DCPAPERS

A Preliminary Study on Metadata Compatibility and Automatic Conversion among LOD Datasets

Yan CONG[1,*,†], Masao TAKAKU[2,†], Yasuyuki MINAMINYAMA[3,†], Takaaki AOKI[1,†] and Shigeki MATSUBARA[1,†]

[1]Nagoya University, Japan
[2]University of Tsukuba, Japan
[3]The University of Tokyo, Japan

Abstract

In recent years, open science has promoted the web-based publication and reuse of research data. Publishing such data typically involves selecting a domain-specific metadata schema and using repositories for distribution. In Linked Open Data (LOD) datasets, domain-specific metadata schemas like VoID and DCAT are commonly used. However, differences in metadata schemas across institutions hinder cross-domain search and reuse due to duplicated terms and semantic ambiguity. The lack of clear metadata guidelines and examples leads to inconsistent metadata creation, ultimately compromising interoperability. This study addresses these issues by proposing a method to enhance compatibility among metadata schemas. Specifically, it focuses on mapping LOD schemas from VoID and DCAT to DataCite metadata schema. Additionally, it explores automating conversion process by pseudocode. This approach aims to align semantics across domain-specific metadata schemas and improve the findability and interoperability of dataset metadata.

Keywords

Metadata Compatibility, Metadata Management, Data Distribution, Dataset Description, VoID, DCAT,

1. Introduction

In recent years, the movement toward open science, promoting the widespread publication of research data on the web and encouraging its reuse has rapidly gained momentum. Various research institutions and data-sharing platforms started to share publicly and provide a wide variety of datasets. However, these research institutions and data-sharing platforms often use different domain-specific metadata schemas to manage data. As a result, cross-institutional search and reuse of data have become difficult.

Moreover, the problems such as overlaps in property names and semantic ambiguities across different metadata schema properties present a significant challenge. For instance, different platforms use distinct property names to refer to the same concept, or the same term may be

DCMI-2025 International Conference on Dublin Core and Metadata Applications
*Corresponding author.
†These authors contributed equally.
cong.y@nagoya-u.jp (Y. CONG); masao@slis.tsukuba.ac.jp (M. TAKAKU); minamiyama@iss.u-tokyo.ac.jp (Y. MINAMINYAMA); aoki.takaaki@nagoya-u.jp (T. AOKI); matubara@nagoya-u.jp (S. MATSUBARA)
0000-0002-6441-4603 (Y. CONG); 0000-0002-2458-6988 (M. TAKAKU); 0000-0002-7280-3342 (Y. MINAMINYAMA); 0000-0002-5926-4903 (T. AOKI); 0000-0003-0416-3635 (S. MATSUBARA)

https://doi.org/10.23106/dcmi.952519392

used with different meanings depending on the context. Such incompatibilities among metadata schema properties reduce the overall interoperability of data and ultimately undermine the reliability of search results.

To address these issues, integrated management with compatibility based on metadata schemas is required. For instance, in Japan, the JPCOAR Schema was established in 2017 as a standardized metadata schema for datasets registered in institutional repositories, and integrated management has been carried out under the leadership of the Japan Consortium for Open Access Repository (JPCOAR)[1].

In addition, the National Diet Library (NDL) has been operating the NDL Search, a platform designed to enable unified search and access to a wide range of information resources held by libraries, academic institutions, and government agencies across the country[2][3].

Meanwhile, Linked Open Data (LOD) has gained attention as an effective approach for enhancing data interoperability and reusability. By assigning URIs and semantic links to structured data published on the Web, LOD enables meaningful connections between disparate datasets, opening up new possibilities for knowledge discovery and data integration.

In this research, the metadata of research data typically associated with datasets, such as numerical tables and images, and stored in various repositories. The "LOD dataset" is a dataset published in RDF format with semantic links, often functioning as a knowledge base with a structured and interlinked data.

Although research data and LOD datasets are often treated separately in practice, they share a common goal of promoting data sharing and reuse. By adopting the DataCite metadata schema[4], LOD datasets can be deposited and disseminated through repositories that support DOI assignment, contributing to improved findability and interoperability. Furthermore, if LOD datasets are properly structured as research data, they can typically be published in a form that not only supports the search results, but also enables reuse by other researchers.

However, semantic and structural differences between heterogeneous metadata standards pose significant challenges for findability and interoperability. This study addresses these challenges by examining metadata compatibility and proposing a method for converting LOD dataset schemas into the DataCite metadata schema[4].

Building on this understanding, this research examines the interoperability of metadata in LOD datasets based on standardized metadata description formats. Specifically, we map the metadata schemas of LOD datasets within the Resource Description Framework (RDF), from the Vocabulary of VoID metadata schema[5] and Data Catalog Vocabulary (DCAT)[6] to DataCite metadata schema[4], which is an internationally recognized metadata standard.

Through this mapping, we compare and analyze the corresponding properties across these schemas. Furthermore, we discuss the feasibility of automated metadata conversions between different metadata schemas by pseudocode.

2. Review of Existing Practices

There are only a limited number of metadata schemas available for integrating datasets published in the LOD.

For example, the LOD Cloud[7] uses dataset metadata schemas like VoID metadata schema[5]

https://doi.org/10.23106/dcmi.952519392

and DCAT[6]. On the other hand, Google Dataset Search[8] mainly uses Schema.org[9], while allowing for the flexible use of other domain-specific metadata schemas.

However, even when using the same dataset, metadata providers sometimes present unclear or incomplete metadata. As a result, metadata creators often have to map data with some level of ambiguity, which leads to reduced compatibility between metadata schemas.

Additionally, since there is no unified rule for metadata in LOD datasets, one-to-one conversions can be challenging when the data structures differ. This can result in data loss or inconsistencies. Therefore, it is crucial to design mappings that clearly define the relationships between metadata schemas and ensure semantic consistency.

Open data portals such as Data.gov[10], the European Data Portal[11], and Japan's Government Open Data Portal[12] use DCAT metadata schema[6] for the metadata of LOD datasets. Platforms like Data.gov[10] and the European Data Portal[11] also provide information on describing LOD datasets.

The National Diet Library (NDL)[2] developed the "Metadata Distribution Guidelines"[3] (first edition) in March 2023 to improve the smooth operation and usability of NDL Search. This guideline clearly outlines the standard procedures and requirements for institutions to provide and link content to NDL Search. They specify the necessary metadata items and their description rules, the scope of information resources eligible for inclusion, and data collection methods that take into account technical and operational requirements to enhance search ability and accessibility. Additionally, NDL Search contributes to the promotion of open data, and the development of these guidelines is considered an advanced initiative led by the government to promote open science.

3. Proposed Method

This research aims to map dataset metadata from VoID metadata schema[5] and DCAT[6] to the international metadata standard DataCite metadata schema[4]. In this section, after introducing the six mandatory properties of DataCite metadata schema[4] that will be used for comparison, the proposed methods are explained in detail.

3.1. DataCite Metadata Schema

DataCite metadata schema[4] is a vocabulary for registering data and citing their metadata. DataCite[13] organization assigns a persistent identifier, Digital Object Identifier (DOI), mainly to research data, regardless of its format or units. DataCite[13] facilitates metadata exchange by connecting research outputs through DOIs. It provides standardized metadata to enhance the discoverability, accessibility, and reusability of research data.

DataCite metadata schema facilitates effective information exchange by offering a clear and consistent structure for describing datasets, native integration with persistent identifiers by DOIs, and wide adoption across major data repositories. Its well-defined metadata schemas reduce ambiguity during metadata transfer, while DOI integration ensures reliable referencing and traceability. These features make DataCite metadata schema[4] not only a strong candidate for standardizing data, but also a suitable choice for helping users discover targets across diverse repositories.

https://doi.org/10.23106/dcmi.952519392

Mandatory Properties

The mandatory properties must be supplied with any initial metadata submission to DataCite, together with their relevant sub-properties. **If one of the required properties is unavailable**, please use one of the standard (machine-recognizable) codes listed in Appendix 3: Standard values for unknown information.

Table 1: DataCite Mandatory Properties

ID	Property	Obligation
1	Identifier	M
2	Creator	M
3	Title	M
4	Publisher	M
5	PublicationYear	M
10	ResourceType	M

Figure 1: Mandatory Properties in DataCite Metadata Schema[4]

Figure1 shows the mandatory properties in DataCite metadata schema[4]. The mandatory properties include identifier, creator, title, publisher, publication year and resource type, and they are strictly defined, enhancing interoperability. We focus on the six mandatory properties defined by DataCite metadata schema[4] and map the properties of VoID and DCAT metadata schemas to them.

3.2. Proposed Method for Mapping

The mapping statuses from each metadata schema to the mandatory properties of DataCite metadata schema[4] will be determined by using the specifications of metadata schema.We defined and classified the mapping statuses as follows:

(1) An equivalent corresponding property can be identified.

(2) Some corresponding properties can be identified, but not all.

(3) A corresponding property cannot be identified. It is necessary to record mapping information in a non-standard format, such as in a “Remarks” field.

In the case of the mapping status (1), an equivalent corresponding property in DataCite metadata schema[4] can be identified by examining the definitions, usage, and examples of the properties listed in the metadata schema specifications. Furthermore, in status (2) and (3), specific dataset description examples will be investigated to confirm the usage of the properties in the metadata schema. In status (2), Some corresponding propertiess can be identified, but not all. Based on the definitions, usage, and examples listed in the metadata schema specifications.In status (3), A corresponding property cannot be identified.If no corresponding properties can be

https://doi.org/10.23106/dcmi.952519392

Table 1
Mapping Status and Results from VoID to DataCite Mandatory Schema

DataCite Mandatory Schema	Section	Mapping Status	Mapping Property	Remarks
Identifier		(2) Some corresponding properties can be identified, but not all.	dcterms:identifier	Mapping can only be performed when an identifier is present and that identifier is a DOI.
Creator	2.2	(1) An equivalent corresponding property can be identified.	dcterms:creator	
Title	2.2	(1) An equivalent corresponding property can be identified.	dcterms:title	
Publisher	2.2	(1) An equivalent corresponding property can be identified.	dcterms:publisher	
PublicationYear	2.2	(1) An equivalent corresponding property can be identified.	dcterms:issued	
ResourceType	4.5	(3) A corresponding property cannot be found.		dcterms:type is could be used in most cases.

identified in the metadata schema specifications, it will be necessary to describe the information in a non-standard format, as in the "Remarks" column.

4. Mapping Results and Automatic Conversion to DataCite Metadata Schema

We map VoID metadata schema[6] and DCAT[6] to DataCite metadata schema[4] to clarify which description elements are applicable and which are not.

4.1. Mapping VoID to DataCite Metadata Schema

Table 1 presents the mapping status and results from VoID metadata schema[5] to six mandatory properties in DataCite metadata schema[4].

First, "Identifier" can be mapped from "dcterms:identifier", only if an identifier is present and that identifier is a DOI. Second, "Creator", "Title", "Publisher" and "PublicationYear" can be mapped from VoID metadata schema[5] that conform to DCMI Metadata Terms [14], namely "dcterms:creator", "dcterms:title", and "dcterms:publisher" and "dcterms:issued". Lastly, "ResourceType", while "dcterms:type" in VoID metadata schema[5] may serve as a corresponding property, the mapping may not be entirely one-to-one, and adjustments may be required based on the repository-specific implementation policies.

https://doi.org/10.23106/dcmi.952519392

Table 2
Mapping Status and Results from DCAT to DataCite Mandatory Schema

DataCite Mandatory Schema	Section	Mapping Status	Mapping Property	Remarks
Identifier	6.4.11	(2) Some corresponding properties can be identified, but not all.	dcterms:identifier	Mapping can only be performed when an identifier is present and that identifier is a DOI.
Creator	6.4.4	(1) An equivalent corresponding property can be identified.	dcterms:creator	
Title	6.4.6	(1) An equivalent corresponding property can be identified.	dcterms:title	
Publisher	6.4.10	(1) An equivalent corresponding property can be identified.	dcterms:publisher	
PublicationYear	6.5.3	(1) An equivalent corresponding property can be identified.	dcterms:issued	
ResourceType	6.4.13	(3) A corresponding property cannot be identified.		dcterms:type is could be used in most cases.

4.2. Mapping DCAT to DataCite Metadata Schema

Table 2 shows the mapping status from DCAT[6] to mandatory properties in DataCite metadata schema[4].

First, "Identifier" can be mapped from "dcterms:identifier" as specified in Section 6.4.11 of DCAT [6], only if an identifier is present and that identifier is a DOI. Second, "Creator" (Section 6.4.4, "dcterms:creator"), "Title" (Section 6.4.6, "dcterms:title"), "Publisher" (Section 6.4.10, "dcterms:publisher"), and "PublicationYear" (Section 6.4.7, release date, "dcterms:issued") are consistently aligned with the corresponding properties in DataCite metadata schema[4]. Lastly, "ResourceType", Section 6.4.13 "Property: type/genre" of DCAT [6] is not directly applicable. However, based on case studies, "dcterms:type" in DCAT[6] can be used to map the property.

4.3. Automatic Conversion from VoID to DataCite Mandatory Schema

We propose an automatic conversion method for mapping VoID metadata schema[5] to DataCite metadata schema[4] by pseudocode.

"dcterms:identifier" is processed as in Figure 2. The process involves extracting the "dcterms:identifier" property contained within the VoID metadata, removing unnecessary whitespace, and normalizing the value. The "extractIdentifier" function checks whether "dcterms:iden-

https://doi.org/10.23106/dcmi.952519392

```
    if voidMetadata.has("dcterms:identifier"):
        identifier = voidMetadata.get("dcterms:identifier")
        return normalizeIdentifier(identifier)
    else:
        raise Error("dcterms:identifier not found in VOID metadata")
function normalizeIdentifier(identifier):
    return identifier.strip()
```

Figure 2: Conversion from VoID to DataCite by Identifier

```
function extractCreator(voidMetadata):
    if voidMetadata.has("dcterms:creator"):
        creators = voidMetadata.getAll("dcterms:creator")
        return [normalizeCreator(c) for c in creators]
    else:
        raise Warning("Creator not found in VOID metadata")
function normalizeCreator(creatorEntry):
    if isURI(creatorEntry):
        return lookupLabelFromURI(creatorEntry)
    else:
        return creatorEntry
```

Figure 3: Conversion from VoID to DataCite by Creator

tifier" exists in the "voidMetadata". If it exists, the function retrieves its value and passes it to the "normalizeIdentifier" function. If "dcterms:identifier" does not exist in "voidMetadata", the function trims leading and trailing whitespace from the string and returns an error.

Figure 3 shows an example "dcterms:creator" property from VoID metadata schema[5] and performs normalization as needed. In the "extractCreator" function, if "dcterms:creator" exists, all of its values are retrieved, and each value is normalized using the "normalizeCreator" function. The helper function "normalizeCreator" determines whether the input is a URI or a string. If it is a URI, it performs a label conversion process (e.g., using "lookupLabelFromURI"); if it is a string, it returns the string as is. If the corresponding property does not exist, a warning is raised.

As shown in Figure 4, the property "dcterms:issued" (date of issue) is used as the corresponding property for PublicationYear in DataCite metadata schema[4]. Since "PublicationYear" (in YYYY format only) is a required field in DataCite metadata schema[4], it is necessary to extract from data-type values accordingly. It is assumed that the date strings follow the ISO 8601 format (e.g., "2025-4-25"). During the conversion process, only the year component is extracted and passed to DataCite metadata schema[4]. If multiple candidate properties are present, the date of issue ("dcterms:issued") is given the highest priority.

As shown in Figure 5, the property in VoID metadata schema[5] corresponding to "Resource-Type" in DataCite metadata schema[4] is considered to be "dcterms:type" . In DataCite metadata schema[4], resource types are strictly defined (e.g., Dataset, Text, Image, etc.), and the val-

https://doi.org/10.23106/dcmi.952519392

```
function extractPublicationYear(voidMetadata):
    if voidMetadata.has("dcterms:issued"):
        issued = voidMetadata.get("dcterms:issued")
        return normalizeYear(issued)
    else:
        raise Exception("Publication year not found in VOID metadata")

function normalizeYear(issuedValue):
    // Extract 4-digit year from date string (e.g., "2025-05-12" → "2025")
    return issuedValue[0:4]
```

Figure 4: Conversion from VoID to DataCite by PublicationYear

```
function extractResourceType(voidMetadata):
    if voidMetadata.has("dcterms:type"):
        rawType = voidMetadata.get("dcterms:type")
        typeLabel = isURI(rawType) ? lookupLabelFromURI(rawType) : rawType
        return mapToDataCiteResourceType(typeLabel)
    else:
        raise Warning("ResourceType not found in VOID metadata")
function mapToDataCiteResourceType(typeLabel):
    mapping = {
        "Dataset": "Dataset",
        "Data Collection": "Dataset",
        "Image": "Image",
        "Text": "Text",
        "Software": "Software",
        "Service": "Other",
        "Ontology": "Other"
    }
    normalized = typeLabel.strip().toLowerCase()
    for key in mapping.keys():
        if key.toLowerCase() in normalized:
            return mapping[key]
    return "Other"
```

Figure 5: Conversion from VoID to DataCite by ResourceType

ues of “dcterms:type” in VoID metadata schema[5] cannot always be used as-is. Therefore, a transformation using a mapping table may be necessary.

Additionally, “dcterms:type” can sometimes be ambiguous or described in a proprietary way, so it is important to define a mapping table in advance.

https://doi.org/10.23106/dcmi.952519392

5. Discussion

In Chapter 5, based on the results of mapping from VoID metadata schema[5] and DCAT[6] to DataCite metadata schema[4], the compatibility between different metadata schemas will be discussed.

5.1. Mapping Based on DataCite Metadata Schema

In Chapter 3, based on DataCite metadata schema[4], a mapping was carried out for the six mandatory metadata properties. This demonstrated the basic potential for compatibility between different metadata schemas. However, in order to achieve high reusability and interoperability, a more comprehensive mapping, including the recommended properties are necessary.

On the other hand, when converting RDF schema to XML schema, a key challenge is that many properties cannot be mapped one-to-one due to structural differences. RDF is based on a flexible graph structure that represents relationships, whereas XML uses a strict hierarchical and ordered structure. We developed mapping rules to carefully convert RDF nodes into nested XML elements, ensuring that the original meanings and relationships are preserved throughout the conversion process. This study shows that there is a balance to be found between covering all possible metadata details and keeping the mapping practical. The six mandatory DataCite metadata schemas offer a simple, common base that helps different datasets work together, but they don't cover all the specific details that researchers in different fields need. Also, because each field uses metadata differently and without a clear standard, it's hard to create one mapping that fits all.

Because of this, we focus on these six mandatory schemas as a practical starting point for making datasets compatible across fields. However, future work needs to add more recommended and field-specific metadata to better capture important details.

To enhance the reusability of metadata, it is essential to provide not only structural consistency but also semantic information. For example, if a manual summarizing the usage rules and guidelines for the research data is provided, it can become an important factor in effective reuse by third parties when mapping the data. But also, we carefully consider the semantic distinctions among candidate properties to make informed mapping decisions.

Based on the above, future research should focus on comprehensive mappings not only for required properties but also recommended properties, along with their qualitative evaluation, and further verification of compatibility across disciplines, then we will implement automatic conversion afterward.

5.2. Duplicate Properties and Semantic Ambiguity in Metadata Schema

Several properties and elements among the metadata schema share similar meanings. For example, in VoID metadata schema[5], there are three candidates related to dates: "dcterms:issued" (date of issue), "dcterms:created" (date of creation), and "dcterms:date" (other related dates). Although these metadata schemas may appear similar at first glance, each is assigned a different meaning, requiring a decision on which metadata schema should be prioritized during mapping to the DataCite metadata schema[4].

https://doi.org/10.23106/dcmi.952519392

The duplication and semantic ambiguity of metadata properties in such datasets can be confusing when performing integrated mappings across data, particularly when extracting academic repositories or research data, which may lead to search omissions. A future challenge is to consider the development of guidelines for mapping. Additionally, by designing criteria for selecting the preferred metadata schema based on the actual metadata description practices and evaluating them empirically, it is expected that more reliable mappings can be achieved.

6. Conclusion

This study clarified the differences and description formats between different metadata schemas, including VoID metadata schema[5], DCAT[6], and DataCite metadata schema[4]. After organizing the compatibility between each schema, the mapping methods were proposed, and the potential for mapping design and automatic conversion for integrated handling was explored.

Future challenges include systematizing mapping rules between different metadata schemas and their automatic conversion, building quality verification mechanisms for integrated metadata, and considering the practical usability and evaluation in real-world operations. Additionally, it will be necessary to explore the compatibility of dataset metadata not only with the DataCite metadata schema[4] but also with other recommended properties, aiming to develop a more versatile integrated mapping. But also, we plan to further investigate how our approach can align with broader metadata crosswalk initiatives and standards efforts such as those by Schema.org, and so on.

Future work aims to improve the robustness of the conversion algorithms by incorporating comprehensive error handling and addressing data quality challenges. We will extend the date parsing logic to recognize common non-ISO formats such as YYYY/MM/DD and MM-DD-YYYY. we plan to enhance the handling of language tags and internationalization to ensure accurate multilingual metadata representation when we design the system.

Acknowledgement

This research was supported by MEXT as “Developing a Research Data Ecosystem for the Promotion of Data-Driven Science”.

References

[1] Hayahiko OOZONO and Tomoko KATAOKA and Nanako TAKAHASHI and Tadasuke TAGUCHI and Yutaka HAYASHI and Yasuyuki MINAMIYAMA, JPCOAR Schema: A brand-new metadata schema for smooth international distribution of scholarly output in Japan , Journal of Information Processing and Management (JOHOKANRI) 60 (2018) 719–729. doi:10.1241/johokanri.60.719.

[2] National Diet Library, NDL Search, https://ndlsearch.ndl.go.jp/, 2023. Accessed: April 12, 2025.

[3] National Diet Library, Metadata Distribution Guidelines, https://ndlsearch.ndl.go.jp/guideline, 2023. Accessed: April 12, 2025.

https://doi.org/10.23106/dcmi.952519392

[4] DataCite, DataCite Metadata Properties, https://datacite-metadata-schema.readthedocs.io/en/latest/properties/, 2025. Accessed: April 8, 2025.
[5] W3C, Vocabulary of Interlinked Datasets (VoID). W3C Recommendation, https://www.w3.org/TR/void/, 2013. Accessed: April 8, 2025.
[6] W3C, Data Catalog Vocabulary (DCAT) Version 3.0. W3C Recommendation, https://www.w3.org/TR/vocab-dcat-3/, 2023. Accessed: April 8, 2025.
[7] LOD Cloud, Linked open data cloud, https://lod-cloud.net/, 2025. Accessed: April 8, 2025.
[8] Google, Dataset Search –Google, https://datasetsearch.research.google.com/, 2025. Accessed: April 8, 2025.
[9] Schema.org, Schema.org - Structured Data on the Web, https://schema.org/, 2025. Accessed: April 8, 2025.
[10] U.S.General Services Administration, Data.GOV, https://data.gov/, 2025. Accessed: April 12, 2025.
[11] Publications Office of the European Union, data.europa.eu: The official portal for european data, https://data.europa.eu/en, 2025. Accessed: April 12, 2025.
[12] Digital Agency, Japan, e-GOV Data Portal, https://data.e-gov.go.jp/info/en, 2025. Accessed: April 12, 2025.
[13] DataCite Metadata Working Group, DataCite Metadata Schema Documentation for the Publication and Citation of Research Data, Version 4.6, https://datacite-metadata-schema.readthedocs.io/en/latest/, 2024. Accessed: April 8, 2025.
[14] Dublin Core Metadata Initiative, DCMI Metadata Terms, https://www.dublincore.org/specifications/dublin-core/dcmi-terms/, 2020. Accessed: April 12, 2025.

https://doi.org/10.23106/dcmi.952519392