Research Metadata Schemas WG Case Statement

This group has been spun off from the RDA Data Discovery Paradigms IG.

WG charter

The widespread use of schema.org[1], DCAT[2] and other vocabularies in web pages to add structured metadata describing research data has brought new opportunities for making these outputs FAIRer. The opportunities include, but are not limited to:

1) Leveraging robust commercial search engines like Google, Yahoo, Bing etc. to facilitate broader discovery of, and access to, research data;
2) Providing a common set of vocabularies to describe research resources, enabling improved metadata interoperability across data repositories, increasing re-use and sharing of captured metadata;
3) Providing a potentially new method for metadata/content syndication among data catalogues and registries, enabling federated search across resources of a specific domain, or related domains relevant to a research need.

However, these opportunities also come with new challenges. Schema.org provides a core, minimalistic vocabulary for describing the kind of entities that most common web applications need. By design, schema.org expects, and has enabled, domains of practice to extend this core (Guha et al, 2016). Like other domains of practice, research data communities have their own needs for extending this core to describe research data and its relationships to other resources. These extensions include specific data types and the properties they possess, domain relevant and type specific to persistent identifiers, etc. There are some communities that are addressing these issues and have planned extensions to the core of schema.org to meet their own community needs, for example, bioschemas.org[3] for life science, and science-on-schema.org[4] for earth and environmental science. According to our recent survey[5] that was carried out by the Data Discovery Paradigms IG, more data repositories are following a similar route by either implementing structured markups in metadata landing pages, or else with planned extensions of schema.org in one way or another.

This proposed working group will provide a platform to complement, build on and extend efforts from bioschema.org, science-on-schema.org and the like communities in applying and extending the core schema.org vocabulary for describing research datasets and related resources (e.g. workflow, software, researchers, etc.). The objectives of this work group are twofold:

1. to identify and bridge gaps in existing schemas commonly used for research data, by bringing together communities who are working with such vocabularies to document research data and related resources;

---

[1] https://schema.org/
[2] DCAT: https://www.w3.org/TR/vocab-dcat-2/
[3] https://bioschemas.org/
[4] Science on schema: https://github.com/ESIPFed/science-on-schema.org/
[5] Survey summary: Current practices in using schemas to describe research datasets: http://bit.ly/2JZxXjK

2. to provide guidelines for those communities whose needs are not addressed by existing metadata schema such as schema.org, and provide guidelines on proposing extensions.

The planned outputs will include:

1. A generic 'conceptual data model' with essential types and properties for research data discovery over the web. The model will be built on bioschemas.org, science-on-schema.org, schema.org, DCAT, DDI-DISCO[6] and SSN[7] schemas from some representative research domains, and data discovery use cases. A research domain can map their schema to the conceptual model when they publish data to the web or exchange metadata between data portals/repositories.
2. A guideline, illustrated with common patterns, of common patterns for publishing metadata landing pages with structured data markups; and a guideline of how to customise the research schemas for target domains with examples.
3. Toolings for making the implementation easier if resources are available. This could include collecting and cataloguing tools that generate, validate and parse schema.org & DCAT markup, etc.

Value proposition

It is expected the proposed work will benefit a range of data stakeholders as follows:

Data providers and data catalogue managers:

- The conceptual data model and the guidelines will help data providers and data catalogue managers to implement structured metadata markups and have their data more findable by data seekers/consumers.
- Being able to adopt or map to a common research schema will make it possible for metadata from one catalogue to be more interoperable with, and reusable by, other data catalogues.

Data seekers:

- It will benefit data seekers/consumers for effective and efficient data search via faceted search and filtering, or other opportunities of either human or machine APIs built on combination of structured data search and keyword search.
- It will also make it easier for people to publish info about datasets and thus increase the range of datasets that are discoverable.
- Since all research data is expected to have some common properties, this work will make it possible to describe these common properties, and standard means through which they can be exposed for discovery.

Data technologies:

- When there is a common way to describe metadata across data catalogues, it provides opportunity for developing applications such as federated search either vertically to a

[6] https://www.ddialliance.org/Specification/RDF/Discovery
[7] https://www.w3.org/TR/vocab-ssn/

discipline or across different disciplines based on research needs, applications that can support a spectrum of data search needs from free text search to SPARQL queries.

Engagement with existing work in the area

The proposed work will be built on existing work of bioschema.org, science-on-schema.org, and a number of mappings to and from schema.org as identified from our recent survey[8] (e.g. DCAT to schema.org[9], DCAT-AP to schema.org[10], ISO 19115 to schema.org[11]). We will also reference data models and schemas from ISO and W3C recommended standards such as DCAT, which is currently being updated. Through exploration of existing work, we will identify common elements across research domains and domain specific elements as well.

The group will work with and encourage collaborations with other RDA WGs/IGs. In particular, we will examine closely the outputs from previous RDA WGs and IGs, including:

- Data Description Registry Interoperability (DDRI) WG
- Research Data Collections WG
- Research Data Repository Interoperability WG
- Metadata Interest Group (MIG)
- Preservation Tools, Techniques and Policies
- (more may be identified …)

The data models and types proposed from these WGs and IGs may directly benefit from the proposed work of this group, which is to extend core schema.org vocabulary to include more essential research resources for discovery and re-use of data.

The group will engage with existing RDA WGs/IGs for clear definition of data types and terms, including:

- Data Type Registries WG & #2
- Data in Context IG
- Domain Repositories IG
- (more may be identified …)

The group will seek to collaborate with domain specific RDA WGs/IGs, for example, International Materials Resource Registries WG and Marine Data Harmonisation IG, these groups could be potential adopters of recommendations from this group.

---

[8] Survey summary: Current practices in using schemas to describe research datasets: http://bit.ly/2JZxXjK

[9] https://w3c.github.io/dxwg/dcat/#dcat-sdo

[10] https://ec-jrc.github.io/dcat-ap-to-schema-org/

[11] https://www.w3.org/2015/spatial/wiki/ISO_19115_-_DCAT_-_Schema.org_mapping

We will also liaise with the Schema.org W3C Community Group[12] to recommend the proposed research schemas.

Work plan

Timing	Duration	Action	Main participants
Oct. 2019	0	RDA P14: Official start of the WG	Session participants
Oct. 2019 - March 2020		Identify common elements across research domains and domain specific elements based on existing work and the survey	Co-chairs and group members
		Draft guidelines for publishing metadata landing pages with structured data markups with the latest version of schema.org	
March/April 2020	6 months	RDA P15: Progress report and seeking feedback	Session participants
April - Oct. 2020		Propose data model and data types for research schemas	Co-chairs, technical advisory group, group members
		Extend guidelines for including research schemas	
		Identify existing toolings that help to map, compile and validate structured markups, collaborate and work with these groups to make tools work for the recommendations.	
Oct. 2020	12 months	RDA P16: Report recommendation draft, early adoption use case(s), seek more adopters	Session participants
Oct. 2020 - March 2021		Revise research schemas	Co-chairs, technical advisory group, group members
		Revise guidelines	
		List/catalogue of toolings	
March/April 2021	18 months	RDA P17: Recommendation output with adoption use cases	Co-chairs and group members

[12] https://www.w3.org/community/schemaorg/

Oct. 2021	24 months	RDA P18: More adoption use cases	Co-chairs and group members

Working group operation, communication and engagement

The group has set up a regular meeting every four week to communicate, discuss and get feedback from group members. Advance notice of each meeting will be sent to the group's mailing list, meeting notes and relevant documents will be made available from the group's wiki page at the RDA website.

We have a plan to organise group sessions at RDA plenaries, will be communicating and promoting the work to communities outside of RDA, and most importantly, we will take feedback and seek consensus to ensure the outputs are in line with community needs.

Apart from having four co-chairs, the group would also like to have a technical advisory group with members representing different standard bodies and disciplines from within and outside of RDA. Chairs and the advisory group will meet regularly (2~3 months) to review work in progress and resolve any technical and engagement issues as they arise.

We will have terms of reference for co-chairs, so each co-chair is aware her/his responsibility and RDA principles of openness and diversity inclusiveness. When there are disagreements and conflicts among co-chairs and group members, co-chairs will ensure different viewpoints are discussed and presented, and work with members and communities to achieve consensus.

Co-chairs of this WG have representatives from domains of life science and earth and environmental science, which have already adopted and extended schema.org. Their participation will ensure lessons learnt and outputs produced from the two communities will be carried over to the WG and the two communities will be consulted and engaged with the latest developments. Co-chairs also include a representative from a potential adopter – the Research Data Australia portal (run by Australian Research Data Commons (ARDC)). ARDC will not only likely adopt the outputs from the WG but also promote outputs to their Australia data providers and international partners such as Korea Institute of Science and Technology Information.

Adoption plan

ELIXIR and ESIP semantics technologies cluster have been working with life science community and earth and environmental science community on adoption of bioschema.org and science-on-schema.org respectively. The effort from the two communities on extension of the schema and guidelines and training on their respective adoption process have laid the foundation for this WG to work on. Having a representative from each of the two communities as co-chair of this WG shows the two communities will support, be engaged, and very likely adopt the conceptual model and guidelines from the WG.

Australian Research Data Commons (ARDC) runs a national data catalogue. The catalog (Research Data Australia) harvests metadata from 101 research organisations from around

Australia. ARDC is exploring how to improve global data discovery by providing optimised national aggregation point for syndication to global information systems (e.g. search engines, Scholix, and vertical discipline portals etc). It is likely that the outputs from the working group will be evaluated and adopted, as the outcome from the working group aligns with the direction ARDC is exploring.

The Arctic Data Committee (ADC) is an international body whose members come from data centers that hold polar data of any kind. Its purpose is to "promote and facilitate international collaboration towards the goal of free, ethically open, sustained and timely access to Arctic data through useful, usable, and interoperable systems". The ADC is comprised of members of the International Arctic Science Committee (IASC), the Sustaining Arctic Observing Networks program (SAON), and Standing Committee on Antarctic Data Management (SCADM). During a meeting in Geneva last fall, the ADC community unanimously agreed that adopting structured metadata àla schema.org was in the community's best interest. As a result, they are awaiting the results of this WG in order to guide development. Towards that end, they have appointed a liaison to this community, both to provide input and to take outputs back for implementation.

The WG chairs and members of technical advisory group will actively engagement communities in and out of RDA to promote the output and encourage more adoptions.

The working group chairs

Leyla Garcia (ELIXIR Hub, UK),
Sarala Dissanayake (DataCite, FREYA, Germany)
Adam Shepherd (Biological and Chemical Oceanography Office (BCO-DMO), US)
Mingfang Wu (Australian Research Data Commons, Australia)

Technical advisory group members

Simon Cox (CSIRO, Australia)
Ruth Duerr (Ronin Institute, US)
Doug Fils (Consortium for Ocean Leadership, US)
Rafael C. Jimenez (Research Informatics at Alzheimer's Research, UK)
Nick Juty (ELIXIR, UK)
Siri-Jodha Khalsa (National Snow and Ice Data Center, University of Colorado, US)
Andrea Perego (European Commission, Joint Research Centre (JRC))

Acknowledgement:

We would like to thank the following group members who have contributed to the writing up of the case statement:

Tim Clark - Massachusetts General Hospital / Harvard Medical School, US
Simon Cox - CSIRO, Australia
Anusuriya Devaraja - PANEAGA, Germany
Ruth Duerr - Ronin Institute for Independent Scholarship, US
Doug Fils - EarthCube Science Support Office, US
Leyla Garcia - ELIXIR Hub, UK

Nick Juty - ELIXIR-UK
Fidan Limani - German National Library of Economics, Germany
Stefanie Kethers - Australian Research Data Commons, Australia
Siri Jodha Khalsa - NSIDC, US
Andrea Perego - European Commission, Joint Research Centre, Italy
Adam Shepherd - Biological and Chemical Oceanography Office (BCO-DMO), US
Andrew Treloar - Australian Research Data Commons, Australia
Mingfang Wu - Australian Research Data Commons, Australia

Reference

Guha, R. V., Brickley, D., & Macbeth, S. (2016). Schema.org: Evolution of Structured Data on the Web. Communications of the ACM, 59(2), 44–51. doi:10.1145/2857274.2857276