Analysis of Springer Nature SciGraph Linked Open Data (Postprint)

Authors: Bai Linlin, Zhu Zhongming, Bai Linlin

Date: 2018-01-10T00:00:00+00:00

Abstract

[Purpose/Significance] The analysis of linked open data provided by the Springer Nature SciGraph platform offers a reference for domestic publishers in leveraging linked data to facilitate linked open practices in academic exchange and semantic publishing, thereby propelling the advancement of China' s open research movement. [Method/Process] A detailed analysis is conducted of the entity objects, vocabularies, and data models published by the Springer Nature SciGraph platform. [Result/Conclusion] By constructing its own ontology, Springer Nature SciGraph employs triples in N-Triples format—a simpler serialization for RDF—to represent data. As one of the world' s largest publishers, Springer Nature' s linked data will undoubtedly provide valuable reference for other publishers aiming to achieve linked openness in research.

Full Text

Analysis of Springer Nature SciGraph Linked Open Data

Bai Linlin[1],[2], **Zhu Zhongming**[1]
[1]Lanzhou Library, Chinese Academy of Sciences, Lanzhou 730000
[2]University of Chinese Academy of Sciences, Beijing 100049

Abstract:
[Purpose/significance] This study analyzes the linked open data provided by the Springer Nature SciGraph platform to offer a reference for domestic publishers seeking to utilize linked data, thereby promoting the practice of linked open research in scholarly communication and semantic publishing, and advancing the open science movement. [Method/process] The paper conducts a detailed analysis of the entity objects, vocabularies, and data models employed by the Springer Nature SciGraph platform. [Result/conclusion] Springer Nature SciGraph represents data by constructing its own ontology and employing the simpler N-Triples serialization format for RDF triples. As one of the world'

s largest publishers, Springer Nature' s linked data implementation provides valuable references for other publishers aiming to realize linked research in the future.

Keywords: Springer Nature; SciGraph; linked open data; open research
Classification Number: G254
Funding: This work was supported by the Chinese Academy of Sciences Documentation and Information Capacity Building Project (Project No. Y6ZG421001).

---

1. Introduction

Linked data, as a best practice for the Semantic Web, has evolved significantly since its proposal in 2006, progressing from simple knowledge bases and vocabularies to diverse domain applications. At the 10th LDOW (Linked Data on the Web) workshop in 2017, L. Jens, A. Sören, C. Sarven, and colleagues summarized the development of linked data over the past decade, highlighting its potential to play a substantial role in scholarly communication in the coming decade [1]. The workshop also introduced the new initiative "pioneering the linked open research cloud" [2], encouraging scholars to apply linked data technologies and best practices to scholarly communication. C. Sarven et al. proposed the Linked Research Principles to promote the openness of linked scholarly knowledge [3]. ScholarlyData.org [4] organizes papers, people, organizations, and events related to Semantic Web conferences using the conference ontology data model [5]. The International Research Data Alliance (RDA), established in 2013 by the European Union, U.S. government, and Australian government [6], promotes research data sharing and exchange through data standards and practices, and has implemented functions connecting researchers, publications, research funding, and research datasets through the Research Data Switchboard and Research Graph projects. To help the research community fully leverage the benefits of open science, Springer Nature has launched the SciGraph linked open data platform [7], integrating data resources from Springer Nature and its partners, including information about research funding agencies, research projects and grants, conferences, research institutions, and publications. This platform makes analyzing information related to Springer Nature publications more accessible, currently containing 155 million triples about objects of academic interest. Additional data, such as citations, patents, clinical trials, and usage metrics, will be released in phases, increasing the number of triples in Springer Nature SciGraph to over 1 billion by the end of 2017 [8-9].

This study examines the linked open data provided by the Springer Nature SciGraph platform, analyzing its published entity objects, adopted ontologies and vocabularies, and data models. The aim is to provide a reference for domestic applications of linked data in scholarly communication and open research, promoting the semantic sharing and internationalization of domestic research data

and further advancing the open science movement.

2. Analysis of Entity Objects in Springer Nature SciGraph

Springer Nature' s linked open data categorizes entities into four main classes: agent, asset, concept, and event. Among these, entities in the concept and event classes are the focus of Springer Nature's publication. The concept class includes subclasses such as annotation, contract, publication, and type. The annotation subclass primarily comprises bibliometric categories, the contract subclass comprises funding categories, the publication subclass includes products (referring to articles, books, book chapters, and journals provided by Springer Nature) and works (monographs, serial publications), and the type subclass includes topics, access types, article types, conference series, product market numbers, and publication status subcategories. The event class includes subclasses such as affiliation, aggregation event, annotation event, conference, contributor, and publication event [8]. The thing class serves as the superclass for all classes.

Before publishing linked data, it is essential to identify the entity types and relationships within the data to be published, following the first principle of linked data publishing: using URIs as names for any resource to ensure resource accessibility. Springer Nature SciGraph employs two main URI patterns:

`http://www.springernature.com/scigraph/things/{datasets}/{scigraphId}`
`http://www.springernature.com/scigraph/things/{datasets}/{topic}`

These patterns are fundamentally similar, with the second pattern primarily used for topic-class entity objects. Both use `http://www.springernature.com/scigraph/` as the base address, which serves as the publishing platform for Springer Nature SciGraph linked open data. The "things" class acts as the superclass for all classes. The "datasets" component represents collections of various entities, such as articles, grants, journals, journalbrands, subjects, contributions, and books, and must be specified to access objects within them. Each object collection must be followed by a corresponding object, represented by "scigraphId" or "topic" in the URI; otherwise, the object cannot be located.

Currently, Springer Nature has published linked open data for entity types including articles, journals, subjects, and grants. The corresponding URIs for each data type are shown in Table 1 .

Table 1. URIs for Entity Objects - article: `http://www.springernature.com/scigraph/things/article`
- journal: `http://www.springernature.com/scigraph/things/journals/{scigraphId}`
- subject: `http://www.springernature.com/scigraph/things/subjects/{topic}`
- grant: `http://www.springernature.com/scigraph/things/grants/{scigraphId}`

Specific examples include: - Subject URI: `http://www.springernature.com/scigraph/things/subjects/ge`
(the subject term "geology"in Springer Nature) - Journal URI: `http://www.springernature.com/scigraph/th`

Springer Nature has established effective, unique URIs for data and created the SciGraph Core Ontology (prefix: sg), which consists of 45 classes and 206 proper-

ties with its own namespace (`http://www.springernature.com/scigraph/ontologies/core/`, prefix sg:). Conceptually, this ontology extends the previous nature.com core ontology [10]. This ontology was constructed for two reasons: first, because appropriate vocabulary could not be found in other ontologies or vocabularies to describe certain data or properties, and second, because classes and properties tailored to the model' s characteristics enable Springer Nature to better describe its data while facilitating external citation. The ontology also establishes links with external resources including the Global Research Identifier Database (GRID) [11], the Australian and New Zealand Standard Research Classification: Fields of Research (ANZSRC-FOR) [12], and DOI [13].

The Australian and New Zealand Standard Research Classification: Fields of Research (ANZSRC-FOR) classifies research and development activities according to the methods used in the R&D process, rather than by R&D units or purposes. The ANZSRC-FOR categories include major research fields, related subfields, and emerging areas investigated by enterprises, universities, colleges, national research institutions, and other organizations.

The Global Research Identifier Database (GRID) provides not only organization IDs and names but also metadata such as data types, hierarchical structures, and locations. It links with GeoNames, WikiData, CrossRef, the Open Funder Registry, the International Standard Name Identifier (ISNI), and other resources, enriching its metadata.

3. Vocabularies in Springer Nature SciGraph

The third principle of linked data publishing is to reuse existing, mature vocabularies to describe resources as much as possible to improve vocabulary interoperability and reduce local vocabulary management. The vocabularies used by Springer Nature SciGraph are divided into two categories: general vocabularies and specialized vocabularies (as shown in Table 2). General vocabularies primarily describe general properties of entities, such as entity types and relationships between entity types, while specialized vocabularies describe specific entities and possess properties characteristic of those entities. As evident from the table, Springer Nature SciGraph reuses general vocabularies to describe entity types, RDF data metadata (VoID [14]), and annotation vocabularies. The only specialized vocabularies used are SKOS and the self-built SciGraph Core Ontology; no other vocabularies are reused to describe resources. However, Springer Nature provides mappings between its ontology and other ontologies including bibo [15], crm (conceptual reference model) [16], dbpedia [17], dbpedia-owl, dc, dcterms, event, fabio (the FRBR-aligned bibliographic ontology) [18], foaf, mesh (medical subject headings) [19], obo (open biomedical ontologies) [20], prism (publishing requirements for industry standard metadata) [21], schema, skos, vcard [22], vivo (integrated semantic framework) [23], and wd (wikidata) [24]. Among these, mappings with dbpedia, mesh, and wd are for subject terms, while others are for classes and properties. This demonstrates that the SciGraph Core Ontology provides comprehensive descriptions of classes

and properties. Although it does not reuse existing mature vocabularies, using a self-built ontology based on specific requirements allows for more accurate description of relevant classes and properties.

Table 2. Vocabularies and Annotations - **OWL (Web Ontology Language)**: A language for publishing and sharing ontologies on the World Wide Web - **RDF (Resource Description Framework)**: A framework for describing Web resources in "subject-predicate-object" triple form - **RDFS (Resource Description Framework Schema)**: An extension vocabulary of RDF that provides data modeling vocabulary for RDF data - **Dcterms (DCMI Metadata Terms)**: An extension vocabulary of Dublin Core - **DC (Dublin Core)**: The Dublin Core vocabulary - **Vann (A Vocabulary for Annotating Vocabulary Descriptions)**: A vocabulary for annotating vocabularies with usage notes and examples - **VoID (Vocabulary of Interlinked Datasets)**: Used to describe metadata for RDF datasets - **SKOS (Simple Knowledge Organization System)**: A knowledge organization system vocabulary for describing controlled vocabulary terms - **sg (SciGraph Core Ontology)**: An ontology built by Springer Nature to describe resources provided on the Springer Nature website

4. Analysis of Springer Nature SciGraph Data Models

Currently, Springer Nature has published linked open data for entity types including articles, journals, subjects, and grants. The relational model among these data types is shown in Figure 1 [Figure 1: see original paper] [8]. Articles link to journals and subjects through `sg:hasJournal` and `sg:hasSubject`, respectively, while grant entities link to articles through `sg:hasFundedPublication`, as illustrated in Figure 2 [Figure 2: see original paper]. This paper disassembles this data model for individual analysis.

4.1 Article Data Model Springer Nature provides 20 properties for the article data model, which are categorized into six groups: type, identifier, label, contributor, publisher, subject information, and source. The `rdf:type` property indicates the type. Identifier properties include scigraphId, digital object identifier (DOI), and DOI link. Label properties include language, title, translated title, abstract, and translated abstract. Publisher information includes publication year, publication year-month, publication date, and Springer Nature webpage. Subject information includes the Australian and New Zealand Standard Research Classification: Fields of Research (ANZSRC-FOR) [11] and Springer Nature' s own subject terms. Source information includes journal, journal volume, and journal issue.

For property values, internal links are established with corresponding instance URIs provided by Springer Nature for contributors, funding data, journals, and subject terms. External links utilize DOI links provided by the Digital Object Identifier system and classification numbers provided by ANZSRC-FOR. Other property values are textual or numeric, as shown in Figure 2.

Springer Nature currently does not provide a data model for contributors. However, based on the published N-Triples [25] format triples, contributor properties include data type, scigraphId, publicly visible name, publicly visible surname, publicly visible given name, sort order (author order in the article), corresponding author status (boolean value true or false), role (values: "author" , "editor" , or "principal investigator"), and affiliation (entities provided by Springer Nature) [26].

4.2 Grant Data Model Springer Nature provides 18 properties for the grant data model, categorized into six groups: type, identifier, label, funding amount, funding period, funding body, and other. The `rdf:type` property indicates the type. Identifier properties include the grant' s scigraphId. Label properties include language, title, translated title, abstract, and translated abstract. Funding amount information includes grant amount and funding currency. Funding period includes start date and end date. Funding body includes funding organization and funded organization. Other properties include contributors related to the grant, research field classification numbers, funded publications, license terms, and webpage.

For property values, internal links are established with corresponding instance URIs for grant-related contributors and funded publication articles provided by Springer Nature. External links utilize URIs for funding and funded organizations provided by the Global Research Identifier Database (GRID) [10] and classification numbers provided by ANZSRC-FOR. Other property values are textual or numeric, as shown in Figure 3 [Figure 3: see original paper].

4.3 Journal Data Model Springer Nature provides 8 properties for the journal data model, categorized into six groups: type, identifier, journal brand, format, active publication status, and historical journal status. The `rdf:type` property indicates the type. Identifier properties include the journal' s scigraphId, ISSN, and DOI. The `sg:hasJournalBrand` property indicates the affiliated journal brand. Format refers to the medium of the journal. Active publication status is indicated by the `sg:isActivePublication` property, while historical journal status is indicated by `sg:isHistoricalJournal`.

For property values, internal links are established with corresponding instance URIs for journal brands and articles published in the journal provided by Springer Nature. Other property values are textual or numeric, where the journal medium property takes text values "Electronic"and "Paper (journals, normal index)" , and both "active publication" and "historical journal" properties take boolean values "true" or "false" , as shown in Figure 4 [Figure 4: see original paper].

Based on N-Triples format triples from Springer Nature' s journal data model, journal brand properties include data type, scigraphId, language, title, abbreviated title, subtitle, version statement, publisher, intellectual property owner,

webpage, date added to Springer Nature database, start year, end year, start volume, end volume, volume count, and open access (values: "Fully Open Access" or "Hybrid (Open Choice)").

4.4 Subject Data Model Springer Nature' s subjects are primarily self-established subject terms, divided into nine major categories: Biological Sciences, Earth and Environmental Sciences, Life Sciences, Physical Sciences, Scientific Community and Society, Social Sciences, Humanities, Business and Trade, and Deprecated subjects. Springer Nature provides 16 properties for the subject data model, categorized into five groups: type, label, identifier, reference, and SKOS representation. The `rdf:type` property indicates the type; `rdfs:label` indicates the label. Identifier properties include the subject' s ID. Reference properties include related relationships and alternative relationships between Springer Nature subject terms. SKOS representation organizes Springer Nature subject terms through the SKOS namespace, including preferred terms, non-preferred terms, definitions, notes, scope notes, subject thesaurus affiliation, hierarchical relationships, and top concepts.

For property values, internal links are established with corresponding instance URIs for related subject terms, alternative subject terms, and replaced subject terms provided by Springer Nature. SKOS is used to organize Springer Nature subject terms, enabling semantic representation. Other property values are textual, as shown in Figure 5 [Figure 5: see original paper].

5. RDF Implementation

The Springer Nature SciGraph linked open data platform currently only provides N-Triples format triples. N-Triples is a simpler serialization format for RDF, a line-oriented format where each triple must be written as a separate line consisting of a subject specifier, predicate specifier, and object specifier, followed by a period. If they have URIs, absolute URI references are enclosed in angle brackets [27]. Figure 6 [Figure 6: see original paper] shows an excerpt of N-Triples format code for the subject term "genomics" from Springer Nature SciGraph.

6. Licensing

Licensing clarifies legal issues in the process of publishing and using linked data, including ownership, publication rights, usage rights, and revenue rights. By defining different permissions for different users, it constructs a protection mechanism for the rational use of linked data to promote data openness, ensure open data security, and improve data reusability. Open licensing is a necessary condition for linked data to truly achieve openness and sustainable development in the Semantic Web environment. By embedding licenses in data, users can access data without seeking permission from data publishers. The license includes all permitted and non-permitted operations for users. Existing licensing

models include the GNU Free Documentation License, Common Documentation License, and Creative Commons License, among others, each capable of protecting different types of open data.

The purpose of Springer Nature SciGraph' s linked data publication is to integrate its research data into the linked data network and enable it to function in the public domain. Therefore, it selects the more general Creative Commons license as its licensing agreement. Data in Springer Nature SciGraph is obtained under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license model [28], which permits copying, distributing, and modifying the work in any medium and for any purpose, including creating derivative works, but prohibits commercial use [29]. The license is represented through the `dcterms:license` property, with the attribute value `https://creativecommons.org/licenses/by-nc/4.0/` indicating the terms of data use.

7. Conclusion

The launch of the Springer Nature SciGraph linked open data platform marks the starting point for achieving linked open research in the publishing domain. The implementation of Springer Nature SciGraph linked open data breaks the original data organizational structure, enabling data linking, interoperability, and data mining functions. It describes publications from a conceptual perspective. Through analysis of the Springer Nature SciGraph linked open data model, it is evident that publishers, as the source of published works, play a significant role in achieving data interoperability and linking through semantic description of publications. China' s publishing domain should learn from this model of achieving linked open research publications. By thoroughly analyzing the entities, properties, and relationships contained in Chinese publications, publishers should select appropriate ontologies or, when facing unique Chinese document types (ancient books, rubbings, etc.), construct their own ontologies to build data models for semantic description and semantic publishing of Chinese publications. Simultaneously, licensing should be designated to clarify legal issues in data usage. Of course, implementation requires software platforms. Considering cost issues and interoperability between different publishers, open-source software should be considered for different disciplines, and publishers within the same discipline should adopt unified data models and unified vocabularies to achieve linked open publications.

Author Contributions: Bai Linlin was responsible for data acquisition, outline development, and manuscript writing; Zhu Zhongming was responsible for manuscript revision.

References

[1] LEHMANN J, AUER S, CAPADISLI S, et al. LDOW2017: 10th workshop on linked data on the Web [EB/OL]. [2017-05-25]. http://events.linkeddata.org/ldow2017/ldow-

10th-workshop.pdf.

[2] From ScholarlyData.org to pioneering the linked open research cloud [EB/OL]. [2017-05-25]. http://aims.fao.org/ar/activity/blog/scholarlydataorg-pioneering-linked-open-research-cloud.

[3] Linked Research [EB/OL]. [2017-05-25]. https://linkedresearch.org/.

[4] Welcome to scholarlydata.org [EB/OL]. [2017-05-25]. http://www.scholarlydata.org/.

[5] The conference ontology [EB/OL]. [2017-05-25]. http://www.scholarlydata.org/ontology/doc/.

[6] About RDA [EB/OL]. [2017-05-25]. https://www.rd-alliance.org/about-rda.

[7] Springer Nature SciGraph: Supporting open science and the wider understanding of research [EB/OL]. [2017-05-25]. http://www.springernature.com/cn/group/media/press-releases/springer-nature-scigraph-supporting-open-science-and-the-wider-understanding-of-research/12129614.

[8] SciGraph Dataset Downloads [EB/OL]. [2017-05-26]. https://github.com/springernature/scigraph/wiki#get started.

[9] 支持开放科研，施普林格·自然推出关联数据平台 [EB/OL]. [2017-05-26]. http://www.toutiao.com/a6397639332211818754/.

[10] Nature.com Ontologies [EB/OL]. [2017-05-27]. https://www.nature.com/ontologies/models/core/.

[11] GRID - Global Research Identifier Database [EB/OL]. [2017-05-27]. https://www.grid.ac/downloads.

[12] Australian and New Zealand standard research classification: fields of research [EB/OL]. [2017-05-28]. https://vocabs.ands.org.au/anzsrc-for.

[13] Digital object identifier system [EB/OL]. [2017-05-29]. http://www.doi.org/index.html.

[14] Vocabulary of interlinked datasets (VoID) [EB/OL]. [2017-05-30]. http://vocab.deri.ie/void#.

[15] Bibliographic ontology [EB/OL]. [2017-05-26]. http://purl.org/ontology/bibo.

[16] Definition of the CIDOC conceptual reference model [EB/OL]. [2017-05-30]. http://www.cidoc-crm.org/html/5.0.4/cidoc-crm.html.

[17] DBpedia [EB/OL]. [2017-05-31]. http://wiki.dbpedia.org/.

[18] FaBiO, the FRBR-aligned bibliographic ontology [EB/OL]. [2017-05-31]. http://www.sparontologies.net/ontologies/fabio/source.html.

[19] MeSH linked data (beta) [EB/OL]. [2017-05-31]. https://id.nlm.nih.gov/mesh/.

[20] Open biomedical ontologies [EB/OL]. [2017-05-31]. https://en.wikipedia.org/wiki/Open_Biomedical_Onto

[21] PRISM metadata [EB/OL]. [2017-06-01]. https://www.idealliance.org/prism-metadata/.

[22] vCard ontology - for describing people and organizations [EB/OL]. [2017-06-02]. https://www.w3.org/TR/vcard-rdf/.

[23] VIVO-Integrated semantic framework [EB/OL]. [2017-06-02]. http://bioportal.bioontology.org/ontologies/ISF?p=classes&conceptid=root.

[24] Entity data [EB/OL]. [2017-06-02]. https://www.wikidata.org/wiki/Special:EntityData/.

[25] N-Triples [EB/OL]. [2017-06-02]. https://en.wikipedia.org/wiki/N-Triples.

[26] scigraph/articles.ttl at master [EB/OL]. [2017-06-02]. https://github.com/springernature/scigraph/blob/m

[27] RDF 1.1 N-Triples [EB/OL]. [2017-06-04]. https://www.w3.org/TR/n-triples/.

[28] Creative commons —attribution-NonCommercial 4.0 International —CC BY-NC 4.0 [EB/OL]. [2017-06-04]. https://creativecommons.org/licenses/by-nc/4.0/.

[29] Creative commons —署名-非商业性使用 4.0 国际—CC BY-NC 4.0 [EB/OL]. [2017-06-04]. https://creativecommons.org/licenses/by-nc/4.0/deed.zh.

Note: Figure translations are in progress. See original paper for figures.