Final Report
Research Data Alliance Research Metadata Schemas Working Group

Guidelines for publishing structured metadata on the web

15 June 2021, V3.0

Mingfang Wu, Nick Juty, RDA Research Metadata Schemas WG, Julia Collins, Ruth Duerr, Chantel Ridsdale, Adam Shepherd, Chantelle Verhey, Leyla Jael Castro

Table of Contents

Executive Summary

The FAIR principles refer frequently to metadata as a key enabler in discoverability, but also having a major role in accessibility and reusability. Publishing structured metadata on the web can provide a simple and efficient means to increase the FAIRness of research resources: it exposes metadata contained in web pages through a formal mechanism, allowing systematic collection and processing by web-based crawlers. Efforts to adopt structured metadata within and across domains would benefit greatly from a set of recommendations that would help ensure consistent implementation leading to enhanced discoverability and accessibility of data. Based on community consultation and subsequent work, this guidelines provides nine recommendations to support the process of publishing structured metadata on the web, namely:

- Recommendation 1: Clarify the purpose(s) of your markup
- Recommendation 2: Identify what resource are to be marked up with structured data
- Recommendation 3: Adopt or develop a crosswalk from a repository schema to markup vocabulary
- Recommendation 4: Incorporate external vocabulary if it helps to improve data discoverability and interoperability
- Recommendation 5: Implement markup syntax consistently by following community practices
- Recommendation 6: Be friendly to web crawlers
- Recommendation 7: Make the best use of available tools for mapping, generating and validating structured data
- Recommendation 8: Document and share every step
- Recommendation 9: Find and join a community, and follow their established practices

Terminology

Crosswalks: Metadata crosswalks translate elements (types and properties) from one schema to those of another. Crosswalks facilitate interoperability between different metadata schemas and serve as a base for metadata harvesting and record exchange[1].

A crosswalk acts as a “mapping of the elements, semantics, and syntax from one metadata scheme to those of another. A crosswalk allows metadata created by one community to be used by another group that employs a different metadata standard” (National Information Standards Organization, 2004, p. 11). Practically, this means that properties in different schema may have different ‘names’, but be conceptually identical. E.g., dcat:Catalog and schema:DataCatalog.

Data repository and data catalogue: Will be used interchangeably in this paper to refer to those cataloguing and publishing metadata. A data repository is a web-enabled or accessible resource where data is hosted. Frequently, these repositories are themselves indexed by other resources, providing a ‘data catalogue’. Data catalogues often do not host the data themselves, but store crucial metadata from referenced repositories, allowing one to identify potentially useful individual repositories from a wider pool. In this document, we see no reason to distinguish between these resource types.

Identifier/Persistent Identifier: An identifier is a label which gives a unique identity to an entity: a person, place, or thing. A persistent identifier reliably points to a digital entity[2].

Type: A type represents an entity or thing when it is conceptualised digitally. This type corresponds to a thing observed in the real world, e.g., type chair or type person.

Property: A property is an attribute or relation that is associated with an entity when it is conceptualised digitally. This attribute can furthermore be assigned a quantitative or qualitative value, which provides a name/value pair. or instance “family_name” as name and “Murdoch” as value

Property Name: the name (or key) of the property.

Property Value: the value of the property.

Instance: an example or single occurrence of something

Metadata Publication/Publishing metadata: In this manuscript, this refers to the publication of metadata embedded in landing web pages, i.e., publication of metadata over

[1]University of Texas Libraries: Crosswalk
[2] https://support.orcid.org/hc/en-us/articles/360006971013-What-are-persistent-identifiers-PIDs-

the web. An alternative expression would be “publishing structured data markup (on the web)”.

Semantic Artefacts: (aka semantic resources, semantic structures or more generally knowledge organisation systems). Semantic artefacts organise knowledge so it becomes interpretable and actionable not only by humans but also by machines. They commonly include concepts together with definitions, equivalences and synonyms, aiming at removing (or at least reducing) ambiguity and establishing explicit semantic relationships such as hierarchical and associative relationships, and presenting both relationships and properties of concepts as part of the knowledge model (Zeng, 2008).

Structured data: In this paper, structured data means structured metadata, that is metadata formatted and presented in a manner to facilitate machine processing, supported by a semantic schema or vocabulary.

Markups: sometimes also called snippets. These represent properties (see ‘property’ above) and are implemented on the web in various formats: RDFa, microdata, JSON-LD, where JSON-LD is the currently preferred format.

Controlled Vocabulary: A controlled vocabulary corresponds to a vocabulary restricted to a set of predefined options, commonly agreed by a community or broadly adopted in a domain.

Schema: Here schema refers to data or knowledge schemata. A data schema corresponds to data structure and organisation described in some formal language, e.g., via types and properties such as “Person” with a “family name” and a “first_name”.

1. Introduction

Over the past decade, we have seen an increasing number of public and domain specific data repositories as data sharing is becoming a common scientific practice. Two of the reasons behind the increase of data sharing and data repositories are improving research reproducibility (Vasilevsky, 2017; Merz, 2020) as well as aligning to Open Science initiatives (Munafò, 2016). For example, re3data.org, the Registry of Research Data Repositories, had 23 repositories when it went online in 2012; the number quickly increased to over 1,200 data repositories from across the globe in three years (Pampel and Vierkant 2015), and, by February 2020, the registry had more than 2450 repositories[3]. While data sharing via data repositories is highly welcomed by the scientific community, it becomes ever more challenging for researchers and the public to discover relevant data, especially when required data for a research comes from several repositories. In addition, data aggregators are required to deal with harvesting metadata from a number of sources using a variety of metadata schemas.

There are different ways to discover data on the web, web search tools being one of the approaches favored by researchers (Gregory, et al. 2019). The Web provides a global platform for discovering data that can still be further exploited. One of the current uses of the Web as a data discovery platform relies on web-based data repositories publishing metadata as part of websites landing pages. Such metadata can be used by search engines to improve data discovery and accessibility for human users. However, not all metadata and metadata formats will be easily understood by search engines and, in general, by machines. For machines to correctly interpret and process the meaning of metadata (and data behind it), we need to mark up metadata with a common vocabulary as well as in a machine-processable encoding, i.e., the markup needs to be semantically structured. This structured markup makes possible both semantic and syntactic interoperability on the web (at least at a basic level as markup metadata commonly targets broad use cases as opposed to domain specific vocabularies with greater expressivity and high complexity).

In the past few years, research data repositories have started adopting structured metadata in their landing pages. It is expected that publishing structured metadata over the web will enhance the FAIRness of metadata, particularly the “Findability” aspect in the FAIR (meta)data principles (Wilkinson et al., 2016). Publishing structured metadata makes data more discoverable by web search tools. It also enables rich display of a search result, making it easier for data seekers to judge the relevance of the presented results in terms of the data behind them – an important step of the information searching process with online web search tools (Turpin et al., 2009). Figure 1 shows a search result corresponding to the query “Satellite ASTER Geoscience Map of Australia” from a general web search tool (Figure 1a) and a dataset search tool (Figure 1b). Compared with the general web search engine, the search result presented from the Google Dataset Search[4] tool clearly shows properties associated with data, enabling users to identify repositories that publish metadata about the same (or similar) datasets.

[3] https://blog.datacite.org/german-research-foundation-to-fund-new-services-of-re3data/
[4] https://datasetsearch.research.google.com/

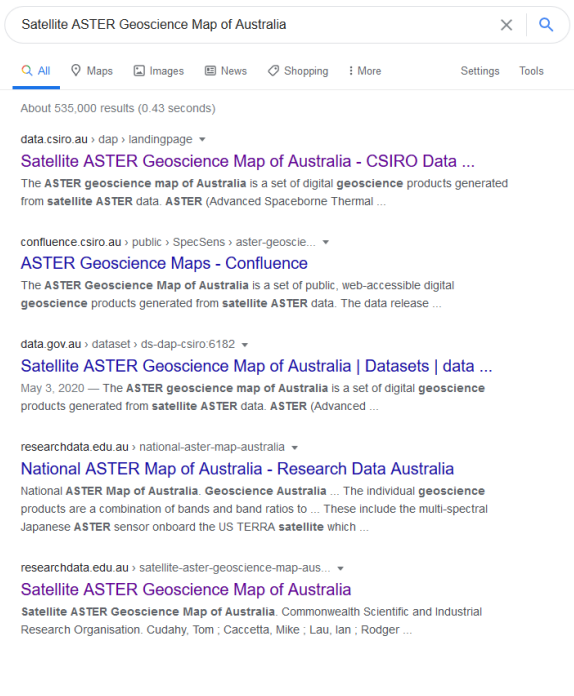


Figure 1a: Search result from google web search engine

Figure 1b: Search result from google dataset search tool

As more data repositories make their data more discoverable by using common vocabularies or schemas, metadata interoperability across repositories will also be enhanced. The research data community can take advantage of such enhanced metadata interoperability; for instance, researchers can explore new methods for metadata syndication and data discovery via the web architecture based on a common vocabulary. If implemented properly, structured data can lead to linked metadata and thus linked (underlining) data, which will enable smart web data applications to perform to their potential. It will also provide opportunities for the research data community to develop innovative search tools such as the initiative of Japan's open data search engines (Keto et al, 2020) and open data discovery infrastructures based on open knowledge maps (Kraker, et al., 2021), enable applications such as aggregated search across resources of a specific domain or related domains relevant to a research need, applications building research knowledge graphs supporting a spectrum of data search needs from free text search, JSON API to SPARQL queries.

In the past years, Schema.org has become a vocabulary commonly used by websites to describe their content and expose the corresponding structured metadata so search engines can better interpret the meaning and data searchers can benefit from more accurate results. Schema.org was originally intended for use in e-commerce applications, largely focusing on domains such as news, movies, products, medical, music etc., but nowadays is also used by libraries around the world to publish bibliography information supporting Linked Data (Godby et al. 2015). Some data repositories, for example NASA, NOAA and Harvard's Dataverse repository, have already

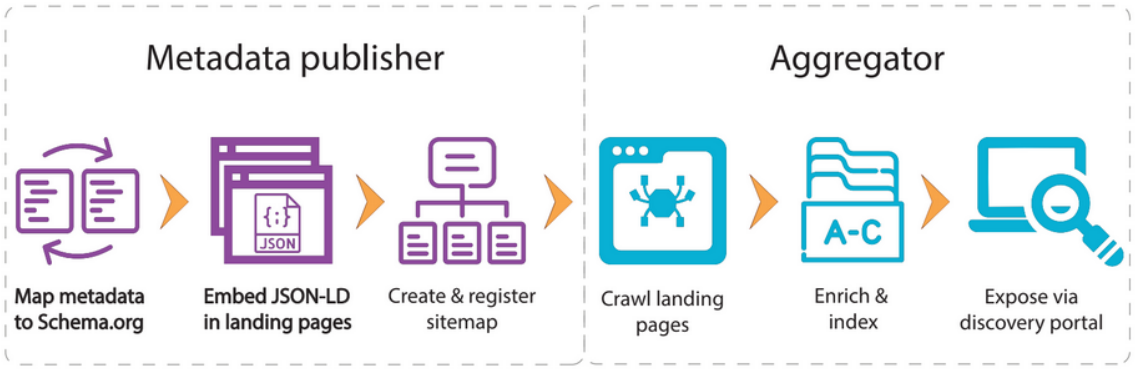
adopted this approach for making their dataset more discoverable on the Web (Noy, 2018), while some other repositories are about to onboard the path. The Research Data Alliance (RDA) Research Metadata Schema Working Group was formed with the purpose of data repositories to exchange experience and lessons learned from publishing structured metadata and to have consistent implementation of the publishing process across repositories. This guideline, as an output of the working group, is to serve the purpose.

2. Process to publish structured metadata

In these recommendations, we assume that repositories have already followed the FAIR (meta)data principles (Wilkinson et al. 2016) and chosen an appropriate data model and schema(s) for describing their data to meet the needs of data management, curation and discovery defined by their community. The chosen schema or schemas may be (i) specific to their domain/disciplines and thus provide a richer and more targeted array of domain-relevant properties, for example, ISO 19115[5] for geographic data and ECRIN metadata schemas for clinical research data (Canham, 2020); (ii) generic schemas, such as Dublin Core, PROV-O, Schema.org, recording those properties that are common across multiple disciplines; or (iii) a combination of domain specific and domain agnostic schemas. Either way, this is the primary step and lays the foundation for publishing structured metadata to the Web.

With that assumption in mind, Figure 2 shows a general process for publishing and consuming structured data. Metadata publishers usually undertake the following four steps (Figure 2a):

1. Develop a crosswalk from a repository's source metadata to Schema.org.
2. Generate markup metadata with Schema.org vocabulary in a commonly adopted format, usually Resource Description Framework in attributes (RDFa), microdata and JavaScript Object Notation for Linked Data (JSON-LD) or Microdata, and embed the markup into the metadata of the landing page.
3. Include URLs of the landing pages into a sitemap, register the sitemap with potential downstream consumers such as web search engine operators, metadata aggregators or application developers.



[5] ISO 19115-1:2014 Geographic information - Metadata: https://www.iso.org/standard/53798.html

Figure 2: a) Left, the publishing process for the publisher, b) Right, the process for aggregators to make structured data discoverable.

Once data repositories provide structured data, a data aggregator will go through the following three steps to consume the structured data (Figure 2b):

1. Send a crawl to fetch each URL from the sitemap.
2. Parse, index and enrich information from the landing page and expose the enriched set as structured data.
3. Make the index (possibly combined with other indexes available to the aggregator) searchable.

During this process, metadata publishers, e.g., data aggregators, can face challenges such as:

- The lack of consistent implementation of structured metadata across data repositories, and guidelines for those who would like to pursue this path. Inconsistent implementation of structured metadata at either the semantic or syntactic level prevents the interoperability and reusability of structured data.
- The Schema.org vocabularies are intentionally minimalistic, for encouraging fast, easy and wide adoption. This strategy works well for data such as movies and cooking recipes. However, research data usually have richer properties and provenance relations with other research objects, and in many cases, there requires the incorporation of external vocabularies and extensions beyond the minimum recommendation, and thus the need of guidelines for achieving this.

3. Data model

To enable repositories to publish and exchange metadata records over the Web, the data model has to be simple to understand and easy to implement. In fact, the Resource Description Framework (RDF) has a simple and abstract data model for representing metadata about web resources and other information[6]. The RDF data model makes statements about a resource, with a statement being expressed as a triple in the form *subject-predicate-object* as shown in Figure 3, where *Subject* and *Object* are web resources and *predicate* specifies the relationship between the two resources. *Predicates* can also be referred to as *properties*. As more resources are described in this way, they can be integrated and linked, forming a web of data, enabling the construction of knowledge graphs and semantic queries.

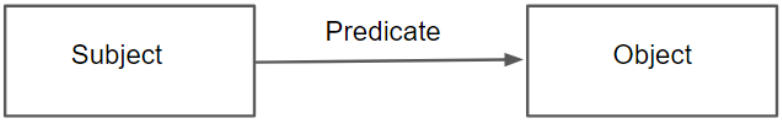


Figure 3: An RDF triple statement

[6] W3C Resource Description Framework (RDF): Concepts and Abstract Data Model https://www.w3.org/2002/07/29-rdfcadm-tbl.html

Several standards have been developed to support the RDF data model, for example, the Web Ontology Language (OWL)[7], Simple Knowledge Organisation System (SKOS)[8], and RDF Schema (RDFs)[9]. However, RDF standards and their serialisation do not necessarily benefit from large scale uptake on web pages, due largely to its rigorous rules and the lack of familiarity or expertise in those people (webmasters) who publish web resources (Guha et al. 2015).

The Schema.org data model, on the other hand, is specifically meant for describing resources that are published on the Web. The data model retains some aspects of RDF but simplifies the vocabulary and its rules, targeting the description of web resources[10] and offering a lightweight semantic option for web data providers. As shown in Figure 4, in the Schema.org data model:

- Each resource or a thing, to be described in a metadata landing page, has a type, for example, a resource can be a type of 'CreativeWork', 'Dataset', 'Software', 'Organisation' or 'Person'. Types are arranged in a multiple inheritance hierarchy where each type may itself be a subclass of multiple types, for example, a dataset is a subclass of 'CreativeWork', which is a subclass of the 'Thing' - the most generic type of item.
- Each type has a set of properties (or attributes), which collectively define a type. For example, a type 'Dataset" has properties such as 'title', 'description', 'subject', 'identifier', 'creator' and so on.
- A property may have simple literal values or instances of other resources with their own types and properties. For example, a resource type 'dataset' has a property 'title' whose expected value is in literal 'text', the 'dataset' has a property 'creator' whose expected values can be a resource instance of the type 'Person' or 'Organisation'.

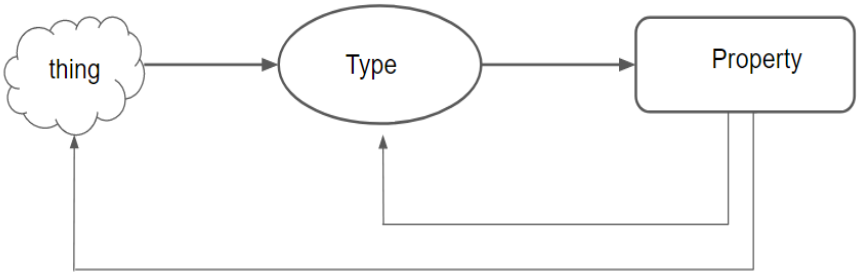


Figure 4: Schema.org data model

Schema.org provides a vocabulary to name the 'type' and the 'property', specifying unambiguously what we are talking about. When we describe an item (e.g., a specific dataset) in the world by assigning the item a type and associated property values, we then create an instance of the type. The Schema.org data model focuses purely on data 'types' and 'properties', and does not extend to specifying whether a property is mandatory nor whether it can be repeated several

[7] https://www.w3.org/OWL/
[8] https://www.w3.org/TR/swbp-skos-core-spec/
[9] https://www.w3.org/TR/rdf-schema/
[10] https://schema.org/docs/datamodel.html

times for the instantiation of a 'type', as in some other schemas. By default, all properties are optional and accept multiple elements. Due to this simplicity, entities and properties as described by other schemes (e.g., ISO19115, DCAT, Dublin Core) can be easily represented or mapped to this model.

The Schema.org data model can be serialised in RDFa, microdata and JSON-LD. These serialisations make it easier to embed the type and properties of a resource item within a HTML page, thus enabling machines to understand the semantic context and building knowledge about the item as described on the resource's HTML page. Due to its simplicity, Schema.org has been widely adopted on the web to expose structured data[11]. If RDFa, microdata and JSON-LD are implemented consistently and compatibly at the syntax level, they can be easily mapped to RDF, retaining the ability to construct web knowledge graphs based on the types and properties, and connections, i.e., relations, across described resources.

Currently, Schema.org vocabulary has about 778 types and 1383 properties. The W3C Schema.org Community Group[12] governs the development and maintenance of the vocabulary. New types and properties can be added if there is community need and support, for example, the new type 'LearningResource' was added as a subtype of 'CreativeWork' in 2020 July release (9.0)[13][14]. As another example, the Bioschemas[15] has done the same for the life science, having successfully incorporated many biomedical terms into the schema.org vocabulary. There are also communities who support the consistent serialistation of the data model. For example, the Schema.org Cluster of the Earth Science Information Partners[16] is working to develop best practices and to provide education and outreach to the Earth science community for web accessible structured data, in order to improve scientific data discovery capabilities.

4. Recommendations

Publishing structured metadata to increase metadata interoperability requires consistent implementation across data repositories to realise its full potential. To that end, the RDA Research Metadata Working Group conducted a community consultation[17], asking participants who were planning to publish structured metadata what they would like to know beforehand (e.g., from others' experience), and to those participants who had already implemented structured metadata, what lessons they could share, particularly pitfalls to avoid. Additional input was also solicited from communities and projects that were active in this area, including Bioschemas[18], Science-on-

[11] What a long, strange trip it's been: https://www.slideshare.net/rvguha/sem-tech2014c/11-07_Rise_of_the_consumers
[12] https://www.w3.org/community/schemaorg/
[13] Schema.org Releases: https://schema.org/docs/releases.html
[14] Learning Resource Metadata is go for Schema: https://blogs.pjjk.net/phil/lrmi-in-schema/
[15] https://bioschemas.org
[16] The ESIP Schema.org Cluster: https://wiki.esipfed.org/Schema.org_Cluster
[17] Requirements/Discussions as captured from the RDA P15
[18] https://bioschemas.org/

Schema.org (Jones, et al, 2021) and various library catalogues on the web. We have coalesced these learnings to derive the following nine recommendations for data repositories, or for anyone who intends to implement structured data in their metadata landing pages, to meet the above challenges as discussed in Section 2.

Recommendation 1: Clarify the purpose(s) of your markup (or why you want to markup your data)

Before publishing structured data, the first question one has to ask is: what are the purposes of adding structured data to resource landing pages? The answer to this question may impact the scope of the task and decisions made at a later stage of the process, for example, which resource objects from a repository should be in scope, which schema, vocabulary and syntactic implementation are appropriate. In general, there are two broad use cases for publishing structured data:

1. For data discovery
   The initial motivation for having structured data came from web search engine operators, whose purpose is to improve data search and result presentation over the web. Repositories need to check what search features are provided by the targeted web search tools, as that can impact on the coverage of structured data. The most common search future is the keyword search. Keyword search indicates topical relevance between searched keyword and searched data, this topic information is usually captured in descriptive metadata such as title, description and keywords. On top of the keyword search, some web data search tools involve advanced search features such as facet filter or facet search along one or more data attributes, to help users to narrow down or broaden up a search, to assist assessment on the relevance or usefulness of a candidate datasets. Other novel advanced data discovery features include the utilisation of data linkage to construct knowledge graphs, for instance, combining Wikidata and Bioschemas data[19]. Novel strategies such as these aim to achieve more precise answers to a search query. And the more discovery features are offered, the more coverage of metadata is required.

2. For exchanging metadata with other repositories
   Embedding structured data in landing pages offers a new way for metadata aggregators to harvest metadata. Currently, if a metadata aggregator harvests metadata from multiple data repositories, or a data repository exports detailed metadata to multiple downstream repositories or catalogues, either the metadata aggregator or the data repository would have to implement and maintain several crosswalks. If both data repositories and aggregators are implementing structured data markup, they would save resources on maintaining crosswalks as they only need to have a crosswalk from their own schemas to/from the common markup vocabularies.

[19] The combination of Wikidata and Bioschemas data is an ongoing project, its current code can be found at https://github.com/elizusha/graph-loader

Aggregators have a similar purpose to the Web discovery applications, that is to make the aggregated metadata (thus data) more discoverable. Aggregators of a specific domain may accommodate and require more detailed metadata than generalist aggregators. So repositories involved in harvesting and exchanging metadata need to understand each other's requirement and potential usage of metadata.

In either use case (or both), **one needs to first identify the purpose and understand the requirement of down streaming metadata consumers, as which impacts on the scope of the project that sets up to publish structured data**. In addition, it is worth noting that the power of structured data lies in its connection to other resources or entities published to the web, for example, a dataset may be a subset or derivative from another dataset; or a dataset may be a secondary product, produced following some software processing, the result of a workflow, etc. Linking to other relevant resources is a good practice for data discovery, metadata exchange and data aggregation.

Recommendation 2: Identify what resources are to be marked up with structured data

More and more data repositories have metadata for not only datasets, but also other research resource objects such as software, models, instruments, samples, etc. These resources are essential for supporting open and reproducible research. Our analysis (Table 1) shows almost every research resource object has a corresponding class from Schema.org.

Table 1: Mapping dataset and related resources to Schema.org components

	Type of resources ("things")	Other standards/Schemas/Schema Class	Schema.org (type)
Primary entity	Catalogue	dcat:Catalog	schema:DataCatalog
	Dataset	dcat:Dataset	schema:Dataset which can contain: ● schema:Article ● schema:Movie ● schema:AudioObject ● schema:ImageObject Or any other schema:Thing
	Software	Codemeta (essentially schema:SoftwareSourceCode, schema:SoftwareApplicat	schema:SoftwareSourceCode schema:SoftwareApplication

		ion)	
	Data service	dcat:DataService	schema:WebAPI
	Publication (grey publication)	DublinCore[20] dcterms:BibliographicResource Bibliographic Ontology (BIBO)[21] bibo:Document bibo:Article bibo:AcademicArticle bibo:Manuscript Semanticscience Integrated Ontology (SIO)[22] sio:publication sio:article sio:peer_reviewed_ article	schema:Book schema:Article:ScholarlyArticle schema:Chapter schema:Poster, schema:Thesis, schema:Report
	Documentation/report	As in publication	schema:Report
	Training material	EDAM ontology[23] edam:TrainingMaterial	schema:Course (training) schema:Text, schema:Publication
	Course	bibo:Event	schema:Course, schema:Course:CourseInstance schema:Event:Hackathon,
Responsibility entity	Person	FOAF[24] foaf:Person	schema:Person
	Organisation	W3C recommendation: The Organization Ontology (ORG)[25] org:FormalOrganization	schema:Organization

[20] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[21] https://bibliontology.com/
[22] https://bioportal.bioontology.org/ontologies/SIO
[23] http://edamontology.org/
[24] http://xmlns.com/foaf/spec/
[25] https://www.w3.org/TR/vocab-org/

	Group	org:Organization	schema:Consortium
	Funding agency	org:Organization Funding, Research Administration and Project Ontology (FRAPO)[26] frapo:FundingAnency	schema:FundingAgency
Subject entity (concept, object, event, place)	Grant	frapo:Grant	schema:Grant
	award	As in the Grant	schema:Award
	Project	As in the Grant	schema:Project, schema:ResearchProject
	Event	bibo:Event	schema:Event
	Instrument	Work in progress within the RDA Persistent Identification of Instruments WG[27]	schema:Instrument

The primary goal of publishing data to the web (or any other platform) is for wider discoverability; however, discoverability is often simply a means for data to be found and reused. One has to determine the necessary properties of a resource, and their relationships to other resources, i.e. data provenance information that helps data consumers to judge the reusability and quality of that resource. The W3C Provenance Incubator Group[28] defines provenance of a resource as:

> 'a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility.' (Gil et al., 2010)

If we treat datasets as primary resources, according to the above definition and the provenance data model[29], then provenance information includes: where (e.g., location) and how (e.g.,

[26] FRAPO, the Funding, Research Administration and Projects Ontology: https://sparontologies.github.io/frapo/current/frapo.html

[27] RDA Persistent Identification of Instruments WG: https://www.rd-alliance.org/groups/persistent-identification-instruments-wg

[28] W3C Provenance Incubator Group Wiki: https://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki

[29] PROV-O: The PROV Ontology: https://www.w3.org/TR/prov-o/

software, instrument, model, sensor) data is captured or produced, as well as who (person or organisation) has been involved in its generation, and for which purpose (e.g., project or grant) it was produced. The question then becomes, where should this provenance, these properties and relationships reside? For simple data discovery, basic metadata is often enough to lead users to the repository website where more information can be found. For many repositories this may be sufficient, at least as a first step. However, to facilitate advanced data discovery, assessment and access services, and to connect related resources wherever they may reside, we highly recommend publishing detailed dataset metadata connecting all related resources into a web of (distributed) data, thereby increasing discovery and access paths to all data.

Another important step when defining what resources will be included as structured data is the persistent identification of those resources. An 'Identifier' is a very special property providing a unique identity that should be included for most, if not all, marked up resources (McMurry, Jut and Blomberg, et al., 2017). An identifier is used to name a resource or a thing uniquely (whether a digital resource or not), a persistent identifier (PID) is guaranteed to be managed and kept up to date over a defined time period. The PID issuing body should provide accompanying documentation to describe the level of persistence. Examples of persistent identifiers include Digital Object Identifiers (DOI), Persistent URL (PURL) based identifiers (eg. identifiers.org, OCLC PURL), and Universal Resource Name (URN) etc. PIDs can be used by both humans and machines to identify resources on the web, and, where applicable, resolve to that resource.

Only the custodial repository should assign a persistent identifier. In situations where metadata records from a repository are harvested and published by multiple aggregators, using the same persistent identifier for a resource enables users to easily identify it as the same resource. For example, in Figure 1b, each of the three blue bars under the title holds a metadata record of the same dataset. In this example, data.csiro.au is the original metadata provider for its own *Aster* dataset, i.e. data.csiro.au is the custodian for that dataset and the corresponding metadata. Such metadata is harvested by aggregators such as researchdata.edu.au and search.datacite.org which keep the dataset DOI assigned by data.csiro.au to refer to it, making it easier for machines to identify the same dataset as the one referred by the original provider and the aggregators.

Recommendation 3: Adopt or develop a crosswalk from your repository schema to Schema.org

In most cases, a crosswalk from a repository schema to Schema.org is required unless Schema.org is already the repository's metadata schema. A schema crosswalk maps the elements from one schema to the equivalent elements in another schema[30]. We recommend the following practice on the crosswalk:

[30] Schema crosswalk: https://en.wikipedia.org/wiki/Schema_crosswalk

First, look for existing crosswalks. If a schema (such as ISO19115:2003, DublinCore) has already been widely adopted by communities, it is likely that a crosswalk already exists. One should first discover and adopt an existing crosswalk, instead of attempting to reinvent the wheel; developing a crosswalk may involve extensive labour on concept mapping and may require community consultation. Some crosswalk collections can help to this end, for example the one created by the Space Physics Archive[33]. Furthermore, reusing an existing crosswalk would ensure that those repositories will align to the same terminologies, allowing better opportunity of integration across repositories and data held. This is beneficial to downstream application developers and users when they search for data across repositories via web data discovery applications. If there is no crosswalk that has exactly the same source schema and target schema as desired, it is still useful to reference existing crosswalks for how properties from two schemas are mapped, especially when one can find a crosswalk that has the same target schema to map to.

Second, make your crosswalk openly available as early as possible. Even if a crosswalk is still under development, it is beneficial to open up a draft crosswalk to the community for feedback, making the crosswalk more adaptable and adoptable. This working group has collected about 15 crosswalks[31]. The 15 source schemas represent general data models (e.g DCAT, DCAT-AP and DataCite) and domain specific ones such as Geographic Information (ISO19115:2003), Bioschemas[32], European Clinical Research Infrastructure Network (ECRIN) (Canham, 2020), and Space Physics Archive Search and Extract (SPASE)[33].

Third, map as many properties as needed for the identified purpose. *Recommendation 1* identifies two main use cases for publishing structured data. If the only purpose is exposing data on web applications (e.g. Google dataset search), a repository only needs to map a minimal set of properties; however, mapping a lot more properties beyond the minimal set is required if a repository needs to exchange metadata with other repositories/aggregators of a specific domain, or to support more sophisticated queries and assist relevance assessment. Even when an aggregator recommends only a minimal set, it does not mean the aggregator is restricted by that set of recommended properties. For example, the Google dataset search guide[34] recommends 20 properties, which do not include 'date updated', 'data provider' and 'data funder' fields often used by repositories or desired by funders; however, Google dataset search does parse and render some of these properties when present as shown in Figure 1b (labelled as 'Dataset updated', 'Dataset provided by', and 'Dataset funder by' respectively).

It is also important to map those properties that are often searched for by data seekers. For example, Kacprzak et al. (2019) found, dataset search queries often include temporal and spatial properties, as well as properties data format and file type. The data format and file type properties are in Schema.org but not in Google dataset search guidelines[36]. If a property is important for a

[31] RDA Research Metadata Schemas WG / Crosswalks: https://github.com/rd-alliance/Research-Metadata-Schemas-WG/tree/master/crosswalks
[32] Bioschemas: https://bioschemas.org/
[33] Space Physics Archive Search and Extract: https://spase-group.org/
[34] https://developers.google.com/search/docs/data-types/dataset

user to judge the relevance or utility of that dataset and this information is missing from a search result, the user may not refer to the source repository to explore further. On the Web, there may be unforeseen consumers who would harvest and parse as much structured data as is available on the Web, in order to provide advanced or domain specific data discovery tools. The more properties that are provided, and the more connectedness of data to other data provenance information, the more discovery paths can be created for data users.

Fourth, take implementations of past versions of source schema or description of legacy data into consideration when adopting or developing a crosswalk. Sometimes, there is a clear mapping at the conceptual level; however, there may exist discrepancies between the latest schema and datasets that were described by following earlier versions of schema and/or implementation guidelines. For example[35], for the latest version of schema: Registry Interchange Format – Collections and Services (RIF-CS V1.6.3), the property RIF-CS:location (type: url with property target=download) (describing the physical and/or electronic locations(s) of a registry object) can be conceptually mapped to Schema:DataDownload:distribution (the description of the location for download of the dataset and the file format for download). However, earlier version of RIF-CS didn't have the target type "download", thus past guidelines from the metadata aggregator Research Data Australia (RDA) instructed their content providers to use the property RIF-CS:location(type=url) to point to the source metadata landing page. Taking this historical development of schema into consideration, it is more appropriate to map the RIF-CS:location (type: url) from earlier versions to Schema:sameAs.

Recommendation 4: Incorporate external vocabularies as necessary

A research data repository may use controlled vocabularies or other semantic resources to:

- **Specify relationships between described resources**, for example, a dataset *is a subset* of another dataset, a dataset *is collected* through a instrument, and then *is cleaned and normalised* by software;
- **Provide the allowed range of a property value**, for example, Library Congress Subject Heading for indicating topics of a library resource, the BODC Parameter Usage Vocabulary (PUV)[36] for labelling scientific variables.

The purpose of using controlled vocabularies is to standardise information, so that there is a shared understanding of the concepts, facilitating interoperability between adopters of those vocabularies, and enabling resources or resources with the same property to be linked thereby improving data discovery.

However, generic schemas such as **Schema.org vocabularies don't enforce constraints or recommend controlled vocabularies for property values or rich relations between resource objects**. This is a deliberate decision as Schema.org is for data from all domains (e.g.,

[35] RIF-CS to Schema.org crosswalk: https://documentation.ardc.edu.au/display/DOC/RIF-CS+to+Schema.org+crosswalk
[36] https://www.bodc.ac.uk/resources/vocabularies/parameter_codes/

news, jobs, music, events, movies, among others), and fewer constraints make it more easily adoptable. However, a data repository can use Schema.org together with vocabularies from other standards or namespaces. The incorporation of external vocabularies into Schema.org may enrich data search interfaces, such as faceted or filter searches (Wu, et al, 2021), as well as to enable APIs such as aggregated search across repositories of a specific domain or related domains.

When repositories plan to include vocabularies and properties outside of Schema.org, it is recommended they **use linked open vocabularies and dereferencable property names** as much as possible. Linked Open Vocabularies are a 'high-quality catalogue of reusable vocabularies to describe Linked and Open Data' (Vandenbussche, et al, 2017). The Linked Open Vocabularies website[37] publishes about 723 vocabularies (e.g SKOS) and 72k terms (e.g., all property names from dcterms). Using linked open vocabulary terms will enable the connection of data from multiple repositories, for example, linking data that are of the same property (e.g datasets of the same subject heading 'climate science', or all datasets from the location X). Furthermore, using dereferencable Uniform Resource Identifiers (URIs) that point to a term or property value will provide unambiguous identification of the reference resource (i.e. does the term "apple" mean fruit in one repository and a corporation in another?), the URLs help provide context to interpret properties precisely.

Recommendation 5: Implement markup syntax consistently by following community practices

Having decided on the scope and the set of properties to be included in the structured data, the next step is to syntactically mark up and serialise the structured data. As discussed in Section 3, Schema.org and its three serialisations, RDFa, microdata and JSON-LD[38], make it easy to embed structured metadata into a resource's web page. These serialisations are to declare the type and the properties of a resource (as shown in Figure 4), as each property is expressed as a pair of `"property name":"property value"`. This recommendation takes JSON-LD as an example, as JSON-LD is designed as a lightweight way to express RDFa and microdata; its adoption is also favoured by major search engines[39] [40].

It is recommended to refer to the implementation guidelines (Jones, et al. 2021) from the ESIP Schema.org cluster[41] for detailed implementation guidelines for each required and recommended data property, for 'dataset' and 'dataCatalogue'. For a complete JSON-LD Syntax, one can refer to W3C recommendation "JSON-LD 1.1 A JSON-based Serialisation for linked data" (W3C, 2021a), and W3C JSON-LD Best Practices for examples (W3C, 2021b).

[37] https://lov.linkeddata.es/dataset/lov
[38] JSON for Linked Data: https://json-ld.org/
[39] Google general structured data guidelines: https://developers.google.com/search/docs/guides/sd-policies
[40] Microsoft Bing - Webmaster tools help and how to: https://www.bing.com/webmasters/help/url-inspection-55a30305
[41] science-on-schema.org

We provide a brief overview below with examples to illustrate 'suboptimal' and good practice' examples when possible. Unlike the 'suboptimal' examples, the good practice examples clearly specify the type of resource object, resource property and property value; since the structured data is primarily for machines to process, machines require proper context to interpret presented data.

- First use the JSON-LD context to declare the namespaces to specify where named properties are defined, as properties from different namespaces may have the same or similar nomenclature but potentially different semantic meaning, or a different range of valid values.
    - Use `@context` to declare namespaces, e.g., `"@context":"https://schema.org/"`
    - Use `@type` to specify the described item, e.g. `"@type":"Dataset"`, or `"@type":"SoftwareSourceCode"`

      The combination of @context and @type will tell a parser of the structured data to interpret described properties from, for example, https://schema.org/Dataset.
    - Use `@context` to declare property terms from namespaces or vocabularies other than Schema.org, for example:

      ```
      "@context":{
            "@vocab":"https://schema.org/",
            "skos":"http://www.w3.org/2004/02/skos/core#"
      }
      ```

      Then reference any property from "skos" with its prefix, for example: `skos:Concept.` Any properties that do not have a prefix (e.g. `"name":"..."`) are defined by the namespace from the `"@vocab"` value.

- Clearly specify the type if a property value is expected to be of that type.
    - E.g, the expected values for the property "creator" are the type "Person" or "Organisation".

Suboptimal example	Good practice example

`"creator":"Peter Smith"`
 	
`Or acceptable example:`
 	
`"creator":{`
`"@type":"Person",`
`"givenName":"Peter",`
`"familyName":"Smith"`
`}`
	`"creator":{`
	`"@type":"Person",`
	`"givenName":"Peter",`
	`"familyName":"Smith",`
	`"sameAs": "http://orcid.org/0000-0000-0000-0000"`
	`}`

- Use an array instead of repeating multiple “property name”:”property value” pairs, where a property may have multiple values.

Suboptimal example	Good practice example
“keywords”:“data science, metadata, structured data” Or “keywords”:“data science”, “keywords”:“metadata”, “keywords”:“structured data”	“keywords”:[“data sciences”, “metadata”, “structured data”]

- Using a structured hierarchy instead of flat one, since reflecting the structure in JSON-LD helps to parse the semantic meaning of each property.

Suboptimal example	Good practice example

“spatialCoverage:”{
“@type”:”Place”,
“latitude”:xx.xxx,
“Longitude”:xx.xx
}
	“spatialCoverage:”{
	“@type”:“Place”,
	“geo”:{
	“@type:“GeoCoordinates”,
	“latitude”:xx.xxx,
	“Longitude”:xx.xx
	}
	}

- Always assign a global persistent identifier (PID) to a resource or a property, where possible. Providing PIDs removes ambiguity around a property/entity, and also helps aggregators link to the ‘source of truth’ when displaying a metadata record,

Suboptimal example	Good practice example

“creator”:{
“@type”:“Person”,
“givenName”:“Peter”,
“familyName”:“Smith”
}
	“creator”:{
	“@type”:“Person”,
	“givenName”:“Peter”,
	“familyName”:“Smith”,
	“sameAs”:”http://orcid.org/0000-0000-0000-0000”
	}

- Use controlled vocabularies and other semantic resources as much as possible. Schema.org introduced two new types, DefinedTerm and DefinedTermSet, for enabling the use of (published) controlled vocabularies to describe a property or property value such as license and keywords etc (Wallis 2018).

Suboptimal example	Good practice example

"keywords": ["geology", "soil sciences"]	"keywords":[
	{
	"@type":DefinedTerm",
	"url":"http://purl.org/au-research/vocabulary/anzsrc-for/2008/0403",
	"Name":"geology",
	"termCode":"0403",
	"inDefinedTermSet":"https://vocabs.ardc.edu.au/repository/api/lda/anzsrc-for/concept"
	},
	{
	"@type":"DefinedTerm",
	"url":"http://purl.org/au-research/vocabulary/anzsrc-for/2008/0503",
	"name":"Soil Sciences",
	"termCode":"0503",
	"inDefinedTermSet":"https://vocabs.ardc.edu.au/repository/api/lda/anzsrc-for/concept"
	},
	{
	"@type":"DefinedTermSet",
	"url":"https://vocabs.ardc.edu.au/repository/api/lda/anzsrc-for/concept",
	"name":"ANZSRC Field of Research Vocabulary Service (ABS 1297.0)"
	}
	]

In this example, it is OK to use text terms for the property "keywords", however, if keyword terms are from a published and community-adopted controlled vocabulary, it is recommended to use the type "DefinedTerm" and its property "url" to specify where the terms are defined, and the property "DefinedTermSet" specifying where the controlled vocabulary is published.

- Use external vocabulary for relation properties if necessary to clearly specify the relationship of a resource in marking up to other related resources, linking related resources on the Web will enhance data discovery (ref. Recommendation 1). As discussed in Section 2, the Schema.org vocabularies are intentionally minimalistic, and don't have

rich relation properties as required by the research community for describing a dataset (ref. Recommendation 4). This shortcoming can be overcome by incorporating terms from external vocabularies, taking the W3C Provenance Ontology (PROV-O) (W3C, 2013) for an example to show relation between a dataset and software that generated the data:

```
{
  "@context": {
    "@vocab": "https://schema.org/",
    "prov": "http://www.w3.org/ns/prov#"
  },
  "@id": "https://doi.org/xx.xxxx",        # dataset
  "@type": "Dataset",
  "Name":"the dataset name”
   ...
  “prov:wasGeneratedBy”:{“@id”:”https://doi.org/xx.xxxx”}   # software
}
```

Recommendation 6: Be friendly to web crawlers

After structured metadata are properly implemented and embedded in a metadata landing page, the next step is to mark the URL (i.e., address) of the landing page into the sitemap of a repository, so that web applications like crawlers can follow the sitemap to find the landing page, and add or update that page in its searchable index. Some repositories that have already implemented structured metadata often complain that not all their landing pages are indexed by a web search engine, and feel frustrated not knowing the reason. Each crawler may have its rules (and limitations) on how and what to follow from a sitemap for optimising their user search experience; repository owners are recommended to check the rules from the target application for instruction on how to construct a sitemap. Failing to follow those rules may result in some metadata landing pages not being indexed. This recommendation addresses only those issues that may require special attention from a data repository. The recommendation does not guarantee each landing page with structured metadata will be indexed by web dataset search tools, however, it may help the diagnosis of why some landing pages are not being indexed.

- A metadata record may go through multiple revisions. A data repository may hold a metadata record for each revision with highly overlapping content (even each version has its own DOI for some repositories). A keyword search resulting in 10 metadata records for the same dataset does not constitute a good user search experience, especially where the latest metadata record version is ranked low on that list. In such instances, **a repository should include only the url of the latest version into its sitemap**, with the landing page including links to all previous versions of that metadata record.
- If changes are made to a metadata record, but the changes are trivial and don’t impact discoverability, then it is recommended not to update the html tag `lastmod`. In other

words, **update the tag `lastmod` only when substantial changes are made to a metadata record**, avoiding unnecessary reindexing

- A crawler may be limited in the number of URLs in a sitemap file, or else in file size it interrogates. Some repositories, especially aggregators, may have numerous metadata records; listing all urls in a sitemap may exceed the limitation of a crawler. In such cases, one can **split a single, large sitemap into several smaller sitemaps, and set up a sitemap index file to point to each sitemap**[42], for example:

```
<?xml version="1.0" encoding="UTF-8"?>
<sitemapindex
xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
   <sitemap>
      <loc>http://www.example.com/sitemap1.xml</loc>
      <lastmod>xxxx-xx-xx</lastmod>     </sitemap>
   <sitemap>
      <loc>http://www.example.com/sitemap2.xml</loc>
      <lastmod>xxxx-xx-xx</lastmod>
   </sitemap>
</sitemapindex>
```

Here the tag `lastmod` is optional, since it indicates the time the corresponding sitemap, not the individual page listed in the sitemap, was modified.

Recommendation 7: Utilise tools that can help you (customise, create, mark up, validate, extract)

There are tools available that can help generate metadata crosswalks, add vocabulary markup to metadata, and to validate resultant markup. In addition to this guidance, the Research Metadata Schemas WG has collected a list of such tools[43]. These tools focus on freely available and/or open source projects. Tools can be grouped into 3 categories - generation, validation, and harvesting.

Generation

Markup generation tools assist with the creation of markup, and in some cases align with certain guidelines or recommendations. Some other generation tools execute crosswalks from other existing meta(data), including sources such as ISO 19115[44], DataCite, or Dublin Core. As

[42] Split up your large sitemap: https://developers.google.com/search/docs/advanced/sitemaps/large-sitemaps

[43] RDA Research Metadata Schemas WG / Tollings: https://github.com/rd-alliance/Research-Metadata-Schemas-WG/blob/master/Toolings/Toolings%20for%20working%20with%20schema.org%20-%2020210128.csv

[44] ISO 19115-1:2014 Geographic information - Metadata: https://www.iso.org/standard/53798.html

indicated in Recommendation 3, the Research Metadata Schemas WG has collected a set of such crosswalks[45], and these crosswalks can be visualised through the tool – Schema Crosswalk Visualisations[46].

These tools include the following:

Tool	Description
CodeMeta generator	For describing software projects with schema.org extensions to SoftwareApplication and SoftwareSourceCode
GeoCodes	For describing scientific datasets using schema.org vocabulary
Schema <Generator>	For describing any schema.org
Dendro	Data management platform supporting multiple ontologies + schema.org metadata

Validation

Validation tools can check if the structured data, either in JSON-LD or RDFa, is formatted correctly. Failing a validation test may result in a web page not being indexed, or not being properly displayed in a search result. These tools include the following:

Tool	Description
Google Structured Data Testing Tool	Provides a display of how Google interprets schema.org markup, including their own recommendations on 'required', and 'recommended' properties/values. The tool accepts URLs or inline markup. This tool is being Deprecated in favor of the Google Rich Results Tool[47].
Science-on-Schema.org Chrome plugin	Will validate the schema.org markup of the current page in Chrome against the science-on-schema.org guidelines (Jones, et al, 2021).

Note: Use Google's tool - Inspect a live URL[48] to find out if a list of URLs from the same domain or an individual URL has been indexed by Google.

[45] Crosswalks from schemas to schema.org: https://github.com/rd-alliance/Research-Metadata-Schemas-WG/blob/master/crosswalks/Crosswalks04092020.csv
[46] Schema Crosswalk Visualisations: https://rd-alliance.github.io/Research-Metadata-Schemas-WG/
[47] Google rich results test tool: https://support.google.com/webmasters/answer/7445569
[48] Google tool: inspect a live URL: https://support.google.com/webmasters/answer/9012289#test_live_page

Harvesting

Harvesting tools focus on the consumption of existing markup. This includes use cases such as validation reporting on existing markup or the aggregation of multiple markups for the construction of knowledge graphs.

Tool	Description
Gleaner	harvesting, validation and indexing of JSON-LD schema.org published in web pages

Recommendation 8: Document and share the whole process

Documenting the Schema.org implementation process, reasoning, and considerations will help existing and new repository staff understand the implementation in a way that allows for future improvements to be implemented effectively and efficiently. Additionally, the documentation will allow easier identification of potential problem areas and future discussions on community best practice. Metadata schemas are reviewed regularly to ensure that the purpose is meeting expectations, and so this will not only improve processes for one particular research community, but also potentially the larger research community.

It is recommended that documentation:

- Specifies each step as discussed in the Recommendation 1 to 8 wherever applicable, including the supporting schemas and crosswalks implemented (i.e., the use different categories, such as mandatory, recommended, optional) so it is clear what the minimum is and how to go beyond
- Provides enough examples (both mapping and implementation), covering common scenarios within the community of use
- Includes information such as which repositories are harvesting your data, and if semantic markup was used by those harvesters. These two things will help new implementers in the same community see what a successful implementation looks like from both the home repository, and the harvesting repository, which can be very useful in the grand scheme of technical implementation.
- Include who a list of adopters of the recommended process especially if the publication process is community-led,
- Considers publishing and making the documentation findable and accessible to wider communities via the web, so other repository owners can learn, follow or adapt their own approach.

Recommendation 9: Identify and join a community, and follow their established practices

It has been emphasised in the previous recommendations that one should not reinvent the wheel. There may already exist communities that provide either a guideline or tools that facilitate steps in the metadata publishing process. Joining and contributing to a well-known and established community has the following advantages:

- It will enable a repository to leverage expertise from that community, thus saving resources and time which would otherwise be expended in exploring routes that may have already been explored
- It will enable consistent implementation at the element, semantic and syntactic level of interoperability, and achieve maximum metadata harmonisation across repositories, aggregators and data discovery service providers;
- Almost all schemas are evolving; a sustainable community will review a schema and its applications (e.g., crosswalk, content generation) regularly in order to meet new requirements, and inform community members of changes. Any schema that requires revision will go through a community consultation process and have strong community support behind a change. Joining such a community will enable your specific use case to be considered; For example, after a community consultation, the bioschemas.org community proposed new types and properties to Schema.org to allow for description of life science resources[49].

The community element is very important whenever exposing structured data as community agreements will guide some of your decisions. Here we include some examples together with at least one of their supported types and a page using it.

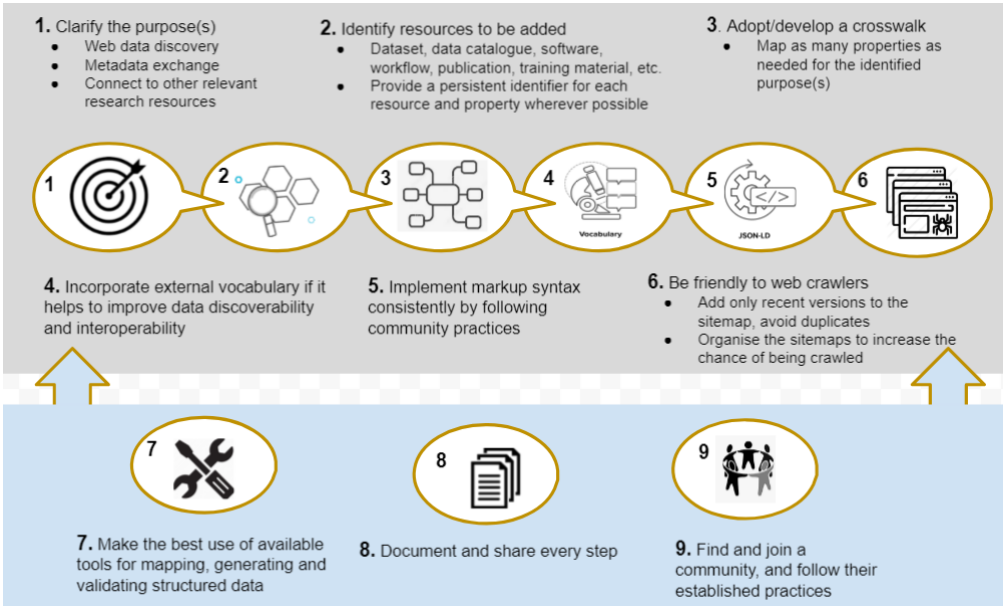
Community	Supported type example	Specification page	Implementation example
Bioschemas	Dataset (adapted from Schema.org)	https://bioschemas.org/profiles/Dataset	http://www.cathdb.info/
Bioschemas	ChemicalSubstance (own type)	https://bioschemas.org/profiles/ChemicalSubstance	https://www.nanocommons.eu/
CodeMeta	SoftwareSourceCode (adapted from Schema.org)	https://codemeta.github.io/terms/	https://github.com/ropensci/codemetar/blob/master/codemeta.json
Science-on-Schema.org	Dataset Data Repository	https://github.com/ESIPFed/science-on-	https://github.com/ESIPFed/science-on-

[49] https://bioschemas.org/types/

	(reuse of Schema.org - ResearchProject, - Organization, - Service)	schema.org/blob/master/guides/Dataset.md	schema.org/blob/master/examples/dataset/full.jsonld
IGSN Global Sample Number (Fils et al. 2020)	Object (own type)	https://github.com/IGSN/igsn-json/tree/master/schema.igsn.org/json/description	https://github.com/IGSN/igsn-json/blob/master/examples/examples.md[50] (Devaraju et al. 2016)

5. Summary

This guideline suggests 9 recommendations that support each stage of the structured data publishing process, as shown in Figure 5. Each recommendation points to available community resources if available. This working group plans to work with potential adopters to validate, enrich or extend the recommendations to make the guidelines more practical to data repositories who plan to publish structured data. Having structured data published semantically and syntactically consistently across repositories will make it easier to harmonise metadata across repositories and build applications at scale, leading to FAIRer metadata, and making data and other web exposed resources more findable by data seekers.



[50] Current IGSN-JSONSchema and examples are provided for (physical) objects from the Earth and Environmental Science domain. Schemas for objects from other domains are under development.

Figure 5: Mapping recommendations to structured data publishing process

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:

- Fotis E. Psomopoulos, Siri Jodha Khalsa, Rafael C Jimenez, Nick Juty and Stephen Richard who helped to set up the Schema.org task force from the RDA Data Discovery Paradigms IG and then this RDA Research Metadata Schemas Working Group;
- Jens Klump (CSIRO) for providing the IGSN-JSONSchema example and references;
- 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), 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 group's survey on "Current practices in using schemas to describe research datasets" and/or crosswalks;
- Karen Payne, Seiya Terada and Chantelle Verhey (Ocean Network, Canada) for developing a suite of tools for visualising the collected crosswalks;
- Joel Benn, Melanie Barlow and Catherine Brady from ARDC for many fruitful discussions, some of which are included in this document.

References:

Barker, P. and Whyte, A. (2020). Harmonizing Metadata for Exchange of FAIR Training Materials. DOI: 10.5281/zenodo.4382676

Greta, B., Caroll, M., Clough, P., Einaudi, M., Ewing, C., Mixter, J., Roy, J., Tomren, H., Washburn, B. and Williams, E. (2021). Transforming Metadata into Linked Data to Improve Digital Collection Discoverability: A CONTENTdm Pilot Project. Dublin, OH: OCLC Research. https://doi.org/10.25333/fzcv-0851.

Canham, Steve. (2020). ECRIN Metadata Schemas for Clinical Research Data Objects Version 5.0 (October 2020) (Version 5.0). Zenodo. DOI: 10.5281/zenodo.1312538

Chan, Lois Mao and Zeng, Marcia Lei (2004). Metadata Interoperability and standardisation - a study of methodology - Part I. Achieving interoperability at the schema level. D-Lib Magazine,12(6). Retrieved fromhttp://www.dlib.org/dlib/june06/chan/06chan.htm

Corcho, O., Kurowski, K., Ojstersek, M., Choirat, C., van de Sanden, M. and Coppens, F. (2020) EOSC Interoperability Framework (V1.0) - 3 May 2020 Draft for community consultation. Available from: https://www.eoscsecretariat.eu/sites/default/files/eosc-interoperability-framework-v1.0.pdf

Devaraju, A., Klump, J. F., Cox, S. J. D., & Golodoniuc, P. (2016). Representing and Publishing Physical Sample Descriptions. *Computers & Geosciences*, *96*, 1–10. https://doi.org/10.1016/j.cageo.2016.07.018

Gil, Y., Cheney, J. Girth, Pl, Hartig, O., Miles, S., Moreau, L. and da Silva, P. P. et al. (2010) Provenance SG Final Report - W3C Incubator Group Report 08 Dec. 2010. Available: https://www.w3.org/2005/Incubator/prov/XGR-prov-20101214/

Godby, C. J., Wang, S. and Mixter, J. K. (2015). Library Linked Data in the Cloud: OCLC's Experiments with New Models of Resource Description. Available from: https://doi.org/10.2200/S00620ED1V01Y201412WBE012

Gogina, Inna. (2016) The World Digital Library: Metadata Crosswalks. Available from: https://innagogina.files.wordpress.com/2016/10/info281-metadata_research-paper.pdf

Gregory, K., Groth, P., Scharnhorst, A., & Wyatt, S. (2020). Lost or Found? Discovering Data Needed for Research Harvard Data Science Review. https://doi.org/10.1162/99608f92.e38165eb

Guha, V., Brickley, D., and Macbeth, S. "Schema.org: Evolution of structured data on the Web: Big data makes common schemas even more necessary". Query, November 2015, https://doi.org/10.1145/2857274.2857276

Fils, D., Klump, J., & Robertson, J. (2020, June). Connecting Data to The Physical World: IGSN 2040 Sprint Outcomes and Recommendations. Technical Report. https://doi.org/10.5281/zenodo.3905364

Freire, N., Charles, V. and Isaac, A. (2018) Evaluation of Schema.org for Aggregation of Cultural Heritage Metadata. In: Gangemi A. et al. (eds) The Semantic Web. ESWC 2018. Lecture Notes in Computer Science, vol 10843. Springer, Cham. https://doi.org/10.1007/978-3-319-93417-4_15

Jones, M. B., Richard, S., Vieglais, D., Shepherd, A., Duerr, R., Fils, D., and McGibbney, L. (2021). Science-on-Schema.org (Version 1.2.0). Zenodo. https://doi.org/10.5281/zenodo.4477164

Kacprzak, E., Koesten, L., Ibáñez, L. D. Blount, T., Tennison, J. and Simperl, E. (2019). Characterising dataset search—An analysis of search logs and data requests. Journal of Web Semantics, Vol.55. pp.37-55. DOI:10.1016/j.websem.2018.11.003

Kato, M. P., Ohshima, H., Liu, Y.-H., & Chen, H. (2020). Overview of the NTCIR-15 Data Search Task. Proceedings of the 15th NTCIR Conference on Evaluation of Information Access Technologies. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings15/pdf/ntcir/01-NTCIR15-OV-DATA-KatoM.pdf

Kraker, P., Schramm, M. and Kittel, C. (2021). Discoverability in (a) crisis. *ABI Technik*, vol. 41, no. 1, 2021, pp. 3-12. https://doi.org/10.1515/abitech-2021-0003

Merz, K. M., Amaro, R., Cournia, Z., Rarey, M., Soares, T., Tropsha, A., Wahab, H. A., & Wang, R. (2020). Editorial: Method and Data Sharing and Reproducibility of Scientific Results. Journal of Chemical Information and Modeling, 60(12), 5868–5869. DOI: 10.1021/acs.jcim.0c01389

McMurry, J. A., Juty, N., Blomberg, N. et al. (2017). Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLOS Biology. June 29, 2017. DOI: 10.1371/journal.pbio.2001414

Munafò, M. (2016). Open Science and Research Reproducibility. eCancer medical science, 10. DOI: 10.3332/ecancer.2016.ed56

National Information Standards Organization. (2004). Understanding Metadata. Retrieved from https://www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf

Noy, N. (2018) Making it easier to discover datasets. Published Sept 5. 2018. Google Blog. Available from: https://www.blog.google/products/search/making-it-easier-discover-datasets/

Pampel H, Vierkant P. (2015) Current Status and Future Plans of re3data.org -Registry of Research Data Repositories. In: Wagner J, Elger K, editors. GeoBerlin2015: Dynamic Earth from Alfred Wegener to today and beyond; Abstracts, Annual Meeting of DGGV and DMG. Berlin, Germany; p. 287—288. Available from:http://gfzpublic.gfz-potsdam.de/pubman/item/escidoc:1369620.

Vandenbussche, P., Atemezing, G. A., Poveda-Villalón, M., Vatant, B. (2017). Linked Open Vocabularies (LOV): A gateway to reusable semantic vocabularies on the Web. Semantic Web 8(3):437-452. Jan. 2017. DOI:10.3233/SW-160213

Taylor, A. (2004). The Organization of Information. 2nd ed. Westport, CN: Libraries Unlimited.

Turpin, A., Scholer, F., Jarveiln, K., Wu, M. and Culpepper, S. J. (2009). Including summaries in system evaluation. In SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference

on Research and development in information retrieval, July 2009 Pages 508–515. https://doi.org/10.1145/1571941.1572029

Vasilevsky, N. A., Minnier, J., Haendel, M. A., & Champieux, R. E. (2017). Reproducible and reusable research: Are journal data sharing policies meeting the mark? PeerJ, 5, e3208. DOI: 10.7717/peerj.3208

W3C (2013). PROV-O: The PROV Ontology. W3C Recommendation 30 April 2013. Retrieved on April 7, 2021, from: https://www.w3.org/TR/prov-o/

W3C (2021a). JSON-LD 1.1: A JSON-based Serialisation for Linked Data. Retrieved on April 7, 2021, from: https://w3c.github.io/json-ld-syntax/

W3C (2021b). JSON-LD Best Practices. W3C Working Group Note 29 March 2021. Retrieved on April 7, 2021, from: https://w3c.github.io/json-ld-bp/

Wallis, R. (2018). Schema.org introduces defined terms. Retrieved on April 7, 2021, from: https://www.dataliberate.com/2018/06/18/schema-org-introduces-defined-terms/

Wilkinson MD, Dumontier M, Aalbersberg IjJ, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 DOI:10.1038/sdata.2016.18

Wu, M., Liu, Y., Brownlee, R., Zhang, X. J. (2021) Evaluating utility of subject headings in a data catalogue. https://drive.google.com/file/d/1dd6F-vNL9S-P2UBnFRGiBkmwWEOUOh-O

Zeng, M. L. (2008). Knowledge Organization Systems (KOS). KNOWLEDGE ORGANIZATION, 35(2–3), 160–182. DOI:10.5771/0943-7444-2008-2-3-160

Appendix A: Survey Summary: Current practices in using schemas to describe datasets

The survey "Current practices in using schemas to describe research datasets" was launched on 27th Feb. 2019. The purpose of the survey was to gather information on existing work involving schemas to describe research data and related resources. The analysis of the survey results will help repositories and the proposed working group understand current practices, identify commonalities, gaps and barriers in using schemas for describing and discovering research datasets; the survey will also inform the work group in planning its objectives and deliverables, along with sharing practices between data repositories.

Twenty organisations/data repositories participated in the survey. The participating repositories cover a wide range of disciplines and metadata schemas. Majority of the repositories (15) have developed mapping between their metadata schema and schema.org, seven of them have implemented structured data markup in metadata landing pages.

Participants reported issues such as rich resource types, relation properties or properties of an entity get lost when they map their metadata to schema.org, The table shows examples of missing properties and types. There are also issues related to use of vocabularies/thesaurus/code lists for property/type values that are not supported by schema.org.

Table: Missing property or relation types in Schema.org for describing dataset as identified from the survey

Resource/Property type	Relation type	Semantic difference
• Scientific measurement • Variable name, variable label, variable definition or description • Environmental entities • Data services / APIs • Tissue samples • Data reuse conditions/consent • Data Controller (legal frameworks • Digital artefacts • Provenance • Licences • Geometries • Some from DataCite: ResourceTypeGeneral, e.g. DataPaper, Model, Workflow • Keyword from external vocabulary (e.g. DefinedTerm, CategoryCode) • Controlled vocabulary from DataCite <dateType>, e.g. Accepted, Available, Copyright, Updated etc.	• Dataset -> FundingAward • Dataset -> Cruise • Dataset -> Funder • Study -> study design • Many from DataCite <relationType>, e.g. isCitedBy, hasVersion, isNewVersionOf, … Issues: • Mapping multiple relation types into one • Not sure if predicates (e.g. in the OBO Foundry Relation Ontology (RO), EnvO, and SWEET) are expressible	For example: • schema:Dataset:name, DataCite:Author:name • Specific term to generic term, e.g. dct:provenance to schema:description

Appendix B: Summary of crosswalks

The group has collected crosswalks from 15 metadata schemas to Schema.org[51]. We have analysed the mappings and classified mapped properties into the following three categories:

Category	**Property name**
Properties can be mapped to those in Schema.org and recommended by Google dataset search guidelines[52]	**In Schema.org**: description, name, identifier, alternateName, sameAs, url, citation, keywords, license, creator, isPartOf, hasPart, version, temporalCoverage, spatialCoverage, measurementTechnique, variableMeasured, url, funder
Properties can be mapped to those in Schema.org	**In Schema.org**: publisher, contactPoint, about or subjectOf, inLanguage, dateCreated, dateCreated, dataPublished, dataModified, copyrightHolder, mentions, isBasedOn, encordingFormat, contentUrl, contentSize, review, contributor, producer, editor, copyrightYear, isAccessibleForFree, position
Properties from DCAT V2/DCAT-AP that have no mapping in Schema.org	**In DCAT V2/DCAT-AP**: rights, accessRights, conformsTo, accuralPerodicity, hasVersion, isVersionOf, provenance

[51] RDA Research Metadata Schemas WG: Crosswalks https://github.com/rd-alliance/Research-Metadata-Schemas-WG/tree/master/crosswalks

[52] https://developers.google.com/search/docs/data-types/dataset