Springer Nature SciGraph Linked Open Data Postprint Analysis

Authors: Bai Linlin, Zhu Zhongming

Date: 2023-10-08T00:00:00+00:00

Abstract

[Purpose / Significance] An analysis of the linked open data provided by the Springer Nature SciGraph platform offers a reference for domestic publishers in utilizing linked data to facilitate linked open practices in scientific research within academic exchange and semantic publishing, thereby advancing the open research movement in China. [Method / Process] This study conducts a detailed analysis of the entity objects, adopted vocabularies, and data models published by the Springer Nature SciGraph platform. [Result / Conclusion] Springer Nature SciGraph represents data using triples in the N-Triples format, a simpler serialization for RDF, by constructing its own ontology. As one of the world's largest publishers, Springer Nature's linked data will undoubtedly provide reference for other publishers aiming to implement linked open data in scientific research.

Full Text

Preamble

ChinaXiv Collaborative Journal [Academic Exploration] 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**] The analysis of the linked open data provided by the Springer Nature SciGraph platform offers a valuable reference for domestic publishers seeking to leverage linked data practices in scholarly communication and semantic publishing. This can help advance the open research movement in China by promoting the adoption of linked open research practices.

[**Method/Process**] This 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 utilizing N-Triples, a simplified serialization format for RDF triples. As one of the world's largest publishers, Springer Nature's linked data implementation will undoubtedly provide important lessons for other publishers aiming to achieve linked open research.

Keywords: Springer Nature; SciGraph; linked open data; open research

1 Introduction

Linked data, as a best practice for the Semantic Web, has evolved from simple knowledge bases and vocabularies to applications across various domains. At the 10th LDOW (Linked Data on the Web) conference in 2017, Jens Lehmann, Sören Auer, Sarven Capadisli, and others summarized the development of linked data over the past decade and highlighted its growing role in scholarly communication in the coming decade [1]. The conference also introduced a new initiative, "Pioneering the Linked Open Research Cloud" [2], encouraging scholars to apply linked data technologies and best practices to academic exchange.

Capadisli et al. proposed the Linked Research Principles to encourage the openness of linked research knowledge [3]. ScholarlyData.org [4] organizes information related to Semantic Web conferences—including papers, people, organizations, and events—using the Conference Ontology data model [5]. In 2013, the European Union, U.S. government, and Australian government jointly established the Research Data Alliance (RDA) [6] to promote the sharing and exchange of research data through data standards and best practices. The Research Data Switchboard and Research Graph projects have enabled the connection of researchers, publications, research funding, and research datasets.

To help the research community fully capitalize on the benefits of open science, Springer Nature launched the SciGraph linked open data platform [7], which integrates 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 it easier to analyze Springer Nature publication information and currently contains 155 million triples about academic entities. Additional data, including citations, patents, clinical trials, and usage statistics, will be released in phases, with the total number of triples in Springer Nature SciGraph expected to exceed 1 billion by the end of 2017 [8-9].

This paper examines the linked open data provided by the Springer Nature SciGraph platform, analyzing its published entity objects, ontologies, vocabularies, and data models. The aim is to provide a reference for the use of linked data in domestic scholarly communication and open research, thereby promoting the semanticization, international sharing, and further development of China's open research movement.

2 Springer Nature SciGraph Entities

Springer Nature's linked open data is organized into four main categories: agent, asset, concept, and event, with concept and event entities being the primary focus of the platform's publications. The concept category includes annotations, contracts, publications, and types at the next level. The annotation subclass primarily comprises citation metrics, the contract subclass focuses on funding, and the publication subclass includes products (referring to articles, book chapters, journals, and monographs provided by Springer Nature), types (including themes, access types, article types, conference series, product market numbers, and publication status subclasses). The event category includes affiliations, aggregated events, annotation events, conferences, contributors, and publication events at the next level. 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. Following the first principle of linked data publishing—using URIs as names for resources to ensure their accessibility—Springer Nature SciGraph employs two main URI patterns:

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

These patterns are functionally identical, with the second pattern primarily used for subject entities. Both use `http://www.springernature.com/scigraph/` as the base address, which serves as the publication platform for Springer Nature SciGraph linked open data. The "thing" class is the superclass for all classes, "datasets" represents collections of various entity types (with possible values including articles, grants, journals, journal-brands, subjects, contributions, books, etc.), and an object must be specified using either "scigraphId" or "topic" to access it.

Currently, the entity types published in Springer Nature SciGraph include articles, journals, subjects, and grants, with their corresponding URIs shown in .

Entity Type	URI Pattern
article	`http://www.springernature.com/scigraph/things/articles/{sc`
journal	`http://www.springernature.com/scigraph/things/journals/{sc`
subject	`http://www.springernature.com/scigraph/things/subjects/{to`
grant	`http://www.springernature.com/scigraph/things/grants/{scig`

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

Springer Nature has not only established effective, unique URIs but also created the SciGraph Core Ontology (prefix: sg), which consists of 45 classes and 206 properties with its own namespace (`http://www.springernature.com/scigraph/ontologies/core/`). Conceptually, this ontology builds upon previous work. The reasons for developing this ontology are twofold: first, existing ontologies or vocabularies lacked appropriate terms to describe certain data or attributes; second, classes and properties tailored to Springer Nature's model enable more accurate data description while facilitating external citation. The ontology also establishes links with external resources such as GeoNames, WikiData, CrossRef, the Open Funder Registry, the International Standard Name Identifier (ISNI), the Global Research Identifier Database (GRID) [11], the Australian and New Zealand Standard Research Classification: Fields of Research (ANZSRC-FOR) [12], and DOI [13] to enrich metadata.

3 Springer Nature SciGraph Vocabulary

The third principle of linked data publishing is to reuse existing, mature vocabularies to describe resources whenever possible, thereby improving vocabulary interoperability and reducing the burden of local vocabulary management. The vocabularies used by Springer Nature SciGraph fall into two categories: general-purpose vocabularies and specialized vocabularies (see). General-purpose vocabularies describe general entity properties, such as entity types and relationships between entity types, while specialized vocabularies describe specific entities and their particular attributes.

Vocabulary	Description
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 using subject-predicate-object triples
RDFS (Resource Description Framework Schema)	An extension vocabulary for RDF that provides data modeling capabilities
Dcterms (DCMI Metadata Terms)	An extended vocabulary of Dublin Core terms
DC (Dublin Core)	The Dublin Core vocabulary
Vann (A Vocabulary for Annotating Vocabulary Descriptions)	A vocabulary for annotating vocabularies with usage examples and documentation
VoID (Vocabulary of Interlinked Datasets)	A vocabulary for describing metadata about RDF datasets
SKOS (Simple Knowledge Organization System)	A vocabulary for representing controlled vocabularies and knowledge organization systems

Vocabulary	Description
sg (SciGraph Core Ontology)	Springer Nature's custom ontology for describing resources on its platform

As shown in the table, Springer Nature SciGraph reuses general vocabularies to describe entity types, RDF dataset metadata (VoID [14]), and annotation terms. The only specialized vocabularies used are SKOS and the custom SciGraph Core Ontology. Although it does not reuse other mature vocabularies for resource description, Springer Nature provides mappings between its ontology and numerous 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.org, SKOS, vCard [22], VIVO (Integrated Semantic Framework) [23], and Wikidata [24]. Notably, mappings to DBpedia, MeSH, and Wikidata are at the class and property level, while others are at the term level. This demonstrates that the SciGraph Core Ontology provides comprehensive class and property descriptions. Although it does not reuse existing mature vocabularies, using a custom ontology tailored to specific requirements allows for more precise description of relevant classes and properties.

4 Springer Nature SciGraph Data Model

Currently, Springer Nature's published linked open data includes four main entity types: articles, journals, subjects, and grants. The relationship model among these data types is shown in [Figure 1: see original paper] [8]. Articles link to journals and subjects via `sg:hasJournal` and `sg:hasSubject`, respectively, while grant entities link to articles through `sg:hasFundedPublication`, as illustrated in [Figure 2: see original paper]. This section breaks down and analyzes these data models individually.

4.1 Article Data Model

Springer Nature provides 20 properties for the article data model, which can be categorized into six groups: type, identifiers, labels, contributors, publisher information, and subject information. The `rdf:type` property indicates the entity 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 month, publication date, and Springer Nature webpage. Subject information includes both the Australian and New Zealand Standard Research Classification: Fields of Research (ANZSRC-FOR) [11] codes and Springer Nature's own subject terms. Source information includes journal, volume, and issue.

In terms of property values, internal links are established to corresponding instance URIs for contributors, funding data, journals, and subject terms provided by Springer Nature. External links utilize the DOI system's DOI link and ANZSRC-FOR classification codes. Other properties take text or numeric values, as shown in [Figure 2: see original paper].

Springer Nature has not yet published a formal data model for contributors. However, based on the published N-Triples [25] format data, contributor properties include data type, `scigraphId`, published name (given name, family name, and full name), order (author order), corresponding author status (boolean true/false), role (with values "author", "editor", or "principal investigator"), and affiliation (as a Springer Nature-provided entity) [26].

4.2 Grant Data Model

Springer Nature provides 18 properties for the grant data model, organized into six categories: type, identifiers, labels, funding amount, funding period, funding organizations, and other properties. 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 funding amount and currency. Funding period includes start year and end year. Funding organization information includes funding organization and recipient organization. Other properties include contributors related to the grant, field of research classification codes, funded publications, license terms, and webpage.

For property values, internal links are made to instance URIs for grant-related contributors and funded publications provided by Springer Nature. External links use URIs from the Global Research Identifier Database (GRID) [10] for funding and recipient organizations, and classification codes from ANZSRC-FOR. Other properties take text or numeric values, as shown in [Figure 3: see original paper].

4.3 Journal Data Model

Springer Nature provides eight properties for the journal data model, divided into six categories: type, identifiers, journal brand, format, active publication status, and historical status. The `rdf:type` property indicates the type. Identifier properties include the journal's `scigraphId`, ISSN, and DOI. The `sg:hasJournalBrand` property indicates the journal brand. Format refers to the medium of the journal. The `sg:isActivePublication` property indicates whether the journal is currently being published, while `sg:isHistoricalJournal` indicates whether it is a historical journal.

For property values, internal links are established to instance URIs for the journal brand and articles published in the journal. Other properties take text or numeric values, where the medium property has text values of "Electronic" or

"Paper (journals, normal index)", and both "is active publication" and "is historical journal" properties take boolean values of "true" or "false", as shown in [Figure 4: see original paper].

Based on the N-Triples format data, journal brand properties include data type, `scigraphId`, language, title, short title, subtitle, imprint, publisher, copyright holder, webpage, date added to Springer Nature database, start year, end year, starting volume, ending volume, volume count, and open access status (with values "Fully Open Access" or "Hybrid (Open Choice)").

4.4 Subject Data Model

Springer Nature's subjects are primarily self-established subject terms, organized into nine main categories: Biological Sciences, Earth and Environmental Sciences, Life Sciences, Physical Sciences, Scientific Community and Society, Social Sciences, Humanities, Business and Trade, and Deprecated terms. The subject data model provides 16 properties, categorized into five groups: type, labels, identifiers, references, and SKOS representation. The `rdf:type` property indicates the type, `rdfs:label` provides the label, and identifier properties include the subject ID. Reference properties include related subjects, replaced subjects, and replacing subjects. SKOS representation organizes Springer Nature subject terms through the SKOS namespace, including preferred labels, alternative labels, definitions, notes, scope notes, subject schemes, hierarchical relationships, and top concepts.

For property values, internal links are made to instance URIs for related, replaced, and replacing subject terms. SKOS is used to semantically represent Springer Nature subject terms, while other properties take text values, as shown in [Figure 5: see original paper].

5 RDF Implementation

The Springer Nature SciGraph linked open data platform currently provides data only in N-Triples format. N-Triples is a simplified, line-based serialization format for RDF where each triple must be written as an independent line consisting of a subject specifier, predicate specifier, and object specifier, followed by a period. If resources have URIs, they are enclosed in angle brackets as absolute URI references [27]. [Figure 6: see original paper] shows a sample of N-Triples 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 establishes a protection mechanism for the rational use of linked data, thereby promoting data openness, ensuring data security, and improving 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 publishers. The license includes all permitted and prohibited operations.

Existing licensing models include the GNU Free Documentation License, Common Documentation License, and Creative Commons License, each providing protection for different types of open data. Springer Nature SciGraph publishes linked data to integrate its research data into the linked data network and enable it to function in the public domain. Therefore, it has chosen the widely applicable Creative Commons license.

Data in Springer Nature SciGraph is available under the Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license model [28], which allows copying, distribution, and modification of the work in any medium and format, and permits adaptation and derivative works, but prohibits commercial use [29]. The license is indicated through the `dcterms:license` property, with the value `https://creativecommons.org/licenses/by-nc/4.0/` specifying the terms of use.

7 Conclusion

The launch of the Springer Nature SciGraph linked open data platform marks a starting point for achieving linked open research in the publishing domain. The implementation of Springer Nature SciGraph breaks down traditional data organizational structures, enabling data linking, interoperability, and mining. By describing publications from a conceptual perspective, the platform facilitates data interoperability and association. Analysis of the Springer Nature SciGraph data model reveals that publishers, as the source of publications, play a crucial role in semantic description for achieving data interoperability and linking.

Chinese publishers should learn from this model of implementing linked open research publications. By thoroughly analyzing the entities, attributes, and relationships contained in Chinese publications, they can select appropriate ontologies or, when dealing with unique Chinese materials (such as ancient texts or rubbings), construct their own ontologies to build data models that enable semantic description and semantic publishing of Chinese publications. Simultaneously, they should specify licensing terms 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, and publishers within the same discipline should adopt unified data models and vocabularies to achieve linked open publications.

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/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] Supporting open research: Springer Nature launches linked data platform [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}}

[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 — Attribution-NonCommercial 4.0 International — CC BY-NC 4.0 [EB/OL]. [2017-06-04]. https://creativecommons.org/licenses/by-nc/4.0/deed.zh.

Author Contributions

Bai Linlin: Responsible for data collection, outline development, and manuscript writing.
Zhu Zhongming: Responsible for manuscript revision.

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