MOD: Metadata for Ontology Description and Publication

Biswanath Dutta
DRTC, Indian Statistical Institute
Bangalore, India
bisu@drtc.isibang.ac.in

Durgesh Nandini
B.C. Roy Engineering College
Durgapur, India
durgeshnandini16@yahoo.in

Gautam Kishore Shahi Dept. of IT, Birsa Institute of Technology, India
gautamshahi16@gmail.com

Abstract

Ontology is an important artifact of Semantic Web applications. Today, there are an enormous number of ontologies available on the Web. Even so, finding and identifying the right ontology is not easy. This is because the majority of ontologies are either not described or described with a general-purpose metadata vocabulary like Dublin Core. On the other hand, ontology construction, irrespective of its types (e.g., general ontology, domain ontology, application ontology), is an expensive affair both in terms of human resources and other infrastructural resources. Hence, the ideal situation would be to reuse the existing ontologies to reduce the development effort and cost, and also to improve the quality of the original ontology. In the current work we present an ontology metadata vocabulary called Metadata for Ontology Description and publication (MOD). To design the vocabulary, we also propose a set of generic guiding principles and a well-established methodology which take into account real concerns of the ontology users and practitioners.

Keywords: metadata, ontology metadata vocabulary, ontology publication, resource description, ontology reuse, ontology library, methodology, metadata design principles, semantic application

1. Introduction

Ontology (a formal, explicit specification of a shared conceptualization (Studer et al., 1998)) construction is an expensive affair both in terms of human and other infrastructural resources. One of the fundamental principles of ontology development is to look for existing ontologies to reuse (Dutta, B. et al., 2015) before deciding to construct one from scratch. In this context, a new type of library that stores ontologies, called an Ontology Library (Ding and Fensel, 2001; d'Aquin & Noy, 2012), plays a crucial role. The goal of an ontology library is to support users to search and retrieve ontologies for the purpose of reusing them. However, in our opinion as ontology practitioners, theoretically the idea of ontology reuse sounds appealing, but in practice it is not easy to implement. There are various reasons why reuse may not be easy to practice. For example, reuse, whether partially or in full, can happen only when there is a match between the user goal of using an ontology and the development goal of an existing ontology. Obrst, et al. (2014) has discussed many such concerns in the form of "what limits ontology reuse?" One of the possible concerns is highly relevant to the current work, i.e., how to find Mr. Right Ontology? To quote them:

> …more than a simple registry of ontologies is needed – there must also be ways of organizing and annotating the ontologies with the appropriate metadata so that users can find the ontologies that match their requirements.

They further state that in addition to notions such as provenance (as captured by Ontology Metadata Vocabulary (OMV) (Hartmann, Jens et al., 2005), which is so far the only existing metadata vocabulary for ontology description), the metadata must include a wider range of features. For instance, metadata from a *development perspective* consists of information such as competency questions, ontological commitments and design decisions; metadata from an

https://doi.org/10.23106/dcmi.952136974

implementation perspective consists of information for reasoning support, languages, rules, conformance to external standards and so forth.

The current work proposes an ontology metadata vocabulary, called Metadata for Ontology Description and publication (MOD). In designing MOD, we have considered the above recommendations of Obrst, et al. (2014) as well as recommendations made by various other ontology practitioners and users in the literature including d'Aquin & Noy (2012). The main contributions of the current work are as follows: proposes an easy to use and well-defined ontology metadata vocabulary MOD, which considers the real concerns of the practitioners and ontology users; proposes a set of *generic* guiding principles and a methodology for designing a metadata vocabulary.

The rest of the paper is organized as follows: section 2 discusses the current state of the ontology libraries. It provides answers to some of the following questions, such as how many metadata elements are used by existing ontology libraries? Do they use any standard vocabularies to describe the ontologies?; section 3 discusses MOD design principles as well as the methodology; section 4 discusses a set of top-level facets, describing the various perspectives of an ontology, that are defined to design the current MOD vocabulary; section 5 provides details of the MOD vocabulary; section 6 discusses some of the related state-of-the-art works. Finally, section 7 concludes the paper.

2. Ontology Metadata in Practice: The Current State of Ontology Libraries

In this section, we present the results of our study of the usage of metadata by the existing ontology libraries on the Web to describe and publish ontologies. Before we discuss the results, we will first briefly define an ontology library and discuss its purpose.

In general, an ontology library is a collection and organization of ontologies. The purpose of an ontology library is to allow users to search, browse, refer and evaluate ontologies for different tasks. The ontology libraries are generally classified into three broad categories: *ontology repository*, *ontology registry* and *ontology directory* (Debashis, N., 2014). We have identified a total of *13* such libraries on the Web. These include Bio-portal, DERI, OBO Foundry, ROMULUS, colore, etc. as shown in Table 1.

TABLE 1: Ontology libraries along with their number of metadata elements

Ontology Library	Number of Elements	Example Elements	Metadata Followed
Bio-Portal (https://bioportal.bioontology.org/)	30	Acronym, People, Number Of Properties, Status, Description	Partially OMV plus own defined elements
Colore (https://code.google.com/p/colore/source/browse/trunk/ontologies/approximate_point)	7	Source Path, File Name, Size, Rev, Author	None
DAML (http://www.daml.org/ontologies/)	12	Link, Description, Submitter, Point of contact, Submitter	None
DERI (http://vocab.deri.ie/)	4	Author, Terms, Last Update, Namespace URI	None
Maven (http://mvnrepository.com/artifact/edu.stanford.protege)	4	Artifact, Last Version, Popularity, Description	None
MISO (http://www.sequenceontology.org/)	6	Definition, Synonyms, DB Xref, Parent, Children	None
MMI (http://mmisw.org/)	22	Full Title, Contact Role, Syntax Format, Authority abbreviation, Contributor, Keywords	None
OBO Foundry (http://www.obofoundry.org)	12	Namespace, Current Activity, Help, Home, Documentation, Contact	None
ONKI (http://onki.fi/en/browser/)	11	Type, URI, Share, superordinate concepts, Coordinate concepts	None

https://doi.org/10.23106/dcmi.952136974

Ontohub (https://ontohub.org/ontologies)	24	Project Name, Description, Institution, URL, task	Partially OMV plus own defined elements
ROMULUS (http://www.thezfiles.co.za/ROMULUS/)	35	Ontology Name, License Description, Project Domain, Creation date, DL expressivity, Number of classes, Number of individuals	Partially OMV plus own defined elements
Schemapedia (http://datahub.io/dataset/schemapedia)	4	Subject, Property, Source	None
SHOE (http://www.cs.umd.edu/projects/plus/SHOE/onts/)	4	Id, Version, Description, Contact	None

To understand the state-of-the-art practices and the metadata usages among the ontology libraries, we have studied each of these 13 libraries thoroughly and have noted the metadata elements they use. We have also tried to find information on whether any of these libraries follow a metadata standard. A consolidated result of our study based on the above parameters is presented in Table 1.

It can be seen from the above Table 1 that except three libraries, namely, Bio-portal, Ontohub and ROMULUS, none of the other libraries use a metadata standard or controlled vocabulary system in describing the ontologies. These three libraries partially use a metadata vocabulary called Ontology Metadata Vocabulary (OMV). In addition to OMV metadata elements, these libraries also use additional self-defined metadata elements. Zubeida and Keet (2013) have observed that OMV is not sufficient for an extensive and descriptive list of metadata for the ontologies. This deficiency in ontology metadata vocabulary may create an obstacle in ontology reuse. It can be further seen from the above table that the usage of a number of metadata elements varies from library to library. The majority of the libraries (70%) are found to be using 15 or fewer than 15 elements. This indicates that the metadata set should not be too large.

By analyzing the above libraries and their metadata, we have also observed that different terms are used in describing similar information in different libraries. For instance, the majority of the libraries have used the term "author" to capture the author information of an ontology, while some of the libraries have used the term "creator" (e.g., ROMULUS). This occurs when we do not use any standard or controlled vocabulary system. The practice of using ad hoc solutions creates obstacles in achieving interoperability among the ontology libraries.

Given the above observations, we have designed MOD as a controlled metadata vocabulary system that can be used by the community. We have tried to provide a minimal set of elements, but keep the essential elements that would be needed to describe an ontology and support ontology reuse.

3. MOD Approach

The MOD approach involves two crucial components: guiding principles and methodology. These are discussed in the following.

3.1 Guiding Principle

To design MOD, we developed some generic principles that acted as guidelines for us in the process of creating the vocabulary. The principles are important to assure the consistency and effectiveness of the vocabulary. The principles are:

1. *Principle of brevity*: The vocabulary should consist of a minimal set of elements maintaining balance between *necessity and sufficiency*.
2. *Principle of clarity*: The metadata elements must be well defined and clear descriptions should be provided.
3. *Principle of simplicity*: The vocabulary should be easy to use.

https://doi.org/10.23106/dcmi.952136974

4. *Principle of authority*: The vocabulary design should be based on a sound methodology in the sense that the inclusion of terms in the vocabulary are justified.
5. *Principle of standardization*: The element names should be standardized. To confirm the standardization, the individual elements should be mapped with the existing standard vocabularies.
6. *Principle of extensibility*: The vocabulary should be extensible.
7. *Principle of usability*: The vocabulary should support the reuse of the described resources. In other words, the vocabulary should allow the creators/developers to highlight the usage and the quality of the resources in a well-defined manner.
8. *Principle of interoperability*: The vocabulary should be interoperable. It should conform to the major knowledge representation languages currently in use for Semantic Web (Berners-Lee, et al., 2001) applications.

3.2 Methodology

To build up MOD, we have used a two-way approach: Top-down approach and Bottom-up approach as discussed below.

Top-down approach

The top-down approach involves looking at the "big picture" of the metadata vocabulary. This is accomplished by defining the top-level facets conceiving the various aspects of the resource to be described. In the current work, the primary resource is an ontology. After defining the aspect, each aspect has to be further analyzed and narrowed down to define the various classes. So the top-down approach proceeds from an abstract level to a concrete level. A further explanation of this step, including the various top-level facets, is contained in section 4.

Bottom-up Approach

The bottom-up approach involves studying and identifying the properties of a resource for search and discovery to facilitate their effective reuse. This is accomplished by analyzing users' ontology search behavior, search criteria and parameters. The extracted properties are further associated with the classes defined in the top-down approach. So the bottom-up approach proceeds from a concrete level to an abstract level.

To understand the users' search behavior, search criteria and parameters, we have conducted a survey. For this, we have used an open-ended questionnaire as a tool. We circulated the questionnaire through email to people who use or deal with ontologies on a regular basis. Participation consisted of researchers and practitioners with diverse educational backgrounds including library and information science, computer science, philosophy, linguistics, etc. The participants were from various countries like India, Italy, Bangladesh, Palestine, etc. After receiving the replies, we have analyzed them and have extracted the key requirements in terms of metadata elements as discussed below.

Two specific questions were asked to the participants. These are:

(1) *How do you search an ontology on the Web or in an ontology library?*
(2) *When you search for an ontology, what is the information you look for before deciding to refer/ consult/ download it?*

We originally sent the questionnaire to a total of 18 people, out of which 12 people responded. As it was an open ended questionnaire, the responses were descriptive. Each of the responses consisted of multiple sentences (aka statements). Each sentences reflect the various actions and concerns of the participants in context of ontology search and retrieval. Table 2 lists the most frequently replied statements. The keywords of the statement have been *italicized*. MOD accommodates all of the essential and most frequently used keywords. These keywords have been

https://doi.org/10.23106/dcmi.952136974

compiled and framed to form the elements of MOD. These responses have not only provided sufficient input for deciding the metadata elements of MOD, but have also provided a potential foundation to outline the multi-faceted approach to the metadata in the early stages of its development.

TABLE 2: Ontology user responses

Statement 1: I look at the ontology descriptors like *domain* details, *number of classes, properties, tools used*. ***Statement 2***: I look for *representations languages* while downloading an ontology. **Statement 3**: I look for SPARQL query file, if any. ***Statement 4***: I would like to see *'user reviews'* with these ontologies, so that I can save a lot of time in understanding the quality of the ontology. ***Statement 5***: I prefer to have a *documentation/ information about the methodology* followed to develop ontology, it will be an additional advantage. ***Statement 6***: I remain curious about the following facts: *top classes, number of classes* and *class* definitions. ***Statement 7***: I look for *types* and *number of relations*. ***Statement 8***: I look for *number of entities* and *description* about each of them. ***Statement 9***: I look for whether I can export the whole or part of the ontology, also look for the *languages and formats* to export.	***Statement 10***: I look for whether the *ontology visualization* feature is supported. ***Statement 11***: I look for the *date on which the ontology was created*. ***Statement 12***: I look for if the ontology was created manually or through some kind of corpus mining, i.e. some information regarding *how the ontology was created*. ***Statement 13***: I look for the *person or organization* that has developed the ontology. ***Statement 14***: I look at the *classes and properties* of the ontology as they are very important in scrutinizing a basic evaluation of ontology; especially in those cases where I am searching for ontology of a known field or domain. ***Statement 15***: I usually search ontology by *topic* and then see the *relevant classes*. Sometimes title does not reflect the relevant ontology classes. ***Statement 16***: Before selecting an ontology to download, I make sure it is in OWL, because of my familiarity with this *ontology language*.

4. Top-level Facets

Following the top-down approach, as discussed above, we have derived a set of top-level facets for the ontology vocabulary. The top-level facets provide a high-level schema of the ontology metadata vocabulary expressing the various aspects of an ontology. These facets are further analyzed to define the classes of MOD discussed in section 5.

To derive the top-level facets, we treat the ontology as a study of subject. In other words, an ontology is at the center of our study. We have studied and analyzed an ontology from multiple perspectives. There are a total of seven aspects that have been identified as follows:

- *General*- an abstraction of the general aspects of an ontology, for instance, the ontologies, ontology type, etc.
- *Ontology Coverage*- an aspect that defines the domain (*a domain is any area of knowledge or field of study that we are interested in or that we are communicating about that deals with specific kinds of entities* (Giunchiglia and Dutta, 2011, Giunchiglia, et al., 2014)) and scope of an ontology.
- *Authority*- describes the agents, like organizations, that own and are responsible for the ontology.
- *Rights*- describes the rights and licenses of an ontology.
- *Environment*- defines the environment in which an ontology has been built, for instance, the tool that is used to build an ontology, the level of formality, and the syntax followed.
- *Action*- an aspect highlighting the applications where an ontology is being applied or used, such as in a project.
- *Preservation*- describes the low level-features of an ontology, for instance, ontology storage, file format, etc.

It can be seen from the above descriptions that each of these aspects is complex in nature. We have further analyzed these aspects and have derived the basic classes of MOD as discussed in the following section.

https://doi.org/10.23106/dcmi.952136974

5. MOD Metadata Model

MOD, the metadata vocabulary, consists of 64 elements. These elements are expressed in terms of Classes, Object properties and Data properties. There are 15 classes including two subclasses, 18 object properties and 31 data properties. The further descriptions on the classes and properties are provided below. In the successive sections we also discuss the various controlled vocabularies that are used to create and standardize the MOD vocabulary.

To make the MOD vocabulary interoperable and conform to the major representation languages currently being used for the Semantic Web applications, we have expressed MOD using OWL. The ontology is available at http://www.isibang.ac.in/~bisu/.

5.1 Classes

MOD consists of 15 classes (*a class is a collection of things sharing common attributes*) presented in Table 3 along with some exemplary class instances. Classes are important in metadata vocabulary as they are required to represent and support the reuse of ontologies (Hartmann, et al., 2005). The classes are derived by analyzing the top-level facets described above. For instance, the top-level facet *Authority* refers to a person or an organization who created and/ or who exercises control over an ontology, an ontology document, etc. In MOD both *Person* and *Organization* are considered as classes and grouped under a general class *Agent*. Similarly, by analysing the facet *Environment*, we have derived the classes like *OntologyTool*, *OntologyLanguage*, and *OntologySyntax*. In a similar fashion, we have analysed all the top-level facets and have derived the following classes shown in Table 3. The classes are presented in the table with the corresponding facets.

TABLE 3: MOD ontology classes

Top-level facet	Class Name	Example of Class Instances
General	Ontology	Space ontology, Food ontology, Fishery ontology,
Authority	Agent Subclass : Organization Subclass : Person	Organization related with the ontology and the person associated with it.
Right	License	Creative Commons, GNU Free Documentation License, GNU General Public License
Scope/Coverage	Domain	Genes, Space, Medicine, Protein
	Ontology type	Application Ontology, General Ontology, Core Reference Ontology
Action	Project	Smart city, Mobility
	Methodology	METHONTOLOGY, YAMO
Environment	Ontology design tool	OntoEdit, Protégé, TopBraid composer
	Ontology design language	RDFS, OWL
	Ontology design syntax	Notation3, Turtle, RDF/XML
Preservation	File Format	.rdf, .gaf
	Level Of Formality	Dictionary, Glossary
	Knowledge Representation Formalism	Frame, Description Logics, First Order Logic.

https://doi.org/10.23106/dcmi.952136974

5.2 Object Property

Object property is a property that connects two resources belonging to two different, or the same, classes. MOD consists of *18* object properties including *creator*, *contributor*, *endorsedBy*, *evaluatedBY, module*, *formalityLevel*, *subject*, *usedIn*, etc. Each object property is defined with its domain and range. For instance, the object property *creator* has a domain class *Ontology* and a range class *Agent*. An object property can have more than one domain and range.

5.3 Data Property

Data property is a property that connects a resource to a data type. The data types are literals. MOD consists of *31* data properties, out of which *21* properties are directly the properties of an ontology resource. The other *ten* properties are the properties of other related resources, for instance, an Agent. Some of the data properties are: *name*, *acronym*, *identifier*, *noOfClasses*, *noOfProperties*, *noOfAxioms*, *naturalLanguage*, *lastUpdated*, *version*, etc. Each data property is specified with its domain and range. For instance, the data property *noOfClasses* has a domain class *Ontology* and a range *integer*. A data property can have more than one domain.

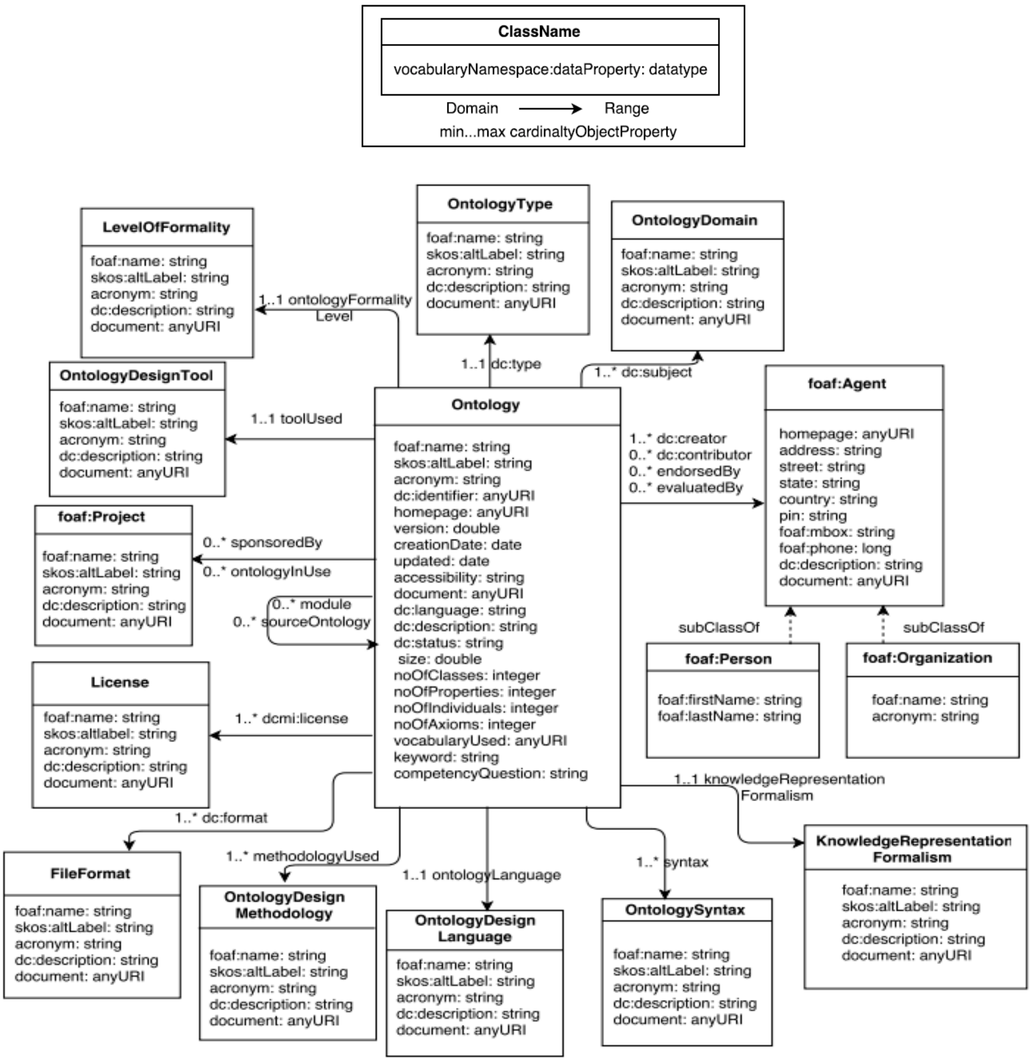

FIG. 1: MOD Overview

https://doi.org/10.23106/dcmi.952136974

5.4 Vocabularies Used

The MOD terms are further standardised by using equivalent terms that are available in the existing metadata standards. Some of the metadata standards that we have used for this purpose are Friend Of A Friend (FOAF) (FOAF, 2014), Dublin Core (DC) (DCMI, 2015), and Simple Knowledge Organization System (SKOS) (SKOS, 2009). This way, MOD not only standardizes the vocabulary, but it also becomes part of the global initiative. This approach also ensures interoperability among the software programs.

Figure 1 above provides an overview of MOD vocabulary in terms classes, data properties and object properties including the constraints on the primary class Ontology. In the figure the prefixes represent the vocabulary namespace URIs. For dc, the namespace URI is http://purl.org/dc/elements/1.1/, for dcmi, the URI is http://purl.org/dc/terms/, for foaf, the URI is http://xmlns.com/foaf/0.1/, for skos, the URI is http://www.w3.org/2004/02/skos/core#.

6. Related Work

Here we will briefly discuss the related metadata standards, especially those that are relevant to the Semantic Web. DC Schema is a vocabulary consisting of a set of terms which can be used for describing web resources (video, images, web pages, etc.), as well as physical resources such as books, magazines, proceedings, journals, CDs, etc. Dublin Core has two sets of metadata, namely, unqualified DC (core elements) and qualified DC. FOAF provides a standard vocabulary to describe people, their activities and their relations to other people and objects. Anyone can use FOAF to describe him or herself. The Organization Ontology (Org, 2014) is a core ontology for organizational structures. It aims to support linked data publishing of organizational information across a number of domains. Its design goals are to allow domain-specific extensions to add classification of organizations and roles, as well as extensions to support neighboring information such as organizational activities. VoID (2011) is an RDF (Resource Description Framework) (RDF, 2014) vocabulary and a set of instructions. It enables the discovery and usage of linked-data sets. RDF Data Cube (2014) Vocabulary provides a means to publish multi-dimensional data, such as statistics, on the web in such a way that it can be linked to related data sets and concepts using an RDF standard. It is a core foundation which supports extension of vocabularies to enable publication of other aspects of statistical data flows or other multi-dimensional data sets.

The above-discussed standards are related to our work in that they are metadata standards to describe the Web resources and are relevant for the Semantic Web applications. However, there is only one work that is very closely related to our work called Ontology Metadata Vocabulary (OMV). It provides a vocabulary for describing the ontologies. The basic differences between MOD and OMV are: MOD provides a minimized and well-defined set of metadata elements, which confirms the *principle of brevity* and *principle of clarity*. MOD elements are mapped and standardised with the other Semantic Web metadata standards. In other words, MOD reuses the existing metadata ontologies, which confirms the *principle of interoperability*. Overall, MOD is a well-guided, refined, easy-to-use standard ontology metadata vocabulary.

7. Conclusion

Metadata is instrumental in finding any kind of resources, whether they are print materials or electronic objects like ontologies, webpages, books, images, audio, video and so forth. Not only does metadata play a role in finding the resources, but can support in decision making to reuse the resources. In this context the current work has significance. MOD can be implemented by ontology libraries, and in general by Web developers, to make an ontology searchable and reusable. In our future work, we plan to pursue the use of MOD in the context of ontology libraries

https://doi.org/10.23106/dcmi.952136974

Acknowledgement

We wish to thank Prof. Fausto Giunchiglia and Prof. Cristina Pattuelli for reviewing the work and providing valuable suggestions. We wish to thank the anonymous reviewers for their time in evaluating our paper patiently and providing critical comments and suggestions, which we have incorporated into this paper. We also thank Prof. A. R. D. Prasad and Prof. Devika P. Madalli for their constant support in conducting the current work.

References

Studer, R. and Benjamins, V.R., Fensel, D. (1998). Knowledge engineering: Principles and methods. In Data and Knowledge Engineering, 25 (1-2), 161-197.

Ding, Y., Fensel, D. (2001). Ontology library systems. The key to successful ontology reuse. In: First Semantic Web Working Symposium, 93–112.

Dutta, B., Chatterjee, U. and Madalli, D. P. (2015). YAMO: Yet Another Methodology for Large-scale Faceted Ontology Construction. Journal of Knowledge Management. 19 (1), 6 – 24.

d'Aquin, M., Noy, N.F. (2012). Where to publish and find ontologies? A survey of ontology libraries. Web Semantics: Science, Services and Agents on the World Wide Web, 11, 96–111.

Obrst, et. al. (2014). Semantic web and big data meets applied ontology. Applied Ontology, 9, 155-170.

Hartmann, Jens et al. (2005). Ontology metadata vocabulary and applications. In Proceedings of OTM Workshop (2005), LNCS 3762, 906–915.

Debashis, N.. (2014). Ontology and ontology libraries: a critical study. In Master's Dissertation (carried under the supervision of Biswanath Dutta). Bangalore, India: DRTC, Indian Statistical Institute, 10-49.

Zubeida C. Khan and Keet, C. M. (2013). The foundational ontology library ROMULUS. Model and Data Engineering. LNCS, 8216, 200-211.

Berners-Lee, Tim, Hendler, James and Lassila, Ora. (2001). The Semantic Web. Scientific American, 29-37.

Giunchiglia, F. and Dutta, B. (2011). DERA: a Faceted Knowledge Organization Framework. Available at: http://eprints.biblio.unitn.it/archive/00002104/

Giunchiglia, F., Dutta, B. and Maltese, V. (2014). From Knowledge Organization to Knowledge representation. Knowledge Organization, 41, 1, 44-56.

OWL. (2004). OWL Web Ontology Language Overview. Retrieved July 9, 2015, from, http://www.w3.org/TR/owl-features/.

FOAF. (2014). Vocabulary Specification 0.99. Retrieved July 9, 2015, from http://xmlns.com/foaf/spec/.

DCMI. (2014). Dublin Core Metadata Element Set, Version 1.1. Retrieved July 9, 2015, from http://dublincore.org/documents/dces/.

SKOS. (2009). SKOS Simple Knowledge Organization System Reference. Retrieved July 9, 2015, from http://www.w3.org/TR/skos-reference/.

Org. (2014). The organization ontology. Retrieved July 9, 2015, from http://www.w3.org/TR/vocab-org/.

VoID. (2011). Describing Linked Datasets with the VoID Vocabulary. Retrieved July 9, 2015, from http://www.w3.org/TR/void/.

RDF. (2014). RDF 1.1 Concepts and Abstract Syntax. Retrieved July 9, 2015, from http://www.w3.org/TR/rdf11-concepts/.

Data Cube. (2014). The RDF Data Cube Vocabulary. Retrieved July 9, 2015, from http://www.w3.org/TR/vocab-data-cube/.

https://doi.org/10.23106/dcmi.952136974