

Guidelines for publishing structured metadata on the web



RDA Recommendation

DOI: [10.15497/RDA00066](https://doi.org/10.15497/RDA00066)

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

Published: 8 October 2021

Version: 3.1

Abstract: The FAIR principles refer frequently to metadata as a key enabler in discoverability, but also having a major role in accessibility, interoperability 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, these guidelines provide nine recommendations to support the process of publishing structured metadata on the web.

Impact: Data repositories would benefit greatly from the recommendations that guide the process of implementing structured metadata. The consistent implementation across data repositories enhances both semantic and syntactic metadata interoperability on the web, which not only makes FAIRer metadata but also enables the creation of better data aggregation and data discovery applications to realise the full potential of open data.

Language: English

License: [Attribution 4.0 International \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/)

RDA webpage: <https://www.rd-alliance.org/group/research-metadata-schemas-wg/outcomes/guidelines-publishing-structured-metadata-web>

Citation and Download: Wu, M., Juty, N., RDA Research Metadata Schemas WG, Collins, J., Duerr, R., Ridsdale, C., Shepherd, A., Verhey, C., & Castro, L. J. (2021). Guidelines for publishing structured metadata on the Web V3.1. Research Data Alliance. DOI: [10.15497/RDA00066](https://doi.org/10.15497/RDA00066).



RESEARCH DATA ALLIANCE

Final Report

Research Data Alliance Research Metadata Schemas Working Group

Guidelines for publishing structured metadata on the web

8 October 2021, V3.1

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

Creative Commons Attribution Only 4.0 license (**CC-BY**)

Table of Contents

Executive Summary	1
Terminology	2
1. Introduction	4
2. Process to publish structured metadata	6
3. Data model	7
4. Recommendations	9
Recommendation 1: Clarify the purpose(s) of your markup (or why you want to markup your data)	10
Recommendation 2: Identify what resources are to be marked up with structured data	11
Recommendation 3: Adopt or develop a crosswalk from your repository schema to Schema.org	14
Recommendation 4: Incorporate external vocabularies as necessary	16
Recommendation 5: Implement markup syntax consistently by following community practices	17
Recommendation 6: Be friendly to web crawlers	21
Recommendation 7: Utilise tools that can help you (customise, create, mark up, validate, extract)	22
Recommendation 8: Document and share the whole process	24
Recommendation 9: Identify and join a community, and follow their established practices	25
5. Summary	26
Acknowledgement:	27
References:	27
Appendix A: Survey Summary: Current practices in using schemas to describe datasets	31
Appendix B: Summary of crosswalks	32

Executive Summary

The FAIR principles refer frequently to metadata as a key enabler in discoverability, but also having a major role in accessibility, interoperability 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, these guidelines provide 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(s) are to be marked up with structured metadata
- 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 metadata
- Recommendation 8: Document and share every step
- Recommendation 9: Find and join a community, and follow their established practices

Terminology

Attribute: A property of an entity.

Community: A body of persons of common and especially professional interests scattered through a larger society¹. For example, this working group is a community of individuals who are interested in or responsible for publishing structured metadata.

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.

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 exchange².

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 schemas 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 entity³.

Instance: An example or single occurrence of something.

Landing page: A web page that a user lands on after clicking a link/URL.

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.

¹ <https://www.merriam-webster.com/dictionary/community>

² University of Texas Libraries: [Crosswalk](#)

³ <https://support.orcid.org/hc/en-us/articles/360006971013-What-are-persistent-identifiers-PIDs->

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

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.

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”.

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 and structured metadata: 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.

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.

1. Introduction

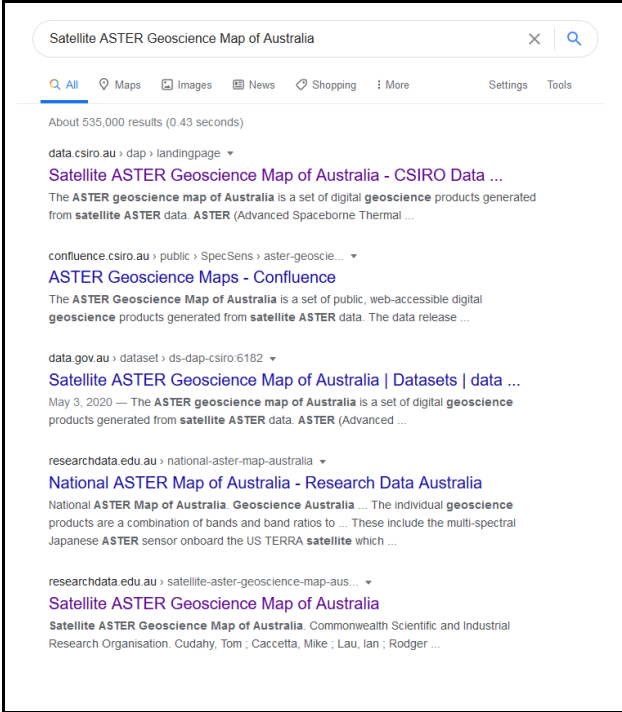
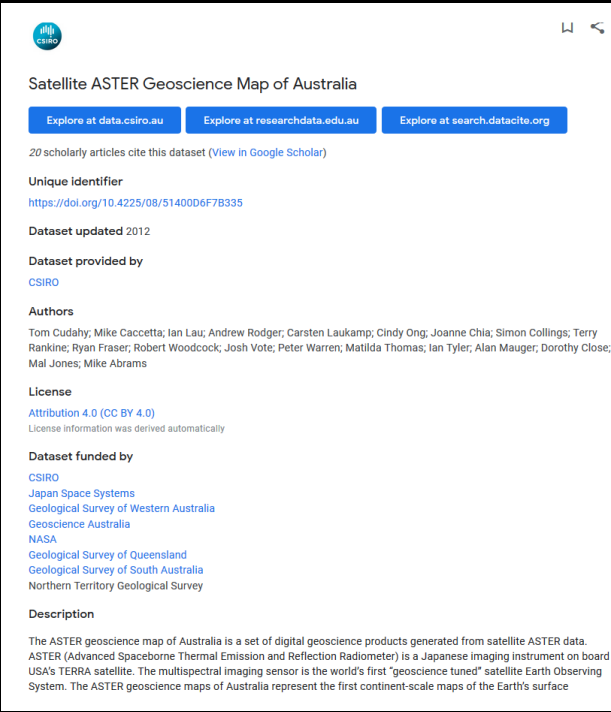
Over the past decade, we have seen an increasing number of public- and domain-specific data repositories, as data sharing is becoming a common scientific practice. Two of the reasons behind the increase of data sharing and data repositories are: improving research reproducibility (Vasilevsky, 2017; Merz, 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 repositories⁴. 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 project 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.

To enable broad discovery and access to research data, some data repositories have begun to leverage data discovery services from global web search engines (e.g. Google, Yahoo, Bing) by embedding structured metadata markup in metadata landing pages using vocabularies such as Schema.org and DCAT. 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 Search⁵ tool clearly shows properties associated with data, enabling users to identify repositories that publish metadata about the same (or similar) datasets.

⁴ <https://blog.datacite.org/german-research-foundation-to-fund-new-services-of-re3data/>

⁵ <https://datasetsearch.research.google.com/>

 <p>The screenshot shows a Google search result page. The search query is "Satellite ASTER Geoscience Map of Australia". The results list several items:</p> <ul style="list-style-type: none"> data.csiro.au › dap › landingpage › Satellite ASTER Geoscience Map of Australia - CSIRO Data ... The ASTER geoscience map of Australia is a set of digital geoscience products generated from satellite ASTER data. ASTER (Advanced Spaceborne Thermal ... confluence.csiro.au › public › SpecSens › aster-geoscie... › ASTER Geoscience Maps - Confluence The ASTER Geoscience Map of Australia is a set of public, web-accessible digital geoscience products generated from satellite ASTER data. The data release ... data.gov.au › dataset › ds-dap-csiro.6182 › Satellite ASTER Geoscience Map of Australia Datasets data ... May 3, 2020 — The ASTER geoscience map of Australia is a set of digital geoscience products generated from satellite ASTER data. ASTER (Advanced ... researchdata.edu.au › national-aster-map-australia › National ASTER Map of Australia - Research Data Australia National ASTER Map of Australia. Geoscience Australia ... The individual geoscience products are a combination of bands and band ratios to ... These include the multi-spectral Japanese ASTER sensor onboard the US TERRA satellite which ... researchdata.edu.au › satellite-aster-geoscience-map-aus... › Satellite ASTER Geoscience Map of Australia Satellite ASTER Geoscience Map of Australia. Commonwealth Scientific and Industrial Research Organisation. Cudahy, Tom ; Caccetta, Mike ; Lau, Ian ; Rodger ... 	 <p>The screenshot shows a detailed dataset page for "Satellite ASTER Geoscience Map of Australia". It includes:</p> <ul style="list-style-type: none"> Buttons to explore the data at data.csiro.au, researchdata.edu.au, and search.datacite.org. 20 scholarly articles cite this dataset (View in Google Scholar). Unique identifier: https://doi.org/10.4225/08/514006F7B335 Dataset updated 2012 Dataset provided by CSIRO Authors: Tom Cudahy; Mike Caccetta; Ian Lau; Andrew Rodger; Carsten Laukamp; Cindy Ong; Joanne Chia; Simon Collings; Terry Rankine; Ryan Fraser; Robert Woodcock; Josh Vote; Peter Warren; Matilda Thomas; Ian Tyler; Alan Mauger; Dorothy Close; Mal Jones; Mike Abrams License: Attribution 4.0 (CC BY 4.0). License information was derived automatically. Dataset funded by CSIRO, Japan Space Systems, Geological Survey of Western Australia, Geoscience Australia, NASA, Geological Survey of Queensland, Geological Survey of South Australia, Northern Territory Geological Survey. Description: The ASTER geoscience map of Australia is a set of digital geoscience products generated from satellite ASTER data. ASTER (Advanced Spaceborne Thermal Emission and Reflection Radiometer) is a Japanese imaging instrument on board USA's TERRA satellite. The multispectral imaging sensor is the world's first "geoscience tuned" satellite Earth observing system. The ASTER geoscience maps of Australia represent the first continent-scale maps of the Earth's surface
<p>Figure 1a: Search result from Google web search engine</p>	<p>Figure 1b: Search result from Google dataset search tool</p>

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 (underlying) 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, and JSON API to SPARQL queries.

Due to its simplicity in implementation, 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 (Guha 2014). 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 bibliographic 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 datasets more discoverable on the Web (Noy, 2018), while some other repositories are about to onboard the path. As Dempsey (2020) points out: discovery often happens elsewhere, it is important that resources from local repositories are effectively represented in external discovery systems used by potential users, and care may be taken that resources are indexed effectively by Google, in order to serve more user communities and their complex data needs.

The Research Data Alliance (RDA) Research Metadata Schema Working Group was formed to facilitate the sharing of data repositories' experiences and lessons learned when publishing structured metadata and to support consistent implementation of the publishing process across repositories. This guideline, as an output of the working group, is to serve those goals.

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 19115⁶ 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.

⁶ ISO 19115-1:2014 Geographic information - Metadata: <https://www.iso.org/standard/53798.html>

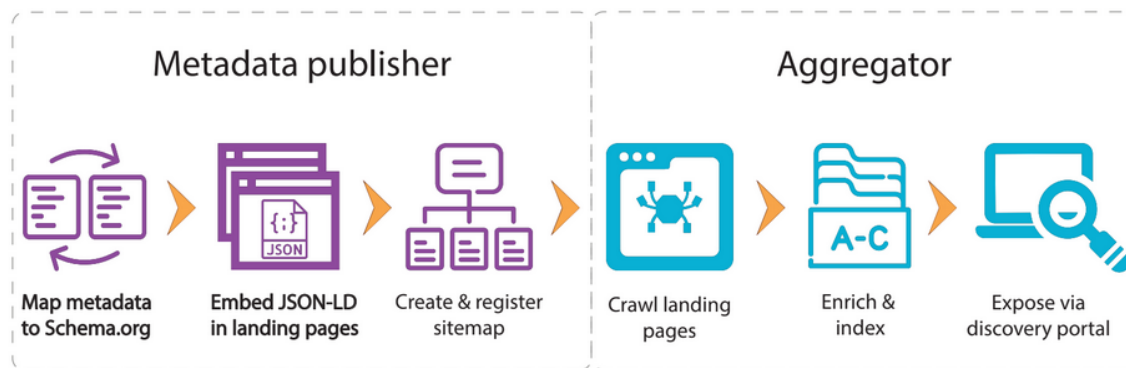


Figure 2: a) Left, the publishing process for the publisher, b) Right, the process for aggregators to make structured data discoverable.

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 crawler 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, they require 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 (W3C RDF, 2002). 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.

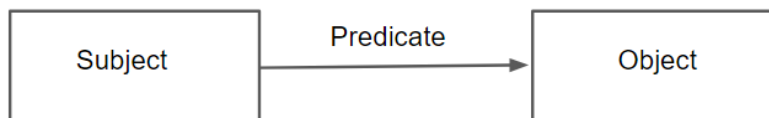


Figure 3: An RDF triple statement

Several standards have been developed to support the RDF data model, for example, the Web Ontology Language (W3C OWL WG, 2012), Simple Knowledge Organisation System (W3C SKOS, 2005), and RDF Schema (W3C RDFs, 2014). 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 resources⁷ and offering a lightweight semantic option for web data providers. As shown in Figure 4, in the Schema.org data model:

- Each resource, or 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'.

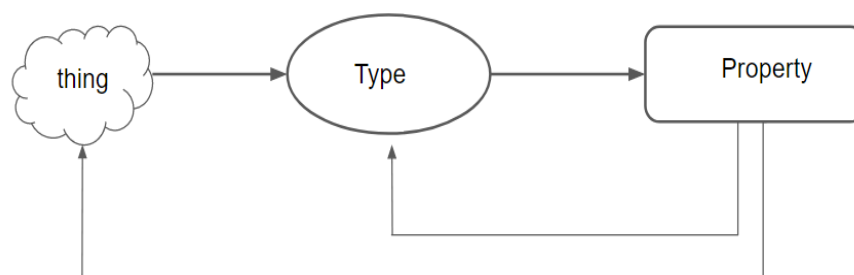


Figure 4: Schema.org data model

⁷ <https://schema.org/docs/datamodel.html>

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 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 (Guha 2014). 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, the Schema.org vocabulary has about 778 types and 1383 properties. The W3C Schema.org Community Group⁸ 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)^{9,10}. As another example, the Bioschemas vocabulary¹¹ has done the same for life science, having successfully incorporated many biomedical terms into the schema.org vocabulary. There are also communities who support the consistent serialisation of the data model. For example, the Schema.org Cluster of the Earth Science Information Partners¹² 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

⁸ <https://www.w3.org/community/schemaorg/>

⁹ [Schema.org Releases: https://schema.org/docs/releases.html](https://schema.org/docs/releases.html)

¹⁰ Learning Resource Metadata is go for Schema: <https://blogs.pjjk.net/phil/lrmi-in-schema/>

¹¹ <https://bioschemas.org>

¹² The ESIP Schema.org Cluster: https://wiki.esipfed.org/Schema.org_Cluster

Metadata Working Group conducted a community consultation¹³, 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 Bioschemas¹⁴, Science-on-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, and which schema, vocabulary and syntactic implementations 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 the coverage of structured data. The most common search feature is keyword search. Keyword search indicates topical relevance between searched keyword(s) 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 data¹⁵. Novel strategies such as these aim to achieve more precise answers to a search query. 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

¹³ [Requirements/Discussions as captured from the RDA P15](#)

¹⁴ <https://bioschemas.org/>

¹⁵ The combination of Wikidata and Bioschemas data is an ongoing project, its current code can be found at <https://github.com/elizusha/graph-loader>

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.

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 requirements and potential usage of metadata.

In either use case (or both), **one needs to first identify the purpose and understand the requirements of downstream metadata consumers, which impacts the scope of the project to set up and 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: <ul style="list-style-type: none"> ● schema:Article ● schema:Movie ● schema:AudioObject ● schema:ImageObject

			Or any other schema:Thing
	Software	Codemeta (essentially schema:SoftwareSourceCode, schema:SoftwareApplication)	schema:SoftwareSourceCode schema:SoftwareApplication
	Data service	dcat:DataService	schema:WebAPI
	Publication (grey publication)	DublinCore ¹⁶ dcterms:BibliographicResource Bibliographic Ontology (BIBO) ¹⁷ bibo:Document bibo:Article bibo:AcademicArticle bibo:Manuscript Semanticscience Integrated Ontology (SIO) ¹⁸ 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 ¹⁹ edam:TrainingMaterial	schema:Course (training) schema:Text, schema:Publication
	Course	bibo:Event	schema:Course, schema:Course:CourseInstance schema:Event:Hackathon,
Responsibility entity	Person	FOAF ²⁰ foaf:Person	schema:Person
	Organisation	W3C recommendation: The Organization Ontology	schema:Organization

¹⁶ <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

¹⁷ <https://bibliontology.com/>

¹⁸ <https://bioportal.bioontology.org/ontologies/SIO>

¹⁹ <http://edamontology.org/>

²⁰ <http://xmlns.com/foaf/spec/>

		(ORG) ²¹ org:FormalOrganization	
	Group	org:Organization	schema:Consortium
	Funding agency	org:Organization Funding, Research Administration and Project Ontology (FRAPO) ²² frapo:FundingAgency	schema:FundingAgency
Other related 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 ²³	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 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 (W3C, 2013), then provenance information includes: where (e.g., location) and how (e.g., software, instrument, model, sensor) data is captured or produced, as well as who (person

²¹ <https://www.w3.org/TR/vocab-org/>

²² FRAPO, the Funding, Research Administration and Projects Ontology: <https://sparontologies.github.io/frapo/current/frapo.html>

²³ RDA Persistent Identification of Instruments WG: <https://www.rd-alliance.org/groups/persistent-identification-instruments-wg>

or organisation) has been involved in its generation, and for what 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, as well as 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 (e.g. 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 and applications to easily identify the authoritative source, and some applications can group duplicate metadata records about the same resource to improve user's data discovery experience. 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. This 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 schema²⁴. We recommend the following practice on the crosswalk:

First, look for existing crosswalks. If a schema (such as ISO 19115:2003, DublinCore, etc.) has already been widely adopted by communities, it is likely that a crosswalk already exists. One

²⁴ Schema crosswalk: https://en.wikipedia.org/wiki/Schema_crosswalk

should first discover and adopt an existing crosswalk, instead of attempting to reinvent the wheel; developing a crosswalk may involve extensive labor 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 Archive³³. 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 that map properties from two schemas, 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 crosswalks²⁵. 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), Bioschemas²⁶, European Clinical Research Infrastructure Network (ECRIN) (Canham, 2020), and Space Physics Archive Search and Extract (SPASE)²⁷.

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 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 guide²⁸ recommends 20 properties, which do not include 'date updated' and 'data provider' fields often used by repositories; however, Google dataset search does parse and render some of these properties when present as shown in Figure 1b (labelled as 'Dataset updated' and 'Dataset provided 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 that dataset search queries often include temporal and spatial properties, as well as the properties data format and file type. The data format and file type properties are in Schema.org, but not in Google Dataset Search guidelines³⁶. If a property is important for a 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

²⁵ RDA Research Metadata Schemas WG / Crosswalks: <https://github.com/rd-alliance/Research-Metadata-Schemas-WG/tree/master/crosswalks>

²⁶ Bioschemas: <https://bioschemas.org/>

²⁷ Space Physics Archive Search and Extract: <https://spase-group.org/>

²⁸ <https://developers.google.com/search/docs/data-types/dataset>

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, 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)²⁹; however, earlier versions of RIF-CS didn't have the target type "download" and 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 an 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)³⁰ 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., news, jobs, music, events, movies, among others), and fewer constraints make it more easily adoptable, though a data repository can use Schema.org together with vocabularies from other standards or namespaces. The incorporation of external vocabularies into Schema.org enriches data search interfaces, such as faceted or filter searches (Wu, et al,

²⁹ RIF-CS to Schema.org crosswalk: <https://documentation.ardc.edu.au/display/DOC/RIF-CS+to+Schema.org+crosswalk>

³⁰ https://www.bodc.ac.uk/resources/vocabularies/parameter_codes/

2021), as well as enables 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 website³¹ 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 markup and serialise the structured data. As discussed in Section 3, Schema.org and its three serialisations, RDFa, microdata and JSON-LD³², 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^{33,34}.

It is recommended to refer to the implementation guidelines (Jones, et al. 2021) from the ESIP Schema.org cluster³⁵ 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).

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

³¹ <https://lov.linkeddata.es/dataset/lov>

³² JSON for Linked Data: <https://json-ld.org/>

³³ Google general structured data guidelines: <https://developers.google.com/search/docs/guides/sd-policies>

³⁴ Microsoft Bing - Webmaster tools help and how to: <https://www.bing.com/webmasters/help/url-inspection-55a30305>

³⁵ science-on-schema.org

data is primarily for machines to process, machines require proper context to interpret presented metadata.

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

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

Suboptimal example	Good practice example
<pre>"keywords": "data science, metadata, structured data"</pre> <p>Or</p> <pre>"keywords": "data science", "keywords": "metadata", "keywords": "structured data"</pre>	<pre>"keywords": ["data sciences", "metadata", "structured data"]</pre>

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

Suboptimal example	Good practice example
<pre>"spatialCoverage": { "@type": "Place", "latitude": xx.xxx, "Longitude": xx.xx }</pre>	<pre>"spatialCoverage": { "@type": "Place", "geo": { "@type": "GeoCoordinates", "latitude": xx.xxx, "Longitude": xx.xx } }</pre>

- 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
<pre>"creator": { "@type": "Person", "givenName": "Peter", "familyName": "Smith" }</pre>	<pre>"creator": { "@type": "Person", "givenName": "Peter", "familyName": "Smith", "sameAs": "http://orcid.org/0000- 0000-0000-0000" }</pre>

- 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
--------------------	-----------------------

<pre>"keywords": ["geology", "soil sciences"]</pre>	<pre>"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)" }]</pre>
---	---

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 an external vocabulary for relation properties, if necessary, to clearly specify the relationship of a resource in marking up to other related resources. 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), while linking related resources on the Web will enhance data discovery (ref. Recommendation 1). This shortcoming can be overcome by incorporating terms from external vocabularies; taking the W3C Provenance Ontology (PROV-O) (W3C, 2013) as an example to show the 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., web 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 properly 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, and may assign a unique DOI to each revision. 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 it can process in a sitemap file, or in the 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**³⁶, 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 was modified, not the individual page listed in the sitemap.

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 markup. In addition to this guidance, the Research Metadata Schemas WG has collected a list of such tools³⁷. 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 19115³⁸, DataCite, or Dublin Core. As indicated in Recommendation 3, the Research Metadata Schemas WG has collected a set of such crosswalks³⁹, and these crosswalks can be visualised through the tool – Schema Crosswalk Visualisations⁴⁰.

³⁶ Split up your large sitemap: <https://developers.google.com/search/docs/advanced/sitemaps/large-sitemaps>

³⁷ RDA Research Metadata Schemas WG / Toolings: <https://github.com/rd-alliance/Research-Metadata-Schemas-WG/blob/master/Toolings/Toolings%20for%20working%20with%20schema.org%20-%2020210128.csv>

³⁸ ISO 19115-1:2014 Geographic information - Metadata: <https://www.iso.org/standard/53798.html>

³⁹ Crosswalks from schemas to schema.org: <https://github.com/rd-alliance/Research-Metadata-Schemas-WG/blob/master/crosswalks/Crosswalks04092020.csv>

⁴⁰ Schema Crosswalk Visualisations: <https://rd-alliance.github.io/Research-Metadata-Schemas-WG/>

These tools include the following:

Tool	Description
CodeMeta generator	For describing software projects with schema.org extensions to <code>SoftwareApplication</code> and <code>SoftwareSourceCode</code>
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 search results. These tools include the following:

Tool	Description
Google Schema Markup Validator⁴¹	Validate all Schema.org-based structured data that's embedded in metadata landing pages, without Google feature specific warnings.
Google Rich Results Tool⁴²	Test your publicly accessible metadata landing page to see what Google rich results can be generated from the structured data the landing page contains.
Science-on-Schema.org Chrome plugin⁴³	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 'Inspect a live URL'⁴⁴ tool to find out if a list of URLs from the same domain or an individual URL has been indexed by Google.

⁴¹ Google tool: Test your structured data:

<https://developers.google.com/search/docs/advanced/structured-data>

⁴² Google rich results test tool: <https://support.google.com/webmasters/answer/7445569>

⁴³ <https://chrome.google.com/webstore/detail/science-on-schemaorg/blpbacoppjgpoedkiglokdhiegajpn?hl=en>

⁴⁴ 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 ⁴⁵	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 periodic assessment and 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:

- Specify each step as discussed in Recommendations 1 to 7 wherever applicable, including the supporting schemas and crosswalks implemented (i.e., the use of different categories, such as mandatory, recommended, or optional), so it is clear what the minimum is and how to go beyond.
- Provide enough examples (both mapping and implementation), covering common scenarios within the community of use.
- Include 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 a list of adopters, especially if the publication process is community-led.
- Consider 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.

⁴⁵ <https://gleaner.io/>

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 be existing communities that provide either a guideline, tools, or both, 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 levels 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, etc.) 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 any 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 resources⁴⁶.

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 example type	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/codemeta/blob/master/codemeta.json
Science-on-Schema.org	Dataset Data Repository (reuse of Schema.org)	https://github.com/ESI-PFed/science-on-schema.org/blob/master	https://github.com/ESI-PFed/science-on-schema.org/blob/master

⁴⁶ <https://bioschemas.org/types/>

	- ResearchProject, - Organization, - Service)	er/guides/Dataset.md	er/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 ⁴⁷ (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 existing community resources. 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.

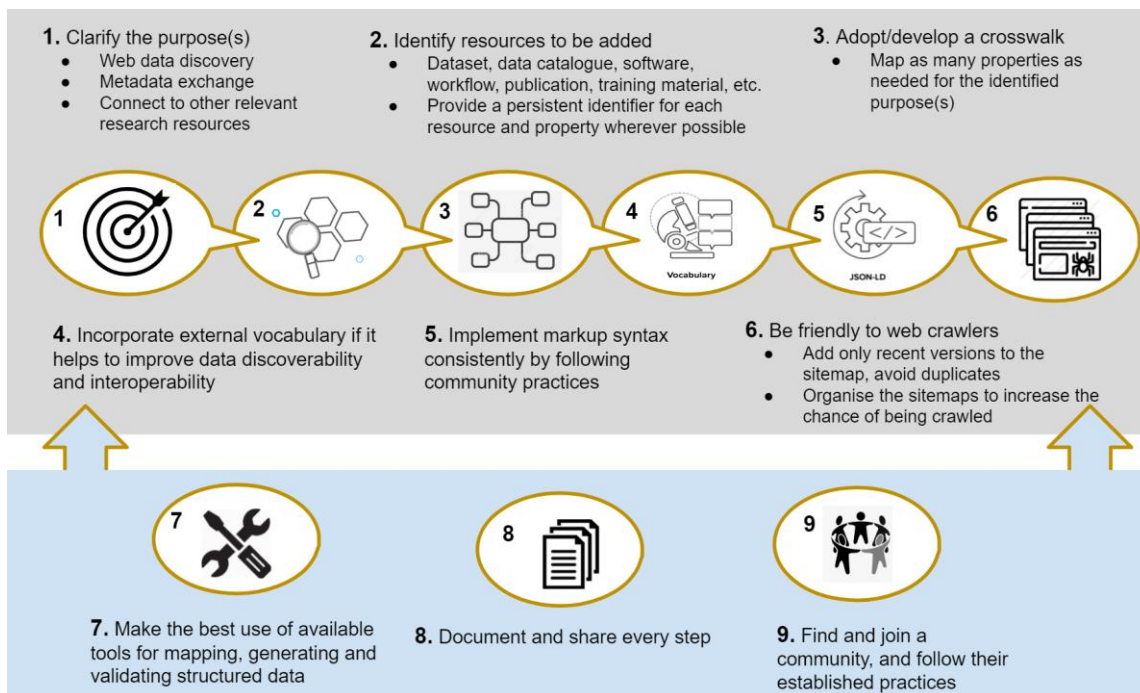


Figure 5: Mapping recommendations to structured data publishing process

⁴⁷ 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.

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.
- The RDA Council for their constructive feedback.

References:

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

Greta, B., Carroll, M., Clough, P., Einaudi, M., Ewing, C., Mixer, 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, S. (2020). ECRIN Metadata Schemas for Clinical Research Data Objects Version 5.0 (October 2020) (Version 5.0). Zenodo. DOI: [10.5281/zenodo.1312538](#)

Chan, L. M. and Zeng, M. L. (2004). Metadata Interoperability and standardisation - a study of methodology - Part I. Achieving interoperability at the schema level. *D-Lib Magazine*,12(6). Retrieved from <http://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>

Dempsey L. (2020). Library Discovery Direction. In McLeish, Simon (ed.). 2020. *Resource Discovery for the Twenty-First Century Library: Case Studies and Perspectives on the Role of IT in User Engagement and Empowerment*. London: Facet Publishing.

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>

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

Gil, Y., Cheney, J. Girth, P., 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, I.. (2016) The World Digital Library: Metadata Crosswalks. Available from: https://innagogina.files.wordpress.com/2016/10/info281-metadata_research-paper.pdf

Guha, R. V. (2014) What a long, strange trip it's been. Retrieved on September 3, 2021 from: https://www.slideshare.net/rvquha/sem-tech2014c/11-07_Rise_of_the_consumers

Guha, R. 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>

Jones, M. B., Richard, S., Vieglaiss, D., Shepherd, A., Duerr, R., Fils, D., and McGibbney, L. (2021). [Science-on-Schema.org](https://doi.org/10.5281/zenodo.4477164) (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](https://doi.org/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](https://doi.org/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](https://doi.org/10.1371/journal.pbio.2001414)

Munafò, M. (2016). Open Science and Research Reproducibility. *eCancer medical science*, 10. DOI: [10.3332/ecancer.2016.ed56](https://doi.org/10.3332/ecancer.2016.ed56)

National Information Standards Organization (NISO). (2004). Understanding Metadata. Retrieved from https://www.liter.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](https://doi.org/10.3233/SW-160213)

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

Turpin, A., Scholer, F., Jarvein, 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](https://doi.org/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 IJ, et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 3:160018 DOI: [10.1038/sdata.2016.18](https://doi.org/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](https://doi.org/10.5771/0943-7444-2008-2-3-160)

W3C OWL Working Group (2012). Web Ontology Language (OWL). 2012-12-11. Retrieved on September 3, 2021, from: <https://www.w3.org/OWL/>

W3C SKOS (2005): SKOS Core Vocabulary Specification, W3C Working Draft 2 November 2005. Retrieved on September 3, 2021, from: <https://www.w3.org/TR/swbp-skos-core-spec/>

W3C RDF Schema 1.1 (2014). W3C Recommendation 25 February 2014, Retrieved on September 3, 2021, from: <https://www.w3.org/TR/rdf-schema/>

W3C RDF (2002). W3C Resource Description Framework (RDF): Concepts and Abstract Data Model. Retrieved on September 3, 2021, from: <https://www.w3.org/2002/07/29-rdfcadm-tbl.html>

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. The 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 losing rich resource types, relation properties or properties of an entity when they map their metadata to schema.org, The table shows examples of missing properties and types. There are also issues related to the 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
<ul style="list-style-type: none"> ● 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. 	<ul style="list-style-type: none"> ● Dataset -> FundingAward ● Dataset -> Cruise ● Dataset -> Funder ● Study -> study design ● Many from DataCite <relationType>, e.g. isCitedBy, hasVersion, isNewVersionOf, ... <p>Issues:</p> <ul style="list-style-type: none"> ● Mapping multiple relation types into one ● Not sure if predicates (e.g. in the OBO Foundry Relation Ontology (RO), EnvO, and SWEET) are expressible 	<p>For example:</p> <ul style="list-style-type: none"> ● schema:Dataset:name, DataCite:Author:name ● Specific term to generic term, e.g. dct:provenance to schema:description

Accepted, Available, Copyright, Updated etc.		
---	--	--

Appendix B: Summary of crosswalks

The group has collected crosswalks from 15 metadata schemas to Schema.org⁴⁸. 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 ⁴⁹	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, encodingFormat, 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, accuralPeriodicity, hasVersion, isVersionOf, provenance

⁴⁸ RDA Research Metadata Schemas WG: Crosswalks <https://github.com/rd-alliance/Research-Metadata-Schemas-WG/tree/master/crosswalks>

⁴⁹ <https://developers.google.com/search/docs/data-types/dataset>