Japan Search RDF Schema: a dual-layered approach to describe items from heterogeneous data sources

Daichi Machiya	Tomoko Okuda	Masahide Kanzaki
National Diet Library, Japan	National Diet Library, Japan	Xenon Limited Partners, Japan
d-machiy@ndl.go.jp	standardization@ndl.go.jp	mkanzaki@gmail.com

Abstract

The National Diet Library, Japan (NDL), with support by Xenon Limited Partners, has designed a new metadata schema based on the RDF model while developing a national platform for metadata aggregation and sharing, "Japan Search". Japan Search collects metadata from libraries, museums, archives, and research institutions across the country, and provides an integrated search service as well as APIs (SPARQL Endpoint and REST-API). The aim of this paper is to introduce the new schema, highlighting its dual-layered data model and the normalization of temporal (When), spatial (Where), and agential (Who) information provided in the source data.

Keywords: National Diet Library; national library; Japanese library; Linked Open Data; metadata aggregation; metadata model; digital cultural heritage

1. Introduction

The NDL has been providing an integrated search service of the materials held by libraries, universities, archives, and museums through NDL Search since 2012. It receives, stores, and outputs metadata based on a Dublin Core application profile, DC-NDL. Since 2015, the NDL has been working closely with related parties including the Intellectual Property Strategy Promotion Bureau in promoting digitization and online access of cultural and scholarly contents held by Japanese institutions. This led to the building of a new cross-sector integrated portal for cultural and scholarly contents. The new portal "Japan Search" made its debut on February 27, 2019.

Japan Search is not merely an integrated search service; it has also implemented several features to motivate potential partner institutions. Among them is the RDF store, where the aggregated metadata are converted to a standard schema based RDF model. The new schema is called "Japan Search RDF Schema", or "JPS RDF" for short. The schema primarily uses schema.org vocabulary, which is convenient for a wide range of users and applications, along with some additional properties defined at Japan Search.

As of April 2019, Japan Search presents RDF descriptions of 16,651,685 items with over 700 million statements.

2. JPS RDF Schema

2.1. Data Model

JPS RDF adopts a data model comprising of two parts: (1) content description with normalized values, and (2) information on access to the content and on the source database from which metadata is submitted to Japan Search. The content description subsequently consists of two layers: a simple description of the content using well-known schema.org properties and corresponding structured descriptions using properties defined in the Japan Search system.

The rationale behind this data model is to enable users to effectively find content matching their interests across heterogeneous domains, while preserving metadata provenance and acknowledgement to the provider of the content. The dual-layered content description part achieves

DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2019

the former through a combination of a simple triple pattern query and a complex refined search. The access and source information invites the users to the provider of the content, and informs the users whether the provider is a holder of the original object or a database publisher of the digitized surrogates. The source data submitted by partner institutions (hereinafter referred to as "source data") will be kept in a JSON format without modifying properties or values. It is then given a URI, and is linked from the description in JPS RDF, in order for users to access the original catalog information.

For the simple description, schema.org properties were chosen after careful comparison with the terms in other major vocabularies such as DCMI Metadata Terms and BIBFRAME. The project preferred the domain-independent nature and extensibility of the schema.org vocabulary. While the structured description could be expressed in several ways, the project decided to introduce its own vocabulary in order to make maximum use of this structure.

2.2. Simple schema.org descriptions and value normalization

When the source data of a certain piece of content is submitted to Japan Search, values that represent time (When) and place (Where) are normalized, so that the data submitted by various institutions can be retrieved in a unified form. In the JPS RDF, a year is the smallest unit of Time (i.e. temporal information), while the prefecture is the smallest unit for Place (i.e. spatial information). Names of people and organizations are also normalized by matching against a set of named entity dictionaries prepared for the mapping systems. The normalized data are given an IRI (International Resource Identifier), and linked to LOD hubs, such as entity URIs in acknowledged authority files, VIAF, and Wikidata (Fig. 1).

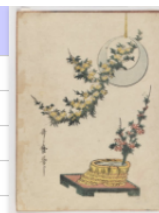| Description of <https://jpsearch.go.jp/entity/chname/喜多川歌麿>	
rdf:type	type:Agent
rdfs:label	"喜多川歌麿"
schema:name	"Kitagawa, Utamaro"@en
	"きたがわ うたまろ"@ja-kana
	"喜多川歌麿"@ja
	"歌麿"@ja
schema:description	"1753?-1806, 江戸時代中期の浮世絵師;「歌麿」はここに含めた"
schema:hasOccupation	<http://ja.dbpedia.org/resource/浮世絵師>
schema:image	<http://commons.wikimedia.org/wiki/Specia ... Kitagawa_Utamaro.jpg>
owl:sameAs	<http://collection.britishmuseum.org/id/person-institution/7171>
	<http://data.bnf.fr/ark:/12148/cb11909802x#about>
	<http://dbpedia.org/resource/Utamaro>
	<http://id.ndl.go.jp/auth/entity/00270473>
	<http://ja.dbpedia.org/resource/喜多川歌麿>
	<http://lod.ac/id/1614>
	<http://viaf.org/viaf/95261927>
	<http://www.wikidata.org/entity/Q272045>

Fig. 1. Normalization of Name "喜多川歌麿 (Kitagawa, Utamaro）"

These normalized values of time, place, and name are described with schema.org properties; e.g. `schema:temporal`, `schema:spatial`, and `schema:contributor` (or `schema:creator`), without distinguishing the relationship of the year, the place, or the name to the content. Any indication of time contained in the source data can be accommodated by the property `schema:temporal`, whether it is the publication date of a book or the excavation date of an

archeological artifact. Similarly, any indication of a place will be accommodated by `schema:spatial`, whether it is the place for collecting a specimen or the place illustrated in a woodblock print. This makes it easy for users to discover the content related to a particular place, period of time, or creator/contributor by a simple triple pattern query.

Although those normalized IRIs have Japanese names as their local parts, they have English labels as well. English-speaking users can take advantage of the normalized values via these labels (`schema:name` with language tag "en") or the above-mentioned LOD links (`owl:sameAs`) (Fig. 2).

```
PREFIX schema: <http://schema.org/>
SELECT ?cho WHERE {
   ?cho schema:spatial/schema:name "Mie"@en .
}
```

Fig. 2. An example of a simple SPARQL query requesting the contents related to Mie prefecture

2.3. Structured description with JPS RDF properties

While simple description using schema.org properties allows casting a wide net, some may want a refined search result using complex conditions. JPS RDF defines properties such as `jps:spatial`, `jps:temporal`, and `jps:agential` (`jps:` is the prefix for the namespace for the properties defined at Japan Search) to express structured description. The structured description complements the corresponding simple description with a property `jps:relationType` that specifies the relationship of the time/place/name to the content it is linked to. The structured description contains the same value as the corresponding simple description in `jps:value`, which designates the representative (normalized) value of this construct. By using the relationship type associated within the `jps:spatial` structured node, complex queries such as "ceramics created in Mie Prefecture" (see the structure in Fig. 3) or "specimens collected in Kanagawa prefecture in the 1930s" can be composed.
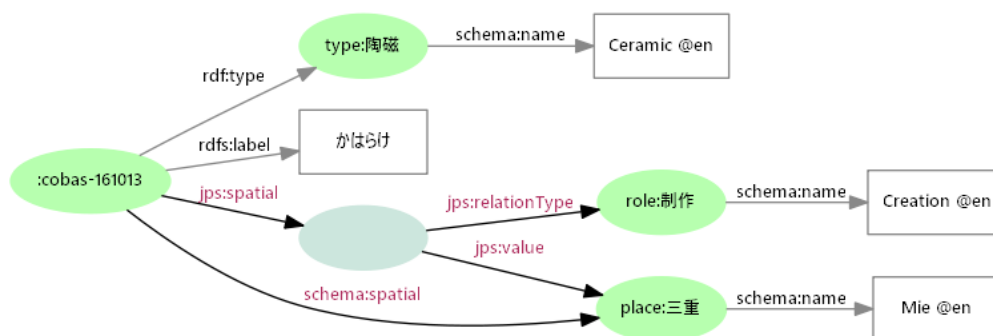

Fig. 3. Structured description of Place using `jps:relationType`

The structured property nodes preserve original values provided by the content provider, so that data consumers can obtain rich descriptions missed during normalization. By adding contextual information, e.g. role (`jps:relationType`), and original values to the relation, the structured properties of JPS RDF are meant to be a sort of "property graph" or annotation of properties. While RDF has no built-in mechanism to annotate a property, the JPS RDF's dual-layered model makes it possible to annotate the relation with detailed information while retaining a simple style of property to facilitate easy and straightforward queries.

✺DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2019

2.4. Access Information and Source Information

Information on access to the content is provided in a structured form as an instance of the "Access Information" class, using the property `jps:accessInfo`. This includes both the access to the original cultural artifact and to its digital surrogates. For example, in the case of a painting, a user can visit the museum with information given by `schema:provider` and find the exhibited original through its catalog number in `jps:contentId`, while obtaining its digitized image with `schema:associatedMedia` or through a landing page designated by `schema:url`. The access information can also accommodate rights and license information.

The class "Source Information" describes structured information on the source data and its provider. The value for `schema:provider` in the source information represents the databases that submit data to Japan Search. The values of `schema:url` here provide the URLs of the web page where the provider has published the source data in their own services. Source information also provides the last date the source data was modified (`schema:dateModified`).

3. SPARQL Endpoint and Use Case Scenarios

The normalized metadata in the JPS RDF store can be retrieved via the Japan Search SPARQL Endpoint (https://jpsearch.go.jp/rdf/sparql/). Since the normalized data in Japan Search are linked to LOD URIs, it is possible to cross-walk Japan Search and other RDF data sources (e,g, Europeana, Wikidata, major national libraries) with SPARQL 1.1 Federated Query. Figure 4 shows the results of such a query, which uses a DBpedia URI to identify 曲亭馬琴 (Kyokutei, Bakin, 1767-1848), then retrieves his works from Europeana and Japan Search. In this way, users can find works of Kyokutei Bakin in a single step: not only those held by Japanese institutions via Japan Search, but also those held in European institutions via Europeana.

SPARQL results:

<< rev 1 - 100 / 1949 next >>

uri	label	image
<http://data.europeana.eu/proxy/provi ... ce_3000095443819>	"Kōdan tsutsumi no io"	
<http://data.europeana.eu/proxy/provi ... ce_3000095474046>	"Mukashigatari shichiya no kura : Shohen : Shichiya no kura"	
<https://jpsearch.go.jp/data/dignl-10303726>	"千代褸良著聞集 12巻 三"	
<https://jpsearch.go.jp/data/dignl-2546342>	"南総里見八犬伝 9輯98巻 [5]"	
<https://jpsearch.go.jp/data/dignl-2551620>	"南総里見八犬伝 9輯98巻 [33]"	

Fig. 4. Federated Query results for works by "曲亭馬琴 (Kyokutei, Bakin)", Europeana and Japan Search

Experiments on utilization of Japan Search have been attempted since its launch. Among such attempts is a provisional analysis of the types of content that characterize each prefecture. The result demonstrates that Tokyo is by far the largest holder of JPS content, with its sheer number of proceedings, publications, and journal articles. The same analysis also reveals that architectural heritage is rather more concentrated in the western part of Japan than in the east. The results of analysis can be effectively visualized using the normalized temporal and spatial data, which can be used to showcase what would be made visible by an aggregated mass of data. The project is aiming to release a new function where online galleries of various content can be generated automatically based on the links contained in JPS RDF.

24

DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2019

4. Conclusion

The JPS RDF is designed to serve a wide range of uses from simple to complex searches, federated query with outside RDF sources, and data analysis of cultural contents. The system and the schema will be continually amended, with the year 2020 as a tentative cutoff date. In the coming years, the NDL will further strengthen cooperation with cultural and scholarly institutions across Japan for the refinement and enrichment of the aggregated data and for the promotion of innovative uses of cultural and scholarly content data.

References

National Diet Library. (2019). Japan Search (beta): For Developers. Retrieved April 25, 2019, from https://jpsearch.go.jp/api.

Kanzaki, Masahide. (2019). Japan Search Unofficial Support Page. https://www.kanzaki.com/works/ld/jpsearch/. Retrieved April 25, 2019 from https://www.kanzaki.com/works/ld/jpsearch/.

DCPAPERS

Proc. Int'l Conf. on Dublin Core and Metadata Applications 2019