Providing Research Graph Data in JSON-LD Using Schema.org

Jingbo Wang
National Computational Infrastructure
143 Ward Road
Acton ACT 2601
Australia
orcid: 0000-0002-3594-1893
Jingbo.Wang@anu.edu.au

Amir Aryani
Australian National University
900 Dandenong Road
Caulfield East, VIC 3145
Melbourne
Australia
orcid: 0000-0002-4259-9774
Amir.Aryani@anu.edu.au

Lesley Wyborn
National Computational Infrastructure
143 Ward Road
Acton ACT 2601
Australia
orcid: 0000-0001-5976-4943
Lesley.Wyborn@anu.edu.au

Ben Evans
National Computational Infrastructure
143 Ward Road
Acton ACT 2601
Australia
orcid: 0000-0002-6719-2671
Ben.Evans@anu.edu.au

ABSTRACT

In this position paper, we describe a pilot project that provides Research Graph records to external web services using JSON-LD. The Research Graph database contains a large-scale graph that links research datasets (i.e., data used to support research) to funding records (i.e. grants), publications and researcher records such as ORCID profiles. This database was derived from the work of the Research Data Alliance Working Group on Data Description Registry Interoperability (DDRI), and curated using the Research Data Switchboard open source software. By being available in Linked Data format, the Research Graph database is more accessible to third-party web services over the Internet, which thus opens the opportunity to connect to the rest of the world in the semantic format.

The primary purpose of this pilot project is to evaluate the feasibility of converting registry objects in Research Graph to JSON-LD by accessing widely used vocabularies published at Schema.org. In this paper, we provide examples of publications, datasets and grants from international research institutions such as CERN INSPIREHEP, National Computational Infrastructure (NCI) in Australia, and Australian Research Council (ARC). Furthermore, we show how these Research Graph records are made semantically available as Linked Data through using Schema.org. The mapping between Research Graph schema and Schema.org is available on GitHub repository. We also discuss the potential need for an extension to Schema.org vocabulary for scholarly communication.

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3038912.3038914

Keywords

Linked Data; JSON-LD; Schema.org; Semantic Web

1. INTRODUCTION

One of the key aspects of research and innovation is the ability to build on the success of prior research projects: hence, it is vital for research outcomes to be discoverable, reusable and trusted. Driven by the rapid development of new data storage technologies, the number of centralised data repositories is growing fast, and researchers, now more than ever, can have access to digital research outcomes stored in them. The problem is that these individual infrastructures often operate in silos and they cannot connect their datasets to related research information on other platforms.

One solution to this problem is the work undertaken by the Data Description Registry Interoperability Working Group of the international Research Data Alliance (RDA). The Working Group has developed the Research Data Switchboard [1] which can connect datasets and related information across research data repositories using information on co-authorship and jointly funded projects. This work was later extended to a distributed graph of scholarly works called Research Graph[1] that includes publications, datasets, grants and researchers' information (see Fig 1). The Research Graph currently connects publications, grants and datasets from a number of significant sized research national infrastructures including the Australian National Data Service (ANDS[2]), Australian National Computational Infras-

[1]http://researchgraph.org
[2]http://ands.org.au

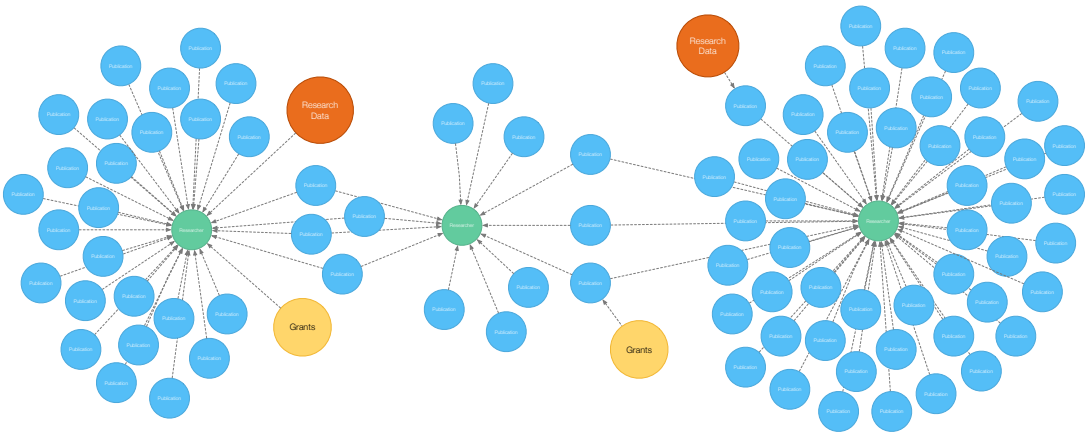

Figure 1: Connection made by Research Graph among grants (yellow), researchers (green), publication (blue) and datasets (red). The relationship among nodes is represented as grant(s) is awarded to researcher(s), and researcher(s) is the author of publication and/or dataset(s), and dataset(s) is referenced in the publication.

tructure (NCI[3]) [6], Dryad[4] (US), CERN InspireHEP[5] (Switzerland), figshare[6] (UK), da|ra and GESIS[7] (Germany), and OpenAIRE[8] (European Infrastructure), as well as ORCID[9], and DataCite[10] which are international.

In this project, we aim to make Research Graph accessible using JSON-LD[11]. JSON-LD is a lightweight Linked Data format. It is easy for humans to read and write, and it is based on the commonly accepted JSON format that provides a way to help make JSON data interoperate at web-scale. The presentation of Research Graph in JSON-LD format will enable third party services to traverse the graph using a JSON-based API. Linked Data will be the key enabler for the interoperability with the new services and it will support more effective discovery of associated research outcomes (articles, books, research datasets, etc.). The JSON-LD format using Schema.org is highly recommended in the data citation roadmap recently [3].

We hope that the BigScholar workshop will help us to find new collaborators in this domain and receive feedback from the community. As part of the workshop presentation, we will demonstrate the connections in the Research Graph and provide examples of the mapping of the data to JSON-LD using Schema.org. We will also discuss a possible extension to Schema.org, such as support for DOI and ORCID identifiers.

Also, we believe utilising Schema.org can provide a better context for some properties and fields in the Research Graph database and make it easier for other systems to digest the Graph records. The goal is to map most of the Research Graph schema using Schema.org vocabularies. For any elements that we could not find a perfect match, we hope to provide a recommendation to extend the Schema.org entities.

[3]http://nci.org.au
[4]http://datadryad.org
[5]https://home.cern
[6]https://figshare.com
[7]http://www.da-ra.de
[8]https://www.openaire.eu
[9]https://orcid.org
[10]https://www.datacite.org
[11]http://json-ld.org

2. SUPPORTING LINKED DATA

Research Graph records are currently available in XML format. The content of the XML values although readable by a human user using HTML and stylesheet transformation, may not necessarily be accessible and actionable by machines through web service interfaces.

The absence of control vocabularies is often the main reason for this lack of interoperability. Machines cannot easily comprehend the text-based content the same way as the human brain decodes and make sense of the free text. This problem motivates the implementation of infrastructures that support Linked Data as they enhance the semantic feature of the online data by transforming data into digital objects based on formal specifications and information models such as Schema.org. In this work, we aim to adopt this practice and provide a more interoperable graph using the Schema.org control vocabulary. By providing semantic enabled capacity of research graph, we hope to increase its discovery and search ability. It offers the potential to apply the intelligent natural language process approach in the future. The combination of web crawling search and machine decoding technology will make research graph available in google-type model, which will greatly increase the visibility by taking advantage of Internet search engine power.

The problem of ambiguity of names can be illustrated using Figure 3 where the metadata from Research Graph has been presented for an NCI dataset in Australia. Properties such as “source: National Computation Infrastructure” can be easily understood by a human user; however, a web-service would not be able to automatically ingest and decode this property and its associated value without a predefine meta model. Even for humans, some terms have ambigu-

ities meanings, unless a dedicated definition was attached to each term. This can be technically solved by assigning a URI to each term as a dereferencing link to denote the concept, such as "source" or "license".

The result is that we bring the linked Research Graph data into a network of standards-based machine interpretable data across different web Services. Given the advantage of being language independent as JSON format, JSON-LD was an obvious choice. We are also working on supporting other RDF-based formats, in particular, VIVO RDF [2] and RMAP [4] with the collaboration of international partners. Furthermore, we are exploring the possibility of integration with Microsoft Academic Graph [5]. As part of the workshop presentation, we will report on the progress of these interoperability projects.

3. RESEARCH GRAPH SCHEMA

Figure 2 describes the Research Graph meta model. This model comprise of four main registry objects also known as nodes: *publication*, *researcher*, *dataset* and *grant*. All these registry objects have three main mandatory properties: *source*, *local_id* and *key*.

- *source* is a namespace for the provider (or publisher) of the metadata (e.g. ORCID and NCI),
- *local_id* is an identifier for the registry object that is unique in the scope of the *source*, and
- *key* is a unique resolvable URL for the registry object that often leads to a landing page on the Research Graph website. The default format of the key is *researchgraph.org/{source}/{local_id}*

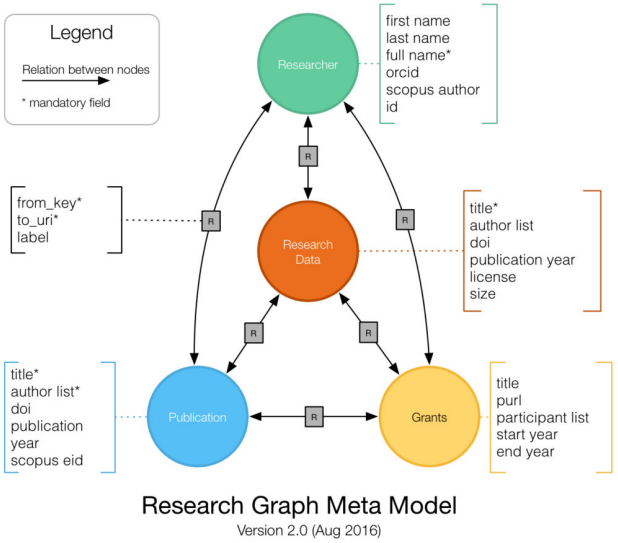

Figure 2: Research Graph Schema v2.0 released in Aug 2016.

These three properties enables the majority of connections in the graph. In addition, the registry objects have properties for ORCID, DOI, PURL and Scopus Identifiers (scopus_eid, scopus_author_id). These identifiers enable extra internal and external connections for registry objects.

The following examples shows the properties of three Research Graph registry objects: one dataset (Figures 3) record from National Computational Infrastructure (NCI), one publication (Figure 4) from CERN, and one grant (Figure 5) from Australian Research Council (ARC) . They are presented in JSON format using Research Graph schema.

```
{"type": "dataset",
 "key":
      "http://researchgraph.org/nci/f3525_9322_8600_7716/",
 "source": "National Computational Infrastructure",
 "local_id": "f3525_9322_8600_7716",
 "last_updated": "2014-12-31",
 "url":
      "http://pid.nci.org.au/dataset/f3525_9322_8600_7716",
 "title": "Coupled Model Intercomparison Project
      (CMIP5)",
 "authors_list": "Evans, Ben",
 "doi": "http://dx.doi.org/10.5072/29/5874605e6b57f",
 "datePublished": "2014-12-31",
 "license":
      "http://dapds00.nci.org.au/thredds/fileServer/
 licenses/license_ua6.txt",
 "megabyte": "1,500,000"
}
```

Figure 3: Example of a dataset registry object from National Computational Infrastructure (NCI) - Australia.

```
{"type": "publication",
 "key": "http://inspirehep.net/record/526765",
 "node_source": "CERN",
 "inspire_id": "526765",
 "title": "Stochastic background of gravitational
      waves",
 "authors_list": [ "Aguiar, O D",
                   "D'Ara\u00fajo, J C N",
                   "Miranda, O D"],
 "local_id": "astro-ph/0004395"
}
```

Figure 4: Example of a publication registry object from CERN InspireHEP.

```
{"type": "Grant",
 "key":
      "http://researchdata.ands.org.au/view/?key=http%3A%
 2F%2Fpurl.org%2Fau-research%2Fgrants%2Farc%2FDP0557498",
 "source": "ANDS",
 "local_id": "DP0557498",
 "title": "The End of the Dark Ages of the Universe",
 "purl": "purl.org/au-research/grants/arc/DP0557498",
 "funder": "Australian Research Council"
```

Figure 5: Example of a grant registry object from Australian Research Council (ARC)

4. MAPPING SCHEMA.ORG VOCABULARY TO SUPPORT JSON-LD

In this pilot project, we aim to use the popular Schema.org vocabulary as the dereferencing tool to support Research Graph schema's definition. In the context mapping, every term is mapped to IRI (Internationalized Resource Identifiers [RFC3987])) in the context so that it is unambiguously identified by an IRI and all values representing IRIs are explicitly marked by the keywords. IRIs are fundamental to Linked Data as that is how most nodes and properties are identified. Figure 6 presents a dataset example in the JSON-LD using the mapping in the context (see the beginning part of the file).

```
{"@context": {
          "@vocab":"http://schema.org/",
          "key": {
               "@id":"http://schema.org/mainEntityOfPage",
                  "@type": "@id"
                     },
          "source":"http://schema.org/sourceOrganization",
          "local_id": {
               "@id":"http://schema.org/disambiguating
                    Description",
                  "@type": "@id"
                       },
          "last_updated":
               "http://schema.org/dateModified",
          "title": "http://schema.org/headline",
          "authors_list": "http://schema.org/author",
             "doi": {
               "@id":"http://schema.org/sameAs",
               "@type": "@id"
               },
          "publication_year":
               "http://schema.org/datePublished",
          "megabyte": "http://ls
              .org/contentSize"
          },
 "@type": "Dataset",
 "key":
      "http://researchgraph.org/nci/f3525_9322_8600_7716/",
 "source": "National Computational Infrastructure",
 "local_id": "f3525_9322_8600_7716",
 "last_updated": "2014-12-31",
 "url":
      "http://pid.nci.org.au/dataset/f3525_9322_8600_7716",
 "title": "Coupled Model Intercomparison Project
      (CMIP5)",
 "authors_list": "Evans, Ben",
 "doi": "http://dx.doi.org/10.5072/29/5874605e6b57f",
 "datePublished": "2014-12-31",
 "license":
      "http://dapds00.nci.org.au/thredds/fileServer/licenses/
  license_ua6.txt",
 "megabyte": "1,500,000"
}
```

Figure 6: Dataset from NCI in JSON-LD format

The complete mapping between Research Graph schema and Schema.org for Research Graph registry objects (Dataset, Publication, Researcher and Grant) is available at GitHub Repository[12]. The summary is available in Table 1. The mapping serves as a term definition file and referenced within JSON-LD files. The context file for the Research Graph can be retrieved from GitHub Repository.

The corresponding one-on-one relationship is defined in the context section of the JSON-LD file. However, adding the mapping in front of each metadata file is tedious and may affect our existing workflow. JSON-LD provides an alternative way to save the mapping as a separate file so that it can be referenced at the beginning of JSON-LD. The content of the mapping can be modified without affecting the actual JSON-LD file itself. In the following modified JSON-LD file (see Figure 7), the context is referenced by adding a single line at the beginning of the JSON file to convert an existing JSON file to a JSON-LD document with minimum interruption to our existing data processing pipeline.

```
{"@context":
     "https://raw.githubusercontent.com/researchgraph/
  schema/master/json-ld/context.jsonld",
 "@type": "Dataset",
 "key":
      "http://researchgraph.org/nci/f3525_9322_8600_7716/",
 "source": "National Computational Infrastructure",
 "local_id": "f3525_9322_8600_7716",
 "last_updated": "2014-12-31",
 "url":
      "http://pid.nci.org.au/dataset/f3525_9322_8600_7716",
 "title": "Coupled Model Intercomparison Project
      (CMIP5)",
 "authors_list": "Evans, Ben",
 "doi": "http://dx.doi.org/10.5072/29/5874605e6b57f",
 "datePublished": "2014-12-31",
 "license":
      "http://dapds00.nci.org.au/thredds/fileServer/
  licenses/license_ua6.txt",
 "megabyte": "1,500,000"
}
```

Figure 7: Concise version of dataset from NCI in JSON-LD format.

5. EXTENSIONS TO SCHEMA.ORG

As part of the mapping exercise we have observed challenges in the following areas: We have managed to successfully map

- Publication ⇒ schema.org/ScholarlyArticle
- Researcher ⇒ schema.org/Person
- Dataset ⇒ schema.org/Dataset

However, the closest that we find to *grant* is schema.org/Action. This is not an exact match, and it is ambiguous. **Therefore we suggest adding a new type to Schema.org for the Research Grants or Research Projects.**

Furthermore, there is no precise method for including common identifiers – ORCID, DOI, Scopus ID(s) and PURL. These identifiers are the key enablers in linking scholarly communications, and it is essential to be able to capture and link them in different registry objects. In the current mapping we are using schema.org/sameAs for all of these identifiers. The implication is that in a large scale graphs with millions of nodes, searching a particular identifier can lead to a technical challenge. **Therefore we believe adding explicit properties for ORCID, DOI and other common**

[12]github.com/researchgraph/Schema/tree/master/json-ld

Table 1: Mapping between Research Graph and Schema.org

A. Mapping for Research Graph mandatory properties

Research Graph Schema	Schema.org Type	Property
key	Thing/CreativeWork/Article/SholarlyArticle Thing/Person, Thing/Action Thing/CreativeWork/Dataset Thing/Action	Schema.org/mainEntityOfPage
source	same as above	Schema.org/publisher for Publication and Dataset Schema.org/affiliation for Researcher and Grant
local_id	same as above	Schema.org/disambiguatingDescription
last_updated	same as above	Schema.org/dateModified

B. Mapping for Research Graph optional properties

Research Graph Schema	Schema.org Type	Property
Publication	Thing/CreativeWork/Article/SholarlyArticle	
title	same as above	Schema.org/headline
doi	same as above	Schema.org/sameAs
publication_year	same as above	Schema.org/datePublished
url	same as above	Schema.org/url
authors_list	Thing/Person	Schema.org/author
Researcher	same as above	
full_name	same as above	Schema.org/name
first_name	same as above	Schema.org/givenName
last_name	same as above	Schema.org/familyName
url	same as above	Schema.org/url
Dataset	Thing/CreativeWork/Dataset	
title	same as above	Schema.org/headline
doi	same as above	Schema.org/sameAs
publication_year	same as above	Schema.org/datePublished
url	same as above	Schema.org/url
license	Thing/CreativeWork	Schema.org/sameAs
megabyte	Thing/CreativeWork/MediaObject	Schema.org/contentSize
Grant	Thing/Action	
title	same as above	Schema.org/headline
Participant_list	same as above	Schema.org/agent
start_year	same as above	Schema.org/startTime
end_year	same as above	Schema.org/endTime
url	same as above	Schema.org/url
funder	Thing/Organization	Schema.org/funder

identifiers to Schema.org/{ScholarlyArticle, Dataset, Person} can extend the functionality of research infrastructures that leverage JSON-LD. A similar extension has already been proposed by BioSchemas community[13]. Their code is available on GitHub[14].

6. CONCLUSION AND FUTURE WORK

In this paper, we have presented a pilot project for adding JSON-LD support for Research Graph data. This project can enable an improved interoperability of connected Research Graph nodes including but not limited to publications from CERN, Dryad, datasets from figshare, da|ra, NCI, Research Data Australia and grants from Australian funders to third party services. We hope the new capability improves the discoverability and reusability of the Research Graph database.

Schema.org is a key enabler in transforming various XML files to JSON-LD. However, our preliminary work identifies a need for extending Schema.org to support widely used identifiers such as DOI, ORCID and PURL. As we are at the early stages of this project, we need feedback and direction from the community and collaboration in this domain, particularly from the service providers who have an interest in research metadata and enabling interoperability between research data infrastructures using JSON-LD. If you are interested in this project, please contact us.

It is promising to demonstrate the possibility and values of converting current Research Graph records into the JSON-LD format. We are getting one step closer to making Research Graph semantically accessible, searchable, and actionable across the web. We will develop an API to con-

[13]http://bioschemas.org/community/index.html
[14]https://github.com/BioSchemas

vert our existing database into the JSON-LD format if it is endorsed by the community to be a useful practice.

7. ACKNOWLEDGMENTS

We would also like to show our gratitude to the Martin Fenner (DataCite, Hannover, Germany - orcid.org/0000-0003-1419-2405) for sharing his insightful comments with us during this research work, and we thank the reviewers for their constructive comments to improve the quality of this work.

8. REFERENCES

[1] A. Aryani. Data description registry interoperability wg: Interlinking method and specification of cross-platform discovery. Technical report, Research Data Alliance, December 2016.

[2] K. Börner, M. Conlon, J. Corson-Rikert, and Y. Ding. Vivo: A semantic approach to scholarly networking and discovery. *Morgan-Claypool*, p.1-175, 2012.

[3] M. Fenner, M. Crosas, J. Grethe, D. Kennedy, H. Hermjakob, P. Rocca-Serra, R. Berjon, S. Karcher, M. Martone, and T. Clark. A data citation roadmap for scholarly data repositories. *Cold spring harbor laboratory*, 2016.

[4] K. Hanson, S. Morrissey, A. Birkland, T. Dilauro, and M. Donoghue. Using rmap to describe distributed works as linked data graphs: Outcomes and preservation implications. *13th International Conference on Digital Preservation, Bern, October 3-6, 2016*, 2016.

[5] A. Sinha, Z. Shen, Y. Song, H. Ma, D. Eide, B. Hsu, and K. Wang. An overview of microsoft academic service (mas) and applications. *In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 243-246*, 2015.

[6] J. Wang, A. Aryani, B. Evans, M. Barlow, and L. Wyborn. Graph connections made by rd-switchboard using nci's metadata. *D-Lib Magazine*, Volume 23(1/2), January/February 2017.