Towards Development of Ontology for the National Digital Library of India

Susmita Sadhu, Poonam Anthony, Plaban Kumar Bhowmick, and Debarshi Kumar Sanyal

Indian Institute of Technology, Kharagpur 721302, India
{susmita90iitkgp, anthony.poonam, plaban, debarshisanyal}@gmail.com

Abstract. In this paper, we present the design of an ontology for describing the resources (mostly educational) in National Digital Library of India (NDLI).With the proposed ontology, We have intended to model a diverse set of scholarly resources present in the NDLI repository, which includes books, research articles, dissertations, videos, simulations, animations, datasets, question papers, model answers and also other digital assets like movies, manuscripts, audio books etc. The designed ontology will act as a data model for the knowledge graph that is the foundation of applications like semantic search and recommendation in NDLI. The ontology has been designed targeting its alignment with Schema.org[1] data model.

Keywords: Ontology, Digital Library, Metadata, RDF/OWL

1 Introduction

The NDLI repository has been constructed with the objective of providing a common platform for facilitating access to digital resources from different source organizations. The repository index is built over the metadata of the contents, while the full-texts reside at the respective host organizations. In most cases, the metadata acquired from the source organizations is rather sparse, which is translated and further augmented to conform to the NDLI metadata schema [2]. The metadata sparsity hinders effective retrieval of documents, resulting in irrelevant documents being ranked higher for certain search queries. This scenario is especially prevalent for longer queries. Moreover, the system currently employs a simple keyword-based search, which falls short when it comes to semantic processing of a natural language query given by a user, such as "all books on computer science authored by a Computer scientist called Ullman". A possible approach to mitigate this issue is to annotate the constituent phrases of the query with their corresponding metadata fields, and subsequently performing a fielded search. Further, we would like to provide support for features like resource

[1] http://schema.org/

[2] https://docs.google.com/spreadsheets/d/1BjXleBDV2QhChiv3TSoi4n5jLwVFy06dV-HLMfyu21w/edit?usp=sharing

recommendation to help boost the learning utility of NDLI. This would require identification of similar/related resources in the repository. Similarity between a pair of resources may be decided by the presence of shared metadata fields or their combinations.

We would focus to leverage the power of knowledge graph based solutions for the aforementioned problems. The knowledge graph can be constructed from the metadata descriptions of the resources. The limitation of sparse metadata may be handled through semantic enrichment of knowledge graph by linking it to other external resources like DBpedia[3], BIO2RDF2[4] and likes. This extended knowledge graph would facilitate semantic search, thus improving document retrieval, as discussed above. We can further augment the knowledge graph by establishing linkages between nodes having similar content, which would provide an effective base for developing a recommendation system, to provide users with a list of suggested resources, related to the one s/he peruses. Additionally, publishing scholarly data as linked data would enable reuse of existing resources. Construction of NDLI knowledge graph would require defining of the underlying ontology for describing the various resources present in the repository.

The metadata schema of NDLI incorporates fields from Dublin Core[5] and Learning Resource Metadata Initiative[6], to comprehensively represent a wide range of learning resources, and resources from general domain, which include contents of varied types like videos, texts, audios, simulations and so on, classified into more than 60 types based on their educational context. Consequently, the metadata schema is complex and contains a number of fields, including some special fields that are resource-specific. A survey of current practices in modern digital libraries indicates a significant drift towards semantic and linked data based solutions[1]. As the general scopes and repository constituents of NDLI vary significantly from the existing LDMs like that of Europeana Data Model (EDM)[2] and Library of Congress[3], the ontological models adopted by them can not be readily applicable to NDLI domain. Schema.org[4] has gained enormous popularity in terms of representing data models for resources on the web. Alignment with its vocabulary not only addresses the issue of interoperability but also provides means to better discoverability of the library resources by search engines.

To our knowledge, none of the existing data models in Digital Library domain seek to align themselves with Schema.org. We would like to adopt the same to enhance Web discoverability of the NDLI resources. Since, the resources in a digital library are predominantly creative works, these can easily be represented using Schema.org vocabulary. However, owing to the diverse nature of resources included in the repository, and various resource-specific fields, the NDLI meta-

[3] http://wiki.dbpedia.org/

[4] http://bio2rdf.org

[5] http://dublincore.org/

[6] http://dublincore.org/dcx/lrmi-terms/

data schema can not be modeled entirely using Schema.org. Hence, it would be necessary to augment our ontology with additional class and property definitions, to represent those elements of the metadata, which Schema.org is unable to capture. In this respect, the contributions of our work are as follows -

1. Design of a custom ontology based on the metadata schema of NDLI, derived from Schema.org
2. Creation of Linked Open Data model, complying to the ontology to facilitate various features for enhancing NDLI as a learning platform.

2 Design of Ontology

2.1 Design Intent

The fields of the NDLI metadata schema can be broadly classified into the following categories: *1. Controlled Vocabulary* - These fields can only take values from a pre-defined vocabulary, e.g., content type, language etc.; *2. Formatted* - Values are of fixed datatype like Date, e.g., date of published etc.; and *3. Free Text* - Fields like author, source, title, abstract etc. The metadata also incorporates hierarchies like DDC[7] and MeSH[8] for classifying subjects. Our ontology design takes into consideration the following aspects of the metadata schema:

1. Class definition for various entities
   (a) Each item in the repository can be modelled as an entity belonging to a particular class like book, article, thesis, video lecture, etc.
   (b) We also define enumeration classes to represent the DDC and MeSH hierarchies, and other controlled vocabulary domains, like content type, language, educational level, etc.
2. Property definition
   Each metadata field is mapped to an ontological property, which is again mapped to OWL supported property taxonomy.
3. Alignment with Schema.org.

2.2 Ontology specification

In the following subsections, we describe the classes and properties that constitute the ontology and discuss its alignment with Schema.org.

Core Ontology

The ontology consists of a top class *CreativeWork*, which is the parent of a number of sub-classes that belong to two broad categories:

1. Learning Resources -
   The class *LearningResource* has been defined as a sub-class of *CreativeWork*, which is further partitioned into a number of sub-classes like Book, Article, VideoLecture, AudioLecture, QuestionPaper etc., to accommodate a wide range of learning resources that reside in the NDLI repository.

[7] https://www.oclc.org/dewey/features/summaries.en.html

[8] https://meshb.nlm.nih.gov/search

2. Non-learning Resources -
   The non-learning type of resources present in the repository are represented using classes like Painting, Movie, MusicAlbum, LegalDocument etc., all of which are defined as sub-classes of *CreativeWork*.

Apart from repository contents, we have also defined various classes to represent other entities, which include:

- Contributors-
  The contributor of a resource is represented using the classes *Author, Editor, Director, Judge*, and so on, derived from *Person* class.
- Organizations -
  Organization class is used to describe entities like sponsor, publisher and source organization of a resource.
- Enumeration classes -
  The ontology consists of some enumeration classes like *AgeRange, ContentType, EducationalLevel, EducationalFrameWork, EducationalUse, FileFormat* to express those metadata fields whose values belong to a controlled vocabulary. For example, *ContentType* is instantiated as *Text, Audio, Video* etc.

Every content is defined as an instance of any of the above sub-classes of *CreativeWork*, while its corresponding metadata fields are represented using different properties included in the ontology design. We have defined three types of properties based on the values of the metadata:

1. *Object Properties*-
   These represent the metadata fields whose value is also an entity. For example, we define properties like *hasAuthor, hasEditor* etc. linking *CreativeWork* to the different contributor classes. Inverse properties like isAuthorOf, isEditorOf, etc have also been defined to capture the inverse relationship between the classes. Properties like *typicalAgeRange, hasContentType, hasEducationalLevel, hasEducationalFrameWork* etc., connect the creative works to the different enumeration classes defined in our ontology.
2. *Data Properties*-
   These include properties like *firstName, lastName, abstract, datePublished, keywords* etc., which are used to represent those metadata whose values are literals.

The ontology, as a whole consists of 50 classes, of which 14 are enumeration classes. A total of 120 properties have been defined to depict the links between these classes, which are comprised of 90 object properties and 30 data properties.

Alignment with Schema.org

We have tried to conform our ontology definition to Schema.org as far as possible, introducing a few new classes and properties in order to adapt to the NDLI metadata schema. The *CreativeWork* class is aligned to Schema.org. However,

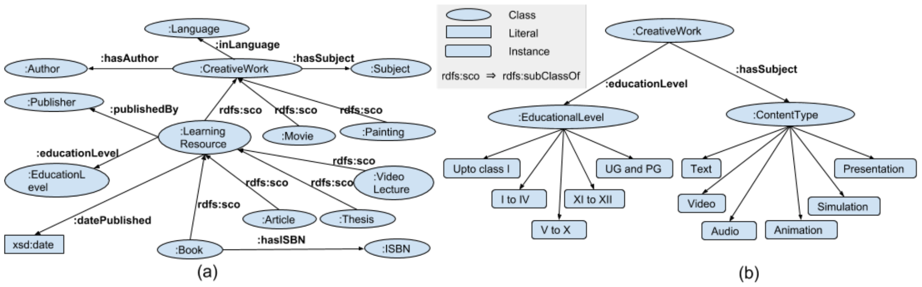

Fig. 1. Overview of classes and properties in ontology

as the NDLI repository is dominated primarily by learning materials, we digressed slightly from Schema.org to define *LearningResource*, as a sub-class of *CreativeWork*. Object properties *hasAuthor* and *hasEditor* of *CreativeWork* are derived from *author* and *editor* properties defined in Schema.org, while properties like *hasDirector, hasJudge* have been introduced to adapt to the NDLI metadata schema. Also, object properties *hasFileFormat, typicalAgeRange*, differ from their Schema.org counterparts, since their range is comprised of enumeration classes. Further, some other properties like *inLanguage, hasPart, isPartOf, keywords, publisher, sponsor* have been preserved according to the Schema.org vocabulary.

Figure 1(a) shows a portion of the ontology, highlighting some of the major classes and properties among them. Table 1 lists some of the properties defined in our ontology, while Figure 1(b) shows some enumeration classes.

Metadata Field	**Property**	**Type**	**Domain**	**Range**
author	hasAuthor	Object Property	CreativeWork	Person
publisher	publishedBy	Object Property	CreativeWork	Publisher
content type	hasContentType	Object Property	CreativeWork	ContentType
isbn	hasISBN	Object Property, Functional Property	CreativeWork	ISBN
date of published	datePublished	Data Property	LearningResource	Date
abstract	abstract	Data Property	CreativeWork	String

Table 1. Properties with types

The knowledge graph construction involves translation of the NDLI data store to a format, conforming to the ontology. We use RDF[9] to model our graph, since it is resilient to change in information schema, and is a standard adopted by all linked-data sources like DBpedia. In Figure 2 a sample graph for an item is portrayed under ontological structure. Every content in NDLI is represented

[9] https://www.w3.org/RDF/

as an instance of the appropriate sub-class of *CreativeWork*, having out-links corresponding to the ontological properties.

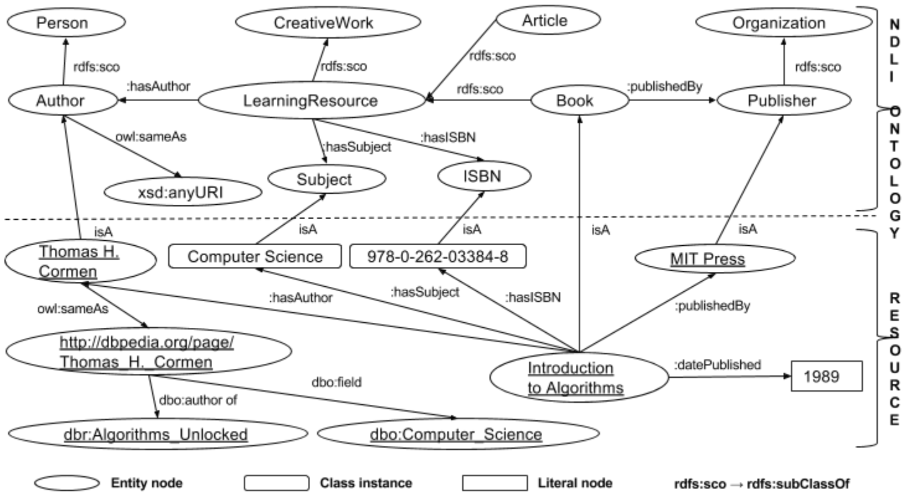

Fig. 2. Knowledge Graph Representation of a resource

3 Conclusion

The ontology designed for NDLI is a rich, extensible one that can model diverse resource types. It is primarily aligned to Schema.org, to support web-compliance and allowing us to link to external data sets. We expect that the ontology-based knowledge graph and applications built around it, will prove fruitful in enriching the learning experience of users in NDLI.

References

1. S. R. Kruk and B. McDaniel, *Semantic digital libraries.* Springer, 2009.
2. M. Doerr, S. Gradmann, S. Hennicke, A. Isaac, C. Meghini, and H. van de Sompel, "The europeana data model (edm)," in *World Library and Information Congress: 76th IFLA general conference and assembly*, pp. 10–15, 2010.
3. A. Kroeger, "The road to bibframe: the evolution of the idea of bibliographic transition into a post-marc future," *Cataloging & classification quarterly*, vol. 51, no. 8, pp. 873–890, 2013.
4. R. V. Guha, D. Brickley, and S. Macbeth, "Schema. org: Evolution of structured data on the web," *Communications of the ACM*, vol. 59, no. 2, pp. 44–51, 2016.