Draft Final Report
Research Data Alliance Research Metadata Schemas Working Group

Guidelines for publishing structured metadata on the web

V1.0

(authors, to be added in the final version)

Table of Contents

Executive Summary

Publishing structured metadata on the web can provide a simple and efficient means to increase the FAIRness of web resources; it exposes metadata contained in web pages through a formal mechanism, allowing systematic collection and processing by web-based crawlers. The FAIR principles refer frequently to metadata as it is a key enabler in discoverability, but also plays major roles in accessibility and reusability. The adoption of structured metadata within and across domains would benefit greatly from recommendations for their consistent implementation across data repositories, in order to achieve full potential of interoperability. Based on community consultation and subsequent works, these guidelines provide ten recommendations to support the process of publishing structured metadata on the web, namely:

- Recommendation 1: Identify the purpose of your markup
- Recommendation 2: Identify what resource objects are to be marked up with structured data
- Recommendation 3: Define the metadata schema and vocabularies to be used for markup
- Recommendation 4: Adopt or develop a crosswalk from a repository schema to markup vocabulary
- Recommendation 5: Incorporate external vocabulary
- Recommendation 6: Follow a consistent implementation of markup syntax
- Recommendation 7: Facilitate access to web crawlers
- Recommendation 8: Utilise tools that can help
- Recommendation 9: Document the whole process
- Recommendation 10: Find some community out there (or create your own)

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

[1]University of Texas Libraries: Crosswalk

[2] 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)”.

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 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, being web search tools one of the approaches favoured 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 opposed to domain specific vocabularies with reach 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/

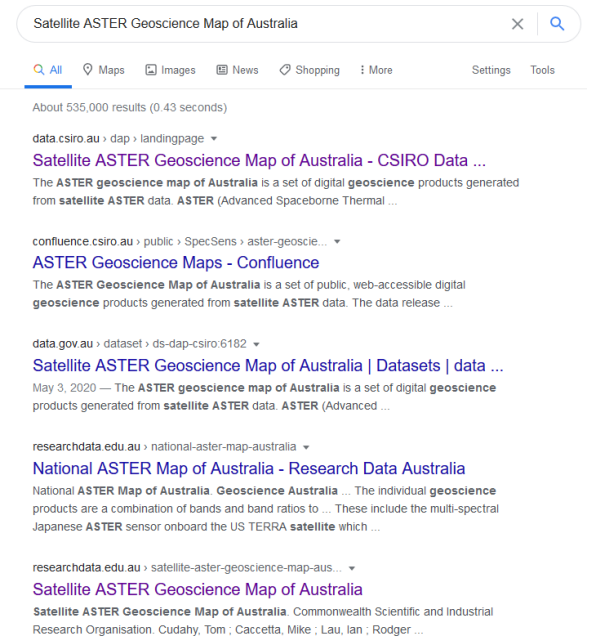

Figure 1a: Search result from google web search engine

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), 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

Improving the availability and consistency of structured data on the web is key to realising the potentials as discussed above. Figure 2 shows a general process for publishing and consuming structured data. Metadata publishers usually undertake the following four steps:

0. Describe repository resources using a suitable metadata schema.
1. Develop a crosswalk from a repository's source metadata to Schema.org, or other community adopted vocabulary such as DCAT if the repository uses a metadata schema other than 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.

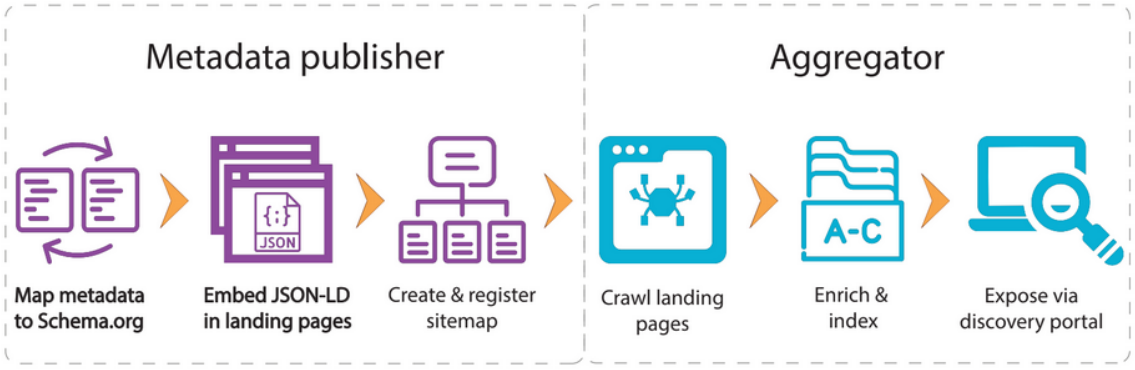

Figure 2: Structured data publishing and aggregating process over the web

Once data repositories provide structured data, a data aggregator will go through the following three steps to consume the structured data:

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) to be 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 semantic or syntactic level will infringe upon the interoperability and reusability of structured data.
- The Schema.org vocabulary adopted or indexed by major web search engines are intentionally minimalistic, for encouraging fast, easy and wide adoption, while leaving room for incorporation of external vocabularies and extensions, if there is a community need.

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[5]. 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.

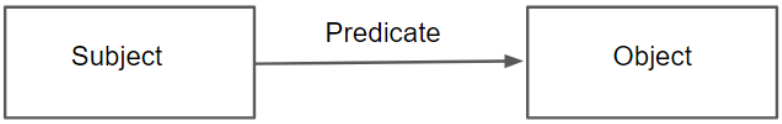

Figure 3: An RDF triple statement

Several standards have been developed to support the RDF data model, for example, the Web Ontology Language (OWL)[6], Simple Knowledge Organisation System (SKOS)[7], and RDF Schema (RDFs)[8]. 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 vocabularies and rules, targeting the description of web resources[9] and offering a lightweight semantic option for web data providers. As shown in Figure 4, in the Schema.org data model:

---

[5] W3C Resource Description Framework (RDF): Concepts and Abstract Data Model https://www.w3.org/2002/07/29-rdfcadm-tbl.html
[6] https://www.w3.org/OWL/
[7] https://www.w3.org/TR/swbp-skos-core-spec/
[8] https://www.w3.org/TR/rdf-schema/
[9] https://schema.org/docs/datamodel.html

- 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’.

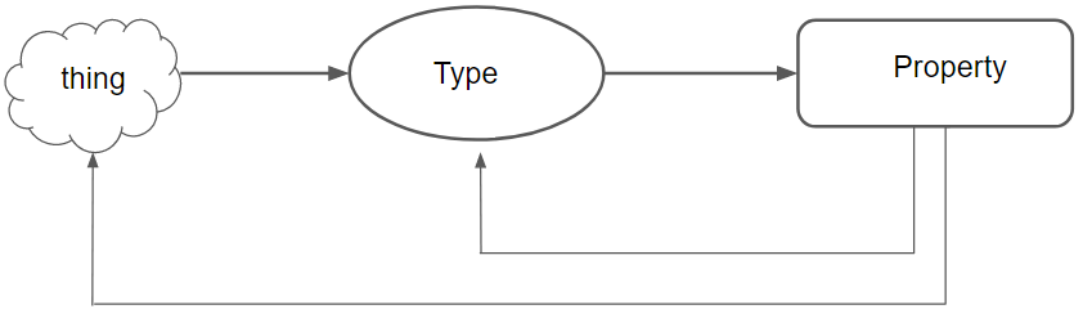

Figure 4: Schema.org data model

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 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[10]. If RDFa, microdata and JSON-LD are

[10] What a long, strange trip it’s been: https://www.slideshare.net/rvguha/sem-tech2014c/11-07_Rise_of_the_consumers

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[11] 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)[12][13]. There are also communities to support the consistent serialistation of the data model, for example, the Schema.org Cluster of the Earth Science Information Partners[14] works on the best practices, education and outreach on the web accessible structured data, for advancing domain-specific needs of improving 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[15], asking participants who were planning to publish structured metadata what they would like to know beforehands (e.g., from others' experience), and to those participants who had already implemented structured metadata, what learnings they could share, particularly pitfalls to avoid. Additional input was also solicited from communities and projects that were active in this area, including Bioschemas[16], Science-on-Schema.org (Jones, et al, 2021) and various library catalogues on the web. We have coalesced these learnings to derive the following ten 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: Identify the purpose 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

[11] https://www.w3.org/community/schemaorg/
[12] Schema.org Releases: https://schema.org/docs/releases.html
[13] Learning Resource Metadata is go for Schema: https://blogs.pjjk.net/phil/lrmi-in-schema/
[14] The ESIP Schema.org Cluster: https://wiki.esipfed.org/Schema.org_Cluster
[15] Requirements/Discussions as captured from the RDA P15
[16] https://bioschemas.org/

structured data:

1. For data discovery by web search engines

   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. Almost all repositories share the same purpose for their data to be as discoverable as possible. A survey shows web search engines are the second most used tool by academics for data search (with searching in literature being the most used) (Gregory, et al. 2019). In fact, some search engines offer tailored search sites for data, e.g., Google DatasetSearch[17].

2. As a way of ingesting metadata to metadata aggregators

   When structured data is published on the web, it can be consumed by anyone who has a desire to implement innovative data search applications. For example, the information retrieval research community has already started a data search track for the purpose of developing a crawler, indexer and search model for open data (Kato et al. 2020). Other efforts in this regard include, for instance, combining Wikidata and Bioschemas data[18]. Novel strategies such as these will likely have a substantial impact on data search and integration.

   Embedding structured data in landing pages offers a new way for metadata aggregators to harvest metadata through web crawling. Currently, if a metadata aggregator harvests metadata from multiple data repositories, or a data repository exports metadata to multiple downstream repositories or catalogues, either the metadata aggregator or the data repository would have to implement and maintain 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 the common markup vocabularies.

Recommendation 2: Identify what resource objects 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

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

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

	Type of resources ("things")	Other standards/Schemas/ Schema Class	Schema.org (type)
Primary entity	Catalogue	dcat:Catalog	schema:DataCatalog
	Dataset	dcat:Dataset	schema:Dataset
	Software	Codemeta (essentially schema:SoftwareSourceCode, schema:SoftwareApplication)	schema:SoftwareSourceCode schema:SoftwareApplication
	Sample	International Geo Sample Number (ISGN)[19]	
	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:publicaton 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

[19] IGSN metadata: https://igsn.github.io/metadata/
[20] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[21] https://bibliontology.com/
[22] https://bioportal.bioontology.org/ontologies/SIO

	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:Organization	schema:Organization
	Group		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	Working in progress from the RDA Persistent	schema:Instrument

[23] http://edamontology.org/
[24] http://xmlns.com/foaf/spec/
[25] https://www.w3.org/TR/vocab-org/
[26] FRAPO, the Funding, Research Administration and Projects Ontology: https://sparontologies.github.io/frapo/current/frapo.html

		Identification of Instruments WG[27]	

The goal of publishing data to the web (or any other platforms) is for wider discoverability; however, discoverability is 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 definition above, and the provenance data model[29], then provenance information includes: where (e.g., location) and how (e.g., software, instrument, models, sensors) data is captured or produced, as well as who (person or organisation) has been involved in its generation, and for which purpose (e.g., projects) it was produced.

It is therefore highly recommended to publish and connect resources, in addition to the dataset itself, that provide provenance information to datasets, improving the likelihood of reproducibility of published datasets, connecting all research related resources into a web of (distributed) data, and increasing discovery paths to the datasets.

Recommendation 3: Define which metadata schema and vocabularies to be used for markup

The selection of a specific metadata schema for a particular repository depends upon:

- Which type of data the repository is going to host
- The user requirements to find, select, and use the data
- The long-term logistics around managing and preserving the data
- The available standard schemas available within that domain

[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/

Schemas range from very generic to extremely discipline specific; discipline-specific schema provide a richer and more targeted array of domain-relevant properties, for example, ISO 19115[30] for geographic data and ECRIN metadata schemas for clinical research data (Canham, 2020). Generic schemas, such as Dublin Core, PROV-O, Schema.org, record only those properties that are common across multiple disciplines. In general, a discipline-specific schema can conceptually be mapped to a generic schema, since the development of the latter would consider a more abstract and wider range of use cases. If both a generic schema and a discipline-specific schema are implemented in this manner, a crosswalk between the two schema types can be properly mapped. There should be no extra burden on metadata capture when a discipline-specific schema/profile is mapped to a generic one. This allows one to express properties and relations at both domain-specific level and at the more abstract level or general level. Either way, data seekers should be able to discover and access the data.

A repository may choose a (discipline specific) schema to meet their needs, and map their chosen schema to another target schema for exchanging metadata and to make data more discoverable. If a repository aims to have web applications to crawl or harvest their structured metadata that are embedded in metadata landing pages, the target schemas currently supported and adopted by most web applications are Schema.org or DCAT. Where properties from a repository schema cannot be clearly mapped to Schema.org or DCAT, web data discovery applications or metadata aggregators usually point back to the original repositories that host the data and more which contain those more specific and granular levels of metadata, as shown in Figure 5a and Figure 5b.

To summarise, the recommendation here is: when a repository chooses a metadata schema and vocabulary for their repository, they should choose the most suitable and community adopted ones for their data and follow the FAIR (meta)data principles (Wilkinson et al., 2016). The more data properties that are captured, the easier it is to map their own schemas to many other general ones.

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

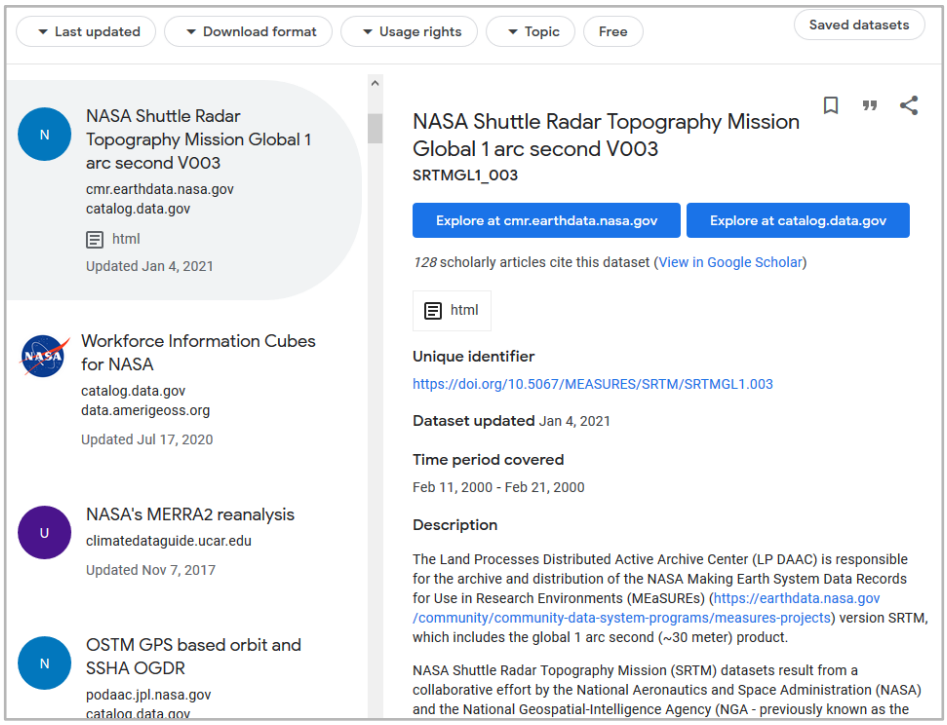

Figure 5a: A metadata record view from an aggregator. The two blues bars under the title point to where the metadata is crawled from

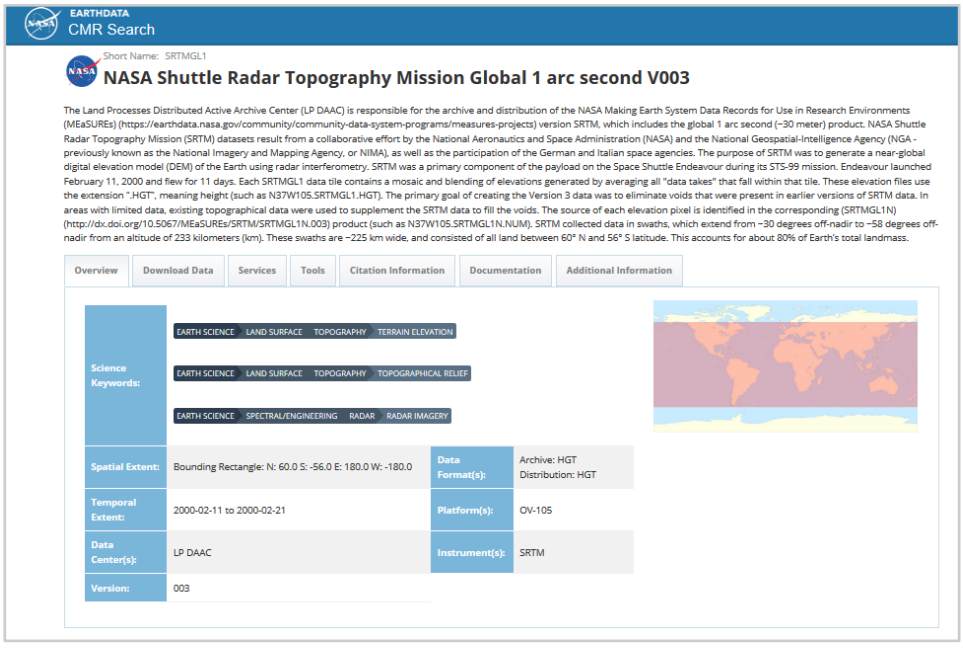

Figure 5b The source metadata record pointed by the aggregator (in Figure 5a)

Recommendation 4: Adopt or develop a crosswalk

In many cases, it may require to do a crosswalk from a repository schema to a more generic markup vocabulary such as Schema.org or DCAT. We recommend the following practice:

First, look for existing crosswalks. If a repository schema has already been widely adopted by communities, it is likely that a crosswalk has already been developed. One should first discover and adopt an existing crosswalk, instead of attempting to reinvent the wheel. This would save valuable resources, since developing a crosswalk may involve extensive labour on concept mapping, and in some cases may require community consultation. Most widely, having the same 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 exists 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, which will make 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 to the destination schema as possible. If a repository has an objective regarding which aggregator it should be harvested by, to make their structured data discoverable, it may map only those properties consumed by the aggregator. Typically, aggregators recommend a set of common properties that could be implemented by the majority of repositories, although that doesn't mean the aggregator is restricted by that set of recommended properties. For example, a record from the Google dataset search (Figure 1), contains 20 properties recommended by the Google dataset search guide[34], which do not include 'date updated', 'data provider' and 'data funder'. However, Google dataset search does parse and render these properties. If this information is important for a user search to judge the relevance of that dataset, and then this information is missing, the user may not refer to the source repository to explore further, as Kacprzak et al. (2019) found that dataset search queries often include

[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

temporal and spatial properties, as well as data format and file type. Apart from foreseen search engines like Google, there may be unforeseen consumers who would harvest structured data as available on the Web, may parse structured data as richly as possible so they can build data discovery tools with more search options apart from keyword search.

Fourth, label a persistent identifier or qualified reference to each resource, including property and property value, with controlled vocabulary. Identifier is a property of a described resource, because of its important uniqueness, it deserves to be discussed separately. 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. Examples of persistent identifiers include Digital Object Identifier (DOI), handle, Universal Resource Name (URN) etc. PIDs can be used by both humans and machines to track, access a resource, link resources, and more importantly, to establish the authenticity of a resource.

The uniqueness and authenticity are important, especially when repositories and aggregators harvest and publish metadata from each other. Using the persistent identifier instead of text to reference a property or related resources enables identification of the same property or term being described in different contexts. For example, in the Figure 2b, three repositories (data.csiro.au, researchdata.edu.au and search.datacite.org) published metadata of the same dataset, while data.csiro.au is the original source and provider of the metadata, and the dataset is downloadable from data.csiro.au, thus DOI points to data.csiro.au. In this example, both repositories researchdata.edu.au and search.datacite.org harvest and publish the metadata from data.csiro.au, because both repositories used the DOI as identifier, the web search tool is easier to identify the three metadata records that describe the same dataset. By attaching the DOI for its source repository, a data discovery tool can avoid duplications in a search result, users are able to identify and follow on to the original repository in order to view more metadata for making relevance assessment or retrieve the data. .

Fifth, connect to other resources/entities. The power of structured data is its connection to other resources or entities published to the web. The connection is through both described property (e.g., the same location, the same creator) and relation to other resources. These related resources may include the same (or different) metadata records, published to different repositories but describing the same dataset. A dataset is a subset or derivative from another dataset, or a dataset that is produced from a software or workflow, etc. These relations should also be included in the crosswalk and mapped to the target schema as much or close as possible.

Sixth, take implementation 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). However, earlier version of

RIF-CS didn't have the target type "download", and past guidelines from the metadata aggregator Research Data Australia (RDA), thus have a large proportion of metadata records in RDA was to use this 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 5: Incorporate external vocabulary

A research data repository may use controlled vocabularies to specify:

- Relation 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;
- A range of property value, for example, Library Congress Subject Heading for indicating topics of a library resource, the BODC Parameter Usage Vocabulary (PUV)[35] for labelling scientific variables.

The purpose of using controlled vocabularies is to standardise information, so that a metadata record is more computationally validatable and interoperable, and content can be better linked and harmonised for improving data discovery.

However, generic schemas such as Schema.org vocabularies don't enforce constraints or recommend controlled vocabularies for property values, don't have rich relations between resource objects that are essential for research provenance and reproducibility. This is a deliberate decision as Schema.org is for data from all domains (e.g.news, jobs, music, event, movie, among others), 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 facet search and filter search (Wu, et al, 2021), as well as 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 the use of 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[36] 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, but of the property (e.g of the same subject heading 'climate science', or all data from the location X), furthermore, that using dereferencable Uniform Resource Identifiers (URIs) points to a term or property value will provide unambiguous identification of the reference

[35] https://www.bodc.ac.uk/resources/vocabularies/parameter_codes/
[36] https://lov.linkeddata.es/dataset/lov

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 6: Follow a consistent implementation of markup syntax

Once a repository has decided on appropriate schemas and vocabularies to structurally describe data at a semantica level, it must decide on the markup language to encode the metadata records with. A repository may encode metadata records in html for displaying for human users in a web browser, or in the Extensive Markup Language (XML) for exchanging metadata records with other repositories. Unfortunately, neither html nor xml is able (or sufficient) to encode the semantic meaning from a chosen schema, or in a consistent way for machine understandability. Thus we need a standard, machine readable format that is capable of marking up semantic meaning of metadata records. When machines can parse and understand the markups, we can develop scalable and intelligent applications on top of that, either displaying for users, exchanging metadata, or supporting more advanced queries that result in more relevant search results than a general web search.

As discussed in Section 3, Schema.org and its three serialisations, RDFa, microdata and JSON-LD, 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 RDAa and microdata, its adoption is also favoured by the popularity of JSON among software engineers and developers.

It is recommended to refer to the implementation guidelines (Jones, et al. 2021) from the ESIP Schema.org cluster[37] for detailed implementation of each required and recommended data properties for dataset and dataCatalogue.

Here is a summary of high-level rules:

- Declare a namespace to specify where named properties are defined, as properties from different properties may have the same name but different semantic meaning, and the type of a described item.
    - Use `@context` to declare namespaces, e.g., `“@context”:``https://schema.org”`
    - Use @type to specify the described item, e.g.`“@type”:``dataset”`

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

        Suboptimal example:

[37] science-on-schema.org

```
"creator":"Peter Smith"
```

Acceptable example:

```
"creator":{@type":"Person", "giveName":"Peter", "familyName":"Smith"}
```

Good practice example:

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

- Use array instead of repeating each "property name":"property value" pairs, when a property has multiple values
  - E.g.,

    Suboptimal example:

    ```
    "keywords":"data science, metadata, structured data"
    ```

    or :

    ```
    "keywords":"data science", "keywords":"metadata", "keywords":"structured data"
    ```

    Good practice example:

    ```
    "keywords":["data sciences", metadata, "structured data"]
    ```

- Using structured hierarchy instead of flat one, as the structure in JSON-LD, helps to parse the semantic meaning of each property.
  - E.g.,

    Suboptimal example:

    ```
    "spatialCoverage:"{"@type":"Place", "latitude":xx.xxx, "longitude":xx.xx}
    ```

    Good practice example:

    ```
    "spatialCoverage:"{"@type":"Place", "geo":{@type:"GeoCoordinates", "latitude":xx.xxx, "longitude":xx.xx}}
    ```

- Always assign a global persistent identifier (PID) if it exists to a resource or a property. Providing PIDs removes ambiguity about a property/entity, also help aggregators link to the source of truth when displaying a metadata record,
  - E.g.,

    Good practice example:

    ```
    "creator":{"@type":"Person", "giveName":"Peter", "familaryName":"Smith", "sameAs":"http://orcid.org/0000-0000-0000-0000"}
    ```

- Use controlled vocabulary and their defined terms as much as possible.
  - E.g.,

Suboptimal example:

```
"keywords": ["geology", "soil sciences"]
```

Good practice example:

```
"Keywords":[
    {"@type":DefinedTerm",
    "url":"http://purl.org/au-
research/vocabulary/anzsrc-for/2008/0403",
     "Name":"geology",
    "termCode":"0403",
    "inDefinedTermSet":"https://vocabs.ardc.edu.au/re
pository/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/re
        pository/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 well 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” where the controlled vocabulary is published.

Recommendation 7: Facilitate access 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 crawler can follow the sitemap to find the landing page, add or update the page into its searchable index. Some repositories who 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, a repository is recommended to check 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 may not guarantee each landing page with structured metadata to 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 overlapped content (even each version has its own DOI for some repositories). A keyword search results in 10 metadata records of the same dataset may not bring in a good user search experience, especially if the latest version that a repository would like a user to find and use is ranked down in the list. In this case, a repository can consider to include only the url of the latest version in a sitemap with the landing page including links to all previous versions.
- If changes are made to a metadata record, but the changes are trivial and don't impact on discoverability, then it is recommended not to update the tag `lastmod`, in another word, update this tag value only when substantial changes are made to a metadata record.
- A crawler may have the limitation on the number of URLs to be listed in and the file size of a sitemap file. A repository (especially an aggregator) may have a very large number of metadata records, listing all urls in a sitemap may exceed the limitation of a crawler. In this case, one can split a big and one sitemap into several smaller sitemaps, and set up a sitemap index file to point to each sitemap[38], 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, it indicates the time the corresponding sitemap, not the individual page listed in the sitemap, was modified.

Recommendation 8: Utilise tools that can help

There are available tools that can help with crosswalk, add vocabulary markup to metadata, and validate the resulted markup. In addition to this guidance, the Research Metadata Schemas WG

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

has collected a list of such tools[39]. 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) sources such as ISO 19115, DataCite, or Dublin Core. As indicated in Recommendation 3, the Research Metadata Schemas WG has collected a set of crosswalks[40], these crosswalks can be visualised through the tool –Schema Crosswalk Visualisations[41]. These tools include the following:

Tool	Description
CodeMeta generator	For describing software projects w. 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 the webpage not being indexed, or not having a proper display as a search result. These tools include the following:

Tool	Description
Google Structured Data Testing Tool	See how Google interprets the schema.org including their own ideas for required, recommended properties. Submit URL or inline markup. Deprecating in favor of the Google Rich Results Tool[42].

[39] 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

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

[41] Schema Crosswalk Visualisations: https://rd-alliance.github.io/Research-Metadata-Schemas-WG/

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

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[43], if one wants to find out if a list of URLs from the same domain or an individual URL has been indexed by Google.

Harvesting

Harvesting tools focus on the consumption of existing markups. This includes use cases such as validation reporting on existing markups or the aggregation of multiple markups for constructing a knowledge graph.

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

Recommendation 9: Document 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 practices. 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 recommending a documentation to:

- Documents 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
- Provide enough examples (both mapping and implementation) covering common scenarios in your community
- Include information such as which repositories are harvesting your data, and if semantic markup was used by their harvesters. These two things will help new implementers in the same community see what a successful implementation looks like from both the home

[43] Google tool: inspect a live URL: https://support.google.com/webmasters/answer/9012289#test_live_page

repository, and the harvesting repository(ies), which can be very useful in the grand scheme of technical implementation.
- If the publication process is community-led, includes who are adopters of the recommended process; that
- Consider publishing and making the documentation findable and accessible to wider communities via the web, so repositories who can learn and may follow or adapt the approach as documented.

Recommendation 10: Find some community out there (or create your own)

It has been emphasised in the previous recommendations that one should not reinvent the wheel if there already exist communities that provide either a guideline or tools that facilitate any step of the publishing process. Joining and contributing to a well-known and well-maintained community has the following advantages:

- It will enable a repository to leverage expertise from community, thus save resources and time to explore routes that may have already been explored by community;
- 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) at a regular time interval in order to meet new requirements, and inform community members of any change. Any schema that requires revision will go through a community consultation process and have a strong community support behind a change, so joining such a community will enable your case being considered; if your case is common among the community, it will be more likely to be considered. For example, after a community consultation, the bioschema.org community proposed new types and properties to Schema.org to allow for description of life science resources[44].

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 types	Page example
Bioschemas	Dataset (adapted from	http://www.cathdb.info/

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

	Schema.org)	
Bioschemas	ChemicalSubstance (own type)	https://www.nanocommons.eu/
CodeMeta	SoftwareSourceCode (adapted from Schema.org)	https://github.com/ropensci/codemetar/blob/master/codemeta.json
Science-on-Schema.org	Dataset Data Repository (reuse of Schema.org -ResearchProject, -Organization, - Service)	https://science-on-schema.org
Learning Resource Metadata Initiative (Phil and Angus, 2020)	LearningResource	https://blogs.pjjk.net/phil/lrmi-examples-for-schema-org/

5. Summary

This guideline suggests 10 recommendations that support each stage of the structured data publishing process, as shown in Figure 6. 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, this will lead to FAIRer metadata, make data and other represent resources more findable by data seekers.

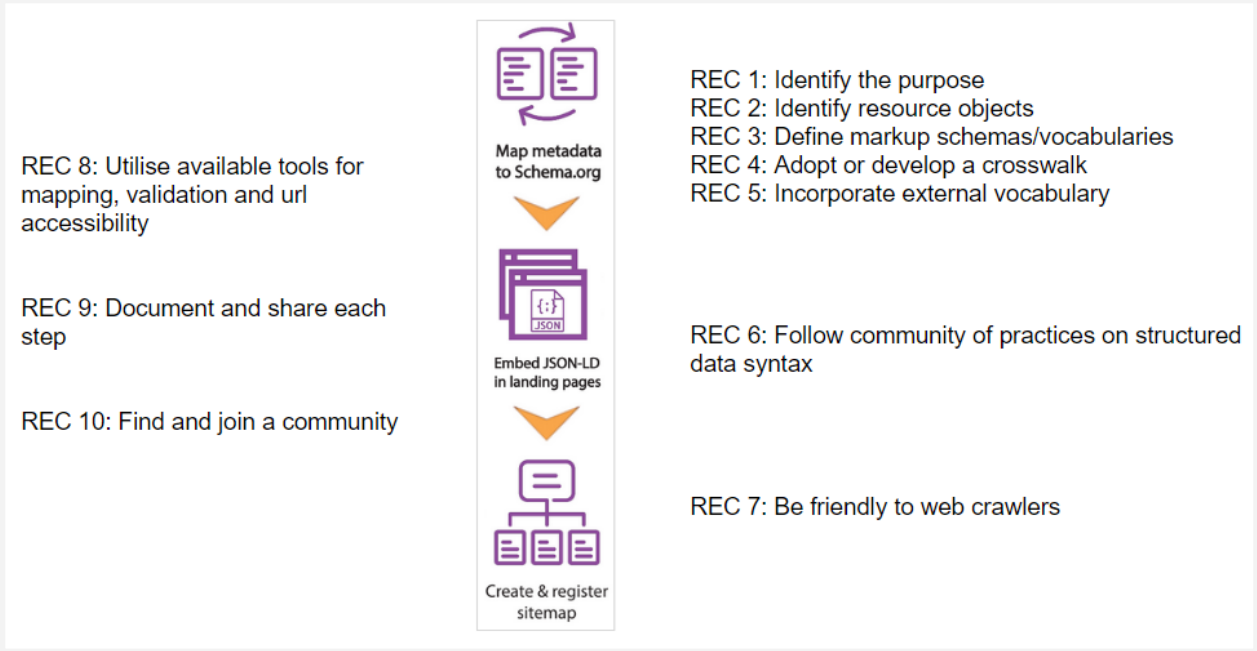

Figure 6: Mapping recommendations to structured data publishing process

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 discussion. 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;
(will add more people who should be thanked here ……).

References

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

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

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

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

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

Munafò, M. (2016). Open Science and Research Reproducibility. Ecancermedicalscience, 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 retrievalJuly 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

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