embargo*'.
- Structural metadata: The seven mapped terms in the structural metadata category describe the citation relation between a dataset and its related academic articles/report ('*citation*'), duplicated datasheet ('*sameAs*'), a clear relation between two datasets ('*isPartof/hasPart*', '*isBasedOn*') and general relation between two datasets ('*isRelatedTo*', '*mentions*').

[2] https://schema.datacite.org/meta/kernel-4.4/doc/DataCite-MetadataKernel_v4.4.pdf

We also examine how many mapped terms are recommended by the Google dataset search guide[8]. The Google guide recommends 23 properties (in italics in Table 2) to be included in structured data. Three of them are required terms ("*name*", "*description*", "*distribution.contentURL*"), while the other 20 are recommended. The 23 terms are distributed among all six NISO metadata categories, and are mapped by more than half of source schemas, especially those falling into the descriptive metadata category. Note that this analysis is on the schema level, and does not take into account whether a repository has implemented a property value at the metadata record level. Benjelloun et al. [3] from Google Research analysed the percentage of datasets in their index with specific properties, showing that the property "*name*" and "*description*" both have 100% coverage, followed by "*provider*" (84.59%), "*keywords*" (80%) and "*URL*" (68.08%), while all other recommended properties had less than 50% coverage (e.g. "*authors*"—14.12%, "*isAccessibleForFree*"—3.04%). This indicates that even if there is a property mappable at the schema level, a repository may decide not to implement that mapping or to populate that property with a value. The reason, most likely, is that the repository does not have sufficient records requiring that property to warrant its implementation.

Table 2. Classification of the mapped Schema.org properties or terms.

NISO Metadata Type		Schema.org properties (The numbers in brackets indicate the number of crosswalks that have a term mapped to the schema.org property. Properties in italics are those recommended by the Google dataset search guide[8].)
Descriptive metadata: For finding or understanding a resource		*Identifier* (14), *name* (14), *description* (14), *creator* (14), *alternameName* (9), datePublished (13), *version* (13), *keywords* (13), about or subjectOf (8), inLanguage (10), *temporalCoverage* (11), *spatialCoverage* (11), *variableMeasured* (7), publisher (12), contributor (10), *funder* (10), producer (8), *measurementTechnique* (6)
Admin. Metadata	**Technical metadata** For decoding and rendering files	*encodingFormat* (13), contentSize (8)
	Rights metadata Intellectual property rights attached to content	copyrightHolder (4), *isAccessibleForFree* (6), *license* (12)
	Preservation Metadata Long-term management of files	*contentUrl* (8), *URL* (14), *distribution* (9), contactPoint (10), copyrightYear (5), dateCreated (11), dateModified (11), expectedArriveFrom/expectedArrivalUntil(4), repeatFrequency (5), includeInDataCatalog (8)
Structural metadata Relationships of parts of resources to one another		*citation* (12), *sameAs* (8), mentions (4), isBasedOn (7), *isPartOf* (10), *hasPart* (10), isRelatedTo (8)

[8] Google Search Central Documentation: Using structured data: https://developers.google.com/search/docs/advanced/structured-data/dataset

[8] https://developers.google.com/search/docs/advanced/structured-data/dataset

Gap analysis

From the survey, there are structural metadata elements that are recommended by source schemas that do not have mappings to Schema.org. These include elements that clearly describe:

- Relationships between datasets, for example: hasVersion, isNewVersionOf, isContinuedBy, isOriginalFromOf, isDerivedFrom (from DataCite);
- Relationships between a dataset and responsible agent, for example: hasFunder, isFundedby, isCompiledBy, isOwnedBy, hasPrincipleInvestigator;
- Relationships between a dataset and the activity by which is was collected, for example: dataset -> Cruise, dataset -> study design; and
- Relationships between a dataset and instrument/software/other services used to produce the data, for example, isProducedBy/produces, isPresentedBy/presents, isOperatedOnBy/operatedOn, isAnnotatedBy/annotate (from RIF-CS).

These structural, relation metadata properties are more granular than the PROV-O Ontology [16]. These gaps reflect both the difference between documentation needed to describe scientific datasets for research and that for more commercial data published on the Web (e.g. movies, businesses, product catalogs, etc.), and the difference between general data schemas and discipline specific schemas.

From information gathered from the survey and through inspection of the source schemas, we observe that:

- Controlled vocabularies, thesauri or code lists are used to specify property values for various elements in the source schema. Schema.org doesn't offer any vocabularies for property values, but the serieralization of Schema.org allows it to incorporate external vocabularies. For example, when populating the property schema:keyword or schema:about, one can specify a text string (either from a vocabulary or not) that can facilitate discovery but not interoperability, while an optimal way is to specify a URI reference to a term from a controlled vocabulary. There is a proposal to add a DefinedTerm element (https://schema.org/DefinedTerm) that could be substituted for plain text values to provide a URI along with the term, but this has not, as yet, been formally adopted into schema.org.
- A controlled vocabulary is a set of pre-defined, authorised terms that are used to specify a property value so that consistency can be achieved within and across repositories. A controlled vocabulary can be standard and controlled by an authoritative organisation (for example, Library of Congress Subject Headings, Australia and New Zealand Standard Research Classification—ANZSRC), a locally defined subset of a standard vocabulary, or a locally defined vocabulary [23]. Ideally, terms in the vocabulary have dereferenceable URIs for unambiguous identification. This case requires a controlled vocabulary to be openly accessible, referenceable and identifiable with a unique and persistent identifier to the vocabulary, for each term in the vocabulary [7]. Research Vocabularies Australia[@] is an example of such a service for finding, accessing and reusing vocabularies.

[@] ARDC Research Vocabularies Australia: https://vocabs.ardc.edu.au/

- There are semantically equivalent properties which are named differently among the schemas. For example, schema:variableMeasured, dats:dimensions, spase:parameter and sosa:observedProperty (DCATv3) all have the same meaning, related to observed or measured data variables. schema.name, and schema:title likewise have equivalent meaning in other schemas.
  Thus, when developing a crosswalk it is necessary to check how each property is defined in each schema, and how it is actually used in the implemented examples. For example, schema.isBasedOn (a resource from which this work is derived or from which it is a modification or adaptation) can be mapped from datacite:IsOriginalFrom, datacite:isSourceOf, datacite:isDerivedFrom, datacite:isVersionOf.
- It is also inevitable that many terms from one schema are mapped to one term in Schema.org, due to Schema.org being a general schema and the simplicity is one of its design rules. For example: the granular relations from RIF-CS:(collection/relatedInfo/isVersionOf, collection/relatedInfo/isEnrichedBy, collection/relatedInfo/isDerivedFrom, collection/relatedInfo/hasValueAddedBy) and datacite:(isOriginal FromOf, isSourceOf, isDerivedFrom) can all be mapped to schema:isBasedOn (A resource from which this work is derived or from which it is a modification or adaptation).
- Rich granular information may be lost where 'many to one' mapping occurs. Whether this loss of information is significant depends on the purpose of a mapping and how this granular information is utilised by a data discovery system. For example, if a use case is to make a dataset widely findable from the web, then adding more descriptive metadata is more important than having a detailed relation; if a use case is to track the history or provenance of a dataset, then this granular relation information is important to have. These two use cases can complement each other: a general repository can have descriptive metadata for discovery and include links so that when a user finds ta potentially relevant dataset, they can follow a link to metadata with more granular contextual information to assess the fitness of the dataset for intended purpose.

3.3 Visualization Tool for Facilitating Mapping

To make the crosswalks more useful for analysis, and for those who are going to do a crosswalk for their own schema, the World Data System—International Technology Office has developed a tool to visualise the above 14 crosswalks (and one from CodeMeta vocabulary to Schema.org)[8]. The tool provides a user-friendly display of the collected crosswalks. By utilising the visualisation tool, crosswalk developers across domains can reference existing mappings, repeating the same types of matches between the Schema.org terms and similar elements found in different metadata schemas, regardless of whether the metadata format is standard or bespoke. The visualization tool is intended as a prototype service for the research data management community, in support of metadata managers who are investigating options for including schema.org markup into existing well formed metadata. The visualizations include various tables, a Sankey diagram, and a Gap Analysis, to support different views for crosswalk inspection. For example: Figure 2 can help to check, given a property from Schema.org, what is its corresponding element in other schemas; and Figure 3 shows these mappings in a 'Filter Table', where a parent type is also shown for properties from Schema.org.

[8] Visualisation of crosswalks: https://rd-alliance.github.io/Research-Metadata-Schemas-WG

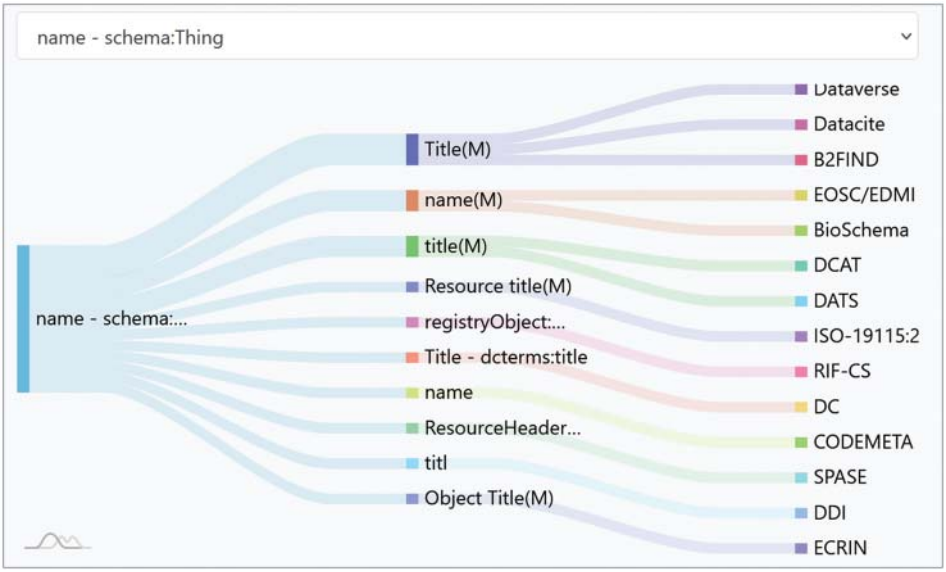


Figure 2. This filter sankey diagram allows a user to choose a schema.org property and see which crosswalked term is connected to which metadata standard. From left to right the labels go schema.org properties, crosswalked terms, then metadata standards.

Filter Table Data

title

Standard	Term	Schema.org crosswalk	Parent Schema
ISO-19115:2003	Resource title(M)	name	schema:Thing
Dataverse	Title(M)	name	schema:Thing
DCAT	title(M)	name(R)	schema:Thing
DATS	title(M)	name	schema:Thing
Datacite	Title(M)	name	schema:Thing
RIF-CS	registryObject:collection:name (Title as displayed in RDA)	name	schema:Thing
DC	Title - dcterms:title	name	schema:Thing
B2FIND	Title(M)	name	schema:Thing

Figure 3. This table is a free text search over both metadata terms and schema.org properties. Wildcard searches are not supported but partial searches are. For example, a search for "publish*" will not return any records, but the search for "publish" will return "datePublished", "publisher", and "Dataset Publisher."

4. DISCUSSION AND CONCLUSION

In summary, through the analysis of the 14 crosswalks, we find most descriptive metadata are mostly interoperable among the schemas and can be mapped to corresponding Schema.org properties. The most inconsistent mapping is the 'Rights' metadata, which requires clearer and consistent definition among the schemas of the terms Rights, License, Copyright Holders, and Data Use Agreement or Conditions, to name a few. The largest gap exists in the Structural metadata elements: first, there is a lack of consistency among the source metadata schemas themselves; and second, there are no rich relation terms in Schema.org. As Structural metadata is important in the linked-data world, the data community needs to agree what Structural metadata from disciplinary schemas could be generalised and applied to all types of data. There also exists a gap in controlled vocabularies to specify various property values, for example, observational variables [17] and a subject classification vocabulary (e.g. Library of Congress Subject Headings) for populating Keyword or Subject elements to describe a dataset.

The gaps are due to the Schema.org design principle that starts simple and increases complexity when community need arises [11]. This challenge is complicated by the fact that relatively simple, domain independent vocabularies satisfy the most common web data search needs, but the research community tends to use more granular and rigorous schema and controlled vocabularies in describing and cataloging research dataset. Lagoze [15] argued that attempting to intermix a single descriptive vocabulary for coarse granularity queries with the complex semantics needed to enable 'drill-down' into more granular queries, leads to metadata sets that are not ideally suited for either purpose; Lagoze advocated for establishing frameworks for the creation of more complex descriptions that can coexist with similar ones as separate packages.

Like any other schemas or vocabularies, Schema.org is evolving. To address the above gaps, the terms schema:DefinedTerm and schema:inDefinedTermSet were introduced as pending changes in Schema.org V12.0, and schema:hasDefinedTerm in Version 13.0[8] to enable the markup of external property names and pre-defined property values from discipline specific vocabularies. This approach balances the simplicity for a general schema and complexity of disciplinary schemas by following some principles that guide the development of metadata schema, especially the modularity principle and the extensibility principle [9]. The recent trend, as observed in the survey and from the development of application profiles by domains (e.g. DCAT-Application Profile and Bioschema profiles,) also follows Duval's metadata development principles.

In summary, we present the analysis of crosswalks to Schema.org from a cross section of domain implemented metadata schema. The analysis is limited by the survey and the conceptual mapping that focuses on the meaning of the elements or properties when mapping between two schemas. This analysis could be enhanced to include the analysis of implemented marked up metadata across repositories to get a more comprehensive picture of the interoperability of published structured metadata on the Web.

[8] https://schema.org/docs/releases.html

ACKNOWLEDGEMENT

This work was developed as part of the Research Data Alliance (RDA) Working Group entitled 'Research Metadata Schemas', and we acknowledge the support provided by the RDA community and structures. We would like to thank members of the group for their support and their thoughtful discussions through plenary sessions and regular monthly calls.

Special thanks go to:

- Joel Benn (ARDC, Australia), Kerrin Borschewski (GESIS, Germany), Steve Canham and Christian Ohmann (University of Dusseldorf, Germany), Baptiste Cecconi (Observatoire de Paris, PSL Research University, France), Douglas Fils (Ocean Leadership, US), Julian Gautier (Harvard University, US), Josef Hardi and John Graybeal (Stanford University, US), Leopold Talirz (EPFL, Switzerland), Chris Hunder (GigaScience Journal), Andrea Perego (European Parliament), Stephen M. Richard (US Geoscience Information Network), Philippe Rocca-Serra and Susanna-Assunta Sansone (Oxford University, UK), Adam Shepherd (WHOI, USA), Matt Styles (Nottingham University, UK), Heinrich Widmann (DKRZ, German), Bruce Wilson (ORNL, USA) and a few anonymous survey participants for contributing to the survey on "Current practices in using schemas to describe research datasets" and/or crosswalks;
- Karen Payne, Seiya Terada and Chantelle Verhey (World Data System—International Technology Office, Canada) for developing a suite of tools for visualising the collected crosswalks.
- ARDC intern Penelope Hagan for initial alignment of 13 crosswalks.

AUTHOR CONTRIBUTION STATEMENT

Mingfang Wu (mingfang.wu@ardc.edu.au) conceptualised and implemented the crosswalk analysis and wrote the original draft, all authors contributed to further conceptualisation and the writing and review of the paper. Stephen Richard (smrTucson@gmail.com) provided the mapping for ISO 19115-1 and contributed text editing and review.

REFERENCES

[1] Ball, A., Greenberg, Jane., Jeffery, K., et al.: RDA Metadata Standards Directory Working Group: Final Report. (2016). Retrieved on 15 Sept. 2021 from: https://www.rd-alliance.org/system/files/MSDWG-Final-Report.pdf

[2] Baca, M. (Editor).: Introduction to Metadata: Third Edition. ISBN:978-1-60606-479-5. (2016). Available online: http://www.getty.edu/research/publications/electronic_publications/intrometadata3'. Retrieved on Oct. 31, 2021.

[3] Benjelloun, O., Chen, S., Noy, N.: Google Dataset Search by the Numbers. arXiv:2006.06894. (2020). https://arxiv.org/pdf/2006.06894.pdf

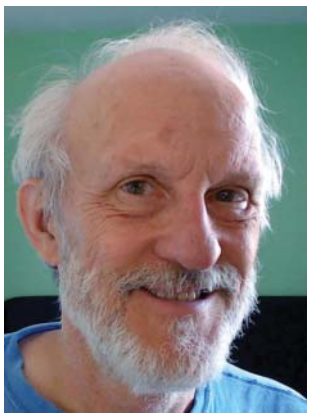
[4] Brickley, D., Murgess, M., Noy, N.: Google Dataset Search: Building a search engine for datasets in an open Web ecosystem. The World Wide Web Conference, San Francisco, CA, USA, May 2019. Pp.1365–1375 (2019)

[5] Canham, S., Ohmann, C.: ECRIN Clinical Research Metadata Schema Version 2 (April 2018) (2.0). Zenodo. (2018). https://doi.org/10.5281/zenodo.1312539
[6] Chan, L.M., Zeng, M.L.: Metadata Interoperability and Standardization – A Study of Methodology I: Achieving Interoperability at the Schema Level. In D-Lib Magazine, Vol.12(6). (2006). ISSN 1082-9873. Available at: https://dlib.org/dlib/june06/chan/06chan.html
[7] Cox, S.J.D., Gonzalez-Beltran, A.N., Magagna, B., Marinescu, M.-C.: Ten simple rules for making a vocabulary FAIR. PLoS Comput Biol 17(6), e1009041 (2021). https://doi.org/10.1371/journal.pcbi.1009041
[8] DataCite Metadata Working Group (DataCite).: DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. 10.5438/0014 (2017)
[9] Duval, E., Hodgins, W., Sutton, S., Weibel, S.L.: Metadata principles and practicalities. D-Lib Mag 8(4), 16 (2002). Available from: http://www.dlib.org/dlib/april02/weibel/04weibel.html.
[10] Fenner, M.: Using Schema.org for DOI registration. DataCite Blog (Jan. 9, 2017). (2017). Available from: https://doi.org/10.5438/0000-00cc
[11] Guha, V., Brickley, D., Macbeth, S.: "Schema.org: Evolution of structured data on the Web: Big data makes common schemas even more necessary". ACMQuery, November 2015, https://doi.org/10.1145/2857274.2857276
[12] Gray, A.J.G., Goble, C.A., Jimenez, R.: Bioschemas: From Potato Salad to Protein Annotation. In International Semantic Web Conference (Posters, Demos at Industry Tracks) (2017)
[13] Habermann, T.: Mapping ISO 19115-1 geographic metadata standards to CodeMeta. PeerJ Computer Science 5, e174 (2019). https://doi.org/10.7717/peerj-cs.174
[14] Jones, M.B., et al.: Science-on-Schema.org v1.2.0 (Version 1.2.0). Zenodo. (2021). https://doi.org/10.5281/zenodo.4477164
[15] Lagoze, C.: Keeping Dublin Core simple: Cross-domain discovery or resource description? D-Lib Magazine 7(1) (2001)
[16] Lebo, T., et al.: PROV-O: The PROV Ontology. (W3C Recommendation). World Wide Web Consortium. (2013). http://www.w3.org/TR/2013/REC-prov-o-20130430/
[17] Magagna, B., et al.: The i-adopt interoperability framework for fairer data descriptions of biodiversity. (2021). DOI:10.5194/egusphere-egu21-13155
[18] Nilsson, M., Baker, T., Johnston P.: Interoperability levels for Dublin Core Metadata. (2009). Available at: https://www.dublincore.org/specifications/dublin-core/interoperability-levels/. Accessed on May 1, 2022
[19] NISO (National Information Standards Organization).: Understanding metadata. Bethesda, MD: NISO Press. (2004). Available from: http://www.niso.org/standards/resources/UnderstandingMetadata.pdf.
[20] Noy, N.: Making it easier to discover datasets. Published Sept 5. 2018. Google Blog. (2018). Available from: https://www.blog.google/products/search/making-it-easier-discover-datasets/
[21] Noy, N.: An analysis of online datasets using dataset search. Google AI Blog. (2020). Available at: https://ai.googleblog.com/2020/08/an-analysis-of-online-datasets-using.html. Accessed on May 1, 2022
[22] Sansone, SA., Gonzalez-Beltran, A., Rocca-Serra, P., et al.: DATS, the data tag suite to enable discoverability of datasets. Sci Data 4, 170059 (2017). https://doi.org/10.1038/sdata.2017.59
[23] Southwick, S.B., Lampert, C.K., Southwick, R.: Preparing Controlled Vocabularies for LInked Data: Benefits and Challenges. Journal of library metadata, 2015-10-02, Vol.15 (3–4), p.177–190 (2015)
[24] Tennant, R.: Different paths to interoperability. Library Journal 126(3), 118–119 (2001)
[25] Willis, C., Greenberg, J., White, H.C.: Analysis and Synthesis of Metadata Goals for Scientific Data. In Journal of American Society for Information Science and Technology 63(8), 1505–1520 (2012). DOI: 10.1002/asi.22683.

[26] Wilkinson, M., Dumontier, M., Aalbersberg, I., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
[27] Wu, M., et al.: Guidelines for publishing structured metadata on the Web. Research Data Alliance. (2021a). DOI: 10.15497/RDA00066
[28] Wu, M., et al.: A Collection of Crosswalks from Fifteen Research Data Schemas to Schema.org. Research Data Alliance. (2021b). https://doi.org/10.15497/RDA00069
[29] Wu, M., et al.: Automated metadata annotation: What is and is not possible with machine learning. Data Intelligence 5(1), 122–138 (2023). doi: 10.1162/dint_a_00162

AUTHOR BIOGRAPHY

Dr. **Mingfang Wu** is senior research data specialist at the Australian Research Data Commons (ARDC). She leads ARDC data discovery projects for making data discoverable by both machine and human users. She has conducted research in the areas of information retrieval; user search behaviour and search context through search log analysis, survey and interview; interfaces supporting exploratory search; and natural language processing.
ORCID: 0000-0003-1206-3431



Dr. **Stephen Richard** is a geoinformatics consultant based in Tucson, Arizona. His background is in geologic mapping and geoscience data management during 24 years at the Arizona Geological Survey. He has participated in geoscience vocabulary development for state and federal geological surveys in the US and the IUGS CGI Geoscience Terminology Working Group. Richard was the editor for the ISO19115-3 XML implementation of ISO 19115 metadata, and has participated in technical development of metadata catalogs for the US National Geothermal Data System and EarthCube DataDiscovery Studio, using ISO 19115 metadata. Recent work has focused on development of schema.org metadata profiles for geoscience datasets for the EarthCube GeoCODES resource and data catalogs, and development of metadata schema for cross-domain sample descriptions and astromaterials analytical data.
ORCID: 0000-0001-6041-5302

Chantelle Verhey is a Research Associate for the World Data System-International Technology Office hosted at Ocean Networks Canada. She has a Masters of Science in Environmental Management from the University of Reading in the UK, and was dedicated to researching Forest fire trends in the Canadian Boreal Forest. After her research was completed, Chantelle moved on to work at the University of Waterloo as a Data Manager for the Polar Data Catalogue. Now, she is combining her research and work experience to enhance data interoperability within the polar scientific community through the use of semantic technologies.
ORCID: 0000-0002-0047-7875

Dr. **Leyla Jael Castro** is currently working as team leader for the Semantic Retrieval research team, part of the Knowledge Management Group, at ZBMED Information Centre for life sciences, focusing on topics such as literature-based information retrieval, recommendation systems, and ontology-based search and categorization. She participates in community efforts such as Bioschemas, and networks such as RDA and ELIXIR.

Dr. **Baptiste Cecconi** is an astronomer working at Observatoire de Paris in Meudon, France. His background is radio astronomy, solar and planetary sciences. He is an active member of his research field's open science alliances: the International Virtual Observatory Alliance (IVOA), the International Planetary Data Alliance (IPDA) and the International Heliophysics Data Environment Alliance (IHDEA). His recent data-related projects are aiming at building interfaces between radio astronomy and neighbouring science fields, focussing on semantic as well as operational interoperability.
ORCID: 0000-0001-7915-5571

Dr. **Nick Juty** is a Senior Research Technical Manager. He is an experienced senior scientist with recent focus on standards adoption across scientific domains. He has played a leading role in delivering an international and cross-disciplinary identification system for scientific data (http://identifiers.org).
ORCID: 0000-0002-2036-8350