ISSN: 2321-7782 (Online)
Volume 3, Issue 6, June 2015
International Journal of Advance Research in Computer Science and Management Studies
Research Article / Survey Paper / Case Study
Available online at: www.ijarcsms.com

Representation of Legacy Bibliographic Data in Food Science for Semantic Web Applications

T. Padmavathi[1]
Food Science & Technology Information Services (FOSTIS/Library)
CSIR-CFTRI, Mysore, India

Dr. M. Krishnamurthy[2]
Documentation Research & Training Centre (DRTC)
Indian Statistical Institute
Bangalore, India

Abstract: ***Food Science database of FOSTIS consists of bibliographic records of publications involving journal articles from various Indian and international journals created by the staff of the library. We would like to discuss the experiences gained in the process of this data representation from its current data format to RDF/OWL-based data. During the course of representation we will specify the entity types in our data and choose a way to model them. For data description we are going to use the most popular and widely implemented vocabularies and domain ontologies. In this way, aligned with the concept of the semantic web, we maintain that the potential of the data would be maximized, as the information in bibliographic records would become easy to share, more visible due to incoming links, and more ready to be processed by web applications. The adoption of semantic web/linked data publishing model for bibliographic data involves moving to a more flexible data format for bibliographic data. While linked data is seen as a pragmatic implementation of the semantic web, we differ that it is not possible to implement the library data vision with the MARC, current standard for bibliographic data. Obviously, there is a need for more web-compatible and web-friendly data model. With this in mind, we proposed linked data as a part of the pragmatic implementation of the vision for bibliographic data on the Web.***

Keywords: bibliographic ontology; legacy bibliographic data; linked data publishing; semantic web technologies; ontology; food science

I. INTRODUCTION

Libraries have a long tradition of data exchange and interoperability, with the use of the Marc formats since the early seventies. But this format is not sufficient to express all the advantages of the data, and cannot be integrated with the Web. The bibliographic data lacks an agreed-upon model and the libraries seems to focus more on the formats than on an core model that can be expressed in different ways. A move to a more model-driven view on bibliographic information would increase the possibilities to interlink the individual parts of a bibliographic record to other entities outside of the library domain particularly within the food science domain. In order to make this bibliographic data openly available, more flexible, easy to share and visible, to increase its suitability for web applications, we adopted semantic web technologies/linked data publishing model for Food science database of FOSTIS.

The adoption of the linked data publishing model proved beneficial for the library and the wider linked data community. In this way, bibliographic data can be available in a data format that is familiar to the experts outside of libraries: the RDF. The libraries play an important role on the web, as the data produced by libraries, archives, digital repositories or academic institutions tend to be maintained by trained professionals and the data outputs are therefore of high quality. These data then have a potential a much-needed backbone of trust for the growth of the so-called semantic web. One of the basic building blocks of the semantic web is the Resource Definition Framework (RDF) (http://www.w3.org/TR/2 004/REC-rdf-concepts-20040210/) that serves for knowledge representation and storage of the represented data. While HTML documents contain information available for human to understand, with the use of RDF the structure of information can be parsed, re-constructed

and processed also by web applications in a more straight-forward manner. RDF data format provides a simple way to connect information resources located in separated datasets by the interlinking of related pieces of information across the Web, from one dataset to another. In this work, we consider reusing the standard vocabularies for the vision towards making the dataset a part of the web of data. In the following sections, we will discuss the related research, ontology design process of representing the legacy data to fully-fledged linked data that is ready to be published on the Web.

II. RELATED RESEARCH

In 1998, Weinstein [1] generated a knowledge base of metadata from a sample of Machine Readable Cataloguing (MARC) records. He implemented the ontology in description logic, a knowledge representation language, and mapped MARC attributes and values to the ontology. In this way, ontology was adapted to describe metadata.

ABC ontology and its model were developed within the Harmony international digital library project in Cornell University. It was designed to offer a common model that would facilitate the ability to exchange metadata ontologies from different domains Lagoze & Hunter [2]. One of the results of the project was a metadata model with more logically described time and entity semantics. Based on this model, a metadata repository of RDF descriptions was built with a search interface on top.

Sure and Iosif [3] compared two ontology-based search tools with a typical keyword-based search tool in terms of search time, mistake-making, usefulness, and the development and maintenance of tools. The first ontology based search engine was used to obtain a picture of the information available in the system knowledge base while the second ontology-based search tool was used to give users semantic context. The search result revealed that the ontology based tools are generally at least as good as the keyword based tool and, to some extent, even superior.

Muller et al [4] have developed Textpresso, a new text-mining system for scientific literature whose capabilities go far beyond those of a simple keyword search engine. Textpresso's two major elements are a collection of the full text of scientific articles split into individual sentences, and the implementation of categories of terms for which a database of articles and individual sentences can be searched. Their ontology comprised of 33 categories of terms. They designed a search engine which enables the user to search for one or a combination of tags and/or keywords within a sentence or document, and as the ontology allows word meaning to be queried, it is possible to formulate semantic queries. Extraction of particular biological facts was accelerated significantly by ontologies.

One of the first metadata applications in the domain of Agriculture, based on Dublin Core, was the AGRIS application profile AGRIS AP [5]. The AGRIS Application Profile is an XML-based bibliographic metadata exchange format that allows sharing of information across dispersed bibliographic systems. This is a major step towards exchanging high-quality and medium-complexity metadata in an application independent format and provides possibilities to offer value-added services, irrespective of how the information was stored locally. The AGRIS Application Profile was used as model for the underlying bibliographic ontology design.

Kim [6] described the design and implementation of an ontology-based Web retrieval (ONTOWEB) system. ONTOWEB allows the semantic search of the Web resources of international organizations such as the World Bank and the Organisation for Economic Co-operation and Development (OECD). The ONTOWEB system has two components: databases and an ontology-based search engine. The ontology-based search engine is a tool used to query the information that has been loaded into the database. In order to evaluate the system, they conducted an experiment to compare the performance of the proposed system with that of Internet search engines in terms of relevance and search time. Their study showed that ontologies can be used not only to improve precision, but also to reduce the search time.

Lim et al [7] propose an ontology-driven knowledgebase specific for Korean mitochondrial single nucleotide polymorphism (SNP). The new ontology is developed by OWL and it is designed to provide enough expressivity for the description of the concepts and properties of the SNP-gene –disease relationship. Using an inference engine, the logical

inconsistency in ontology was readily detected and new facts were inferred. Finally, complex queries are supported through the user interface.

Sini et al [8] described the work done within the Food and Agriculture Organization of the United Nations (FAO) on providing an ontology-based navigation for the Food, Nutrition and Agriculture (FNA) Journal. Their aim was to provide navigation with more efficient and effective browsing of the Food and Nutrition Publications using a knowledge model to guide the user with concepts and relationships relevant to a specific subject area. In their study, data from two different bibliographical databases was merged, unified and presented to the user with improved services. A preliminary metadata merge was needed to combine all the information into one system in order to produce a metadata-ontology. Resource Description Framework Schema (RDFS) was chosen to exploit semantic relationships and the creation of a multilingual concept-based advanced search.

Noh [9] presented the construction of an ontology targeting scholarly journal articles and evaluated its performance. They used Protégé to construct an ontology and utilized the inverted file index to compare the performance. The concept ontology was manually established and bibliography ontology was automatically constructed to produce an OWL concept ontology and OWL bibliography ontology. They compared performance of the knowledge base of the ontology, using the Jena search engine with the performance of an inverted index file using the Lucene search engine.

Costa et al [10] explores the use of complex relationships (otherwise known as Semantic Associations) available in ontologies with the addition of information presented in documents. They introduce a conceptual framework and its current implementation to support the representation of knowledge sources, where every knowledge source is represented through a vector (named Semantic Vector - SV). The novelty of their work addresses the enrichment of such knowledge representations, using the classical vector space model concept extended with ontological support, which means to use ontological concepts and their relations to enrich each SV. Their approach takes into account three different but complementary processes using the following inputs: (1) the statistical relevance of keywords, (2) the ontological concepts, and (3) the ontological relations.

Reyes-Ortiz et al [11] proposed a model of representation of the ontology-based information to make it manageable for an expert system that supports decision-making in medical diagnosis. Their representation is based on ontologies that provide a mechanism for structuring knowledge to become computer-understandable information, shared by information systems, formalized and using a common vocabulary. Their ontological model is able to infer a list of clinical diagnoses from the data of signs, symptoms, risk factors and medical background.

III. Ontology Design Process

In this study, we adopted Methontology which is most widely and generally used methodology for building an ontology. This methodology includes the phases of specification, knowledge acquisition, conceptualization, formalization, implementation and evaluation was essential to achieve the objective of research.

A. Ontology Specification

Food Science and Technology (FST) are a multidisciplinary areas with several core and peripheral areas. In this regard, building an ontology for the FST field would enable researchers, scientists and other technical staff to semantically search the FOSTIS knowledge base (FSWKB). This study targeted the Food science database of FOSTIS which focuses on FST research consisting of bibliographic records of publications involving journals articles from various Indian and international journals. The database includes the articles' bibliography, author, contents, abstract and links to full text information. The database is updated continuously and it contains records available online with retrospective since 2004 (see fig. 1 and 2).

B. Knowledge acquisition

Bibliographic ontology is to model and to describe the bibliographic data and its related component parts semantically. In order to conceptualize the ontology, domain analysis was performed to consider reusing existing ontologies and to acquire knowledge from the domain as the first step. In order to standardise vocabulary some of the well-established ontology specifications, namely, KA2 initiative, ONTOWEB, Marcont (http://marcont.corrib.org/), Biblioontology (http://bibliontology.com/), Dublin Core and FAO journal were studied. Taking into account the specifics of our dataset and reflecting the types of information contained in it, we have considered the use of Bibliographic Ontology (bibo). <http://purl.org/ontology/bibo/> Specification licensed under a Creative Commons License which acts as the basis for constructing ontologies. This specification has been inspired by many existing document description metadata formats namely Dublin Core and FOAF (from "friend of a friend") <http://xmlns.com/foaf/0.1/> is an RDF based schema to describe authors in a semantic way. Bibliographic Ontology relies heavily on W3C's RDF technology, an open Web standard that can be freely used by anyone. The ontology schema was modified to fulfil architectural and system requirements.

C. Conceptualization

There are eleven specific tasks during the conceptualization stage to develop a conceptual model of the ontology. We adopted only selected tasks which are core expressions for conceptualization. The activities we have identified for the domain are:

1. Build Glossary of Terms
2. Build Concepts Taxonomies
3. Build Binary Relation Diagrams
4. Build Concept Dictionary
5. Describe Binary Relations
6. Describe Instance Attributes
7. Describe Class Attributes
8. Describe Constraints
9. Describe Formal Axioms
10. Describe Rules
11. Describe Instances

1. Glossary construction: We enumerated glossary of terms from some of the well-established ontology specifications. The glossary includes all types and forms found in and belong to the domain. We arrived at 50 terms viz., resource, person, language, organization, publisher, article, book, book chapter, title etc. Table 1 shows the glossary that is relative to bibliographic information class.

TABLE I

Glossary of Bibont

Class Name	Subclass	Meaning
Collection	Periodical (which includes books, journals and thesis)	collection of documents or collections
	Series (A loose, thematic, collection of Documents, often	

	Books.)	
	Website (A group of Webpages accessible on the Web.)	
Document	Article (A scholarly academic article, typically published in a journal.)	bounded physical representation of body of information designed with the capacity to communicate.
	Book (A written or printed work of fiction or nonfiction, usually on sheets of paper fastened or bound together within covers.)	
	Collected Document (A document that simultaneously contains other documents which includes Edited Books, Issues etc)	
	Document Part (a distinct part of a larger document or collected document.)	
	Manuscript (An unpublished Document, which may also be submitted to a publisher for publication.)	
	Patent (A document describing the exclusive right granted by a government to an inventor to manufacture, use, or sell an invention for a certain number of years.)	
	Reference Source (A document that presents authoritative reference information)	
	Thesis (A document created to summarize research findings associated with the completion of an academic degree; which includes Master Degree thesis as well as Ph.D thesis)	
Document status		The status of the publication of a document
Organization	State Government, Central Government or private organization	The status of the publication of a document
Person	-	-
Resource		-

2. Concept classification tree: We selected terms that describe objects having independent existence rather than terms that describe these objects. These terms are classes in the ontology and are anchors in the class hierarchy. Most of the remaining terms are properties of these classes.

3. Building the concept lexicon: This process aims to express all concepts included in the domain to be built, including instances of class, concepts of class, property of instances and relations between concepts.

4. Building the class property table: The class properties expressed in the concept lexicon is described. Here we used the class property of Bibliographic Ontology Specification and was modified to suit our system requirements.

5. Instance property: The properties of instances expressed in the concept lexicon are described. Here we used the instance property of Bibliographic Ontology Specification and was modified to suit our system requirements.

6. Relations: In this stage, the relation between class and instance is described. An example of relation name for the publication class is depicted in the figure 1

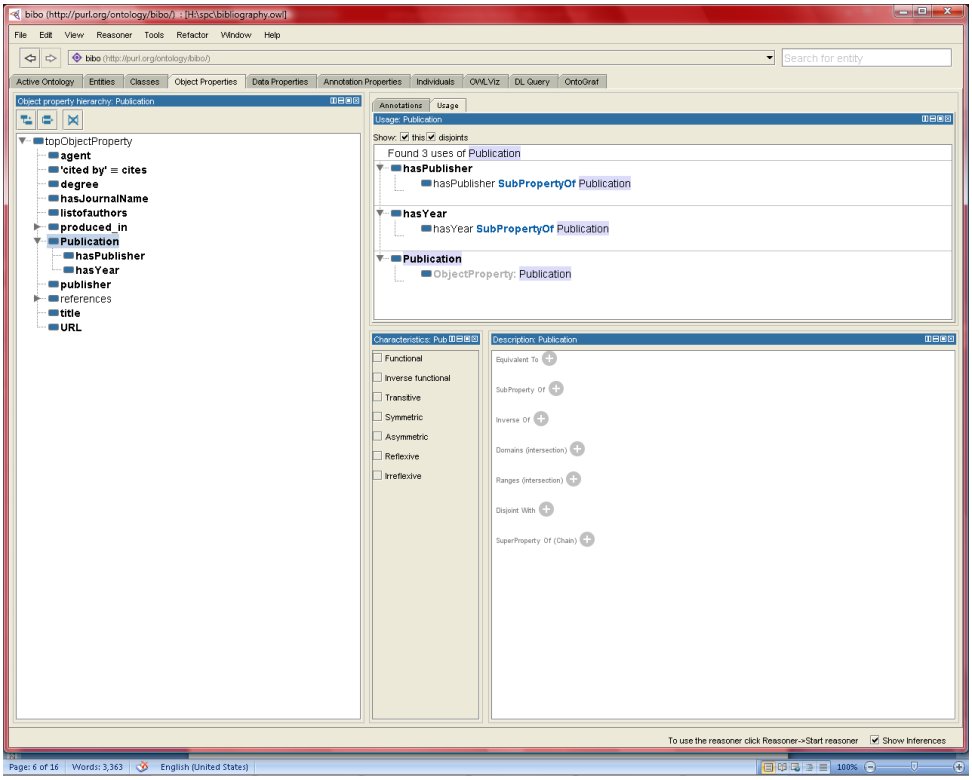
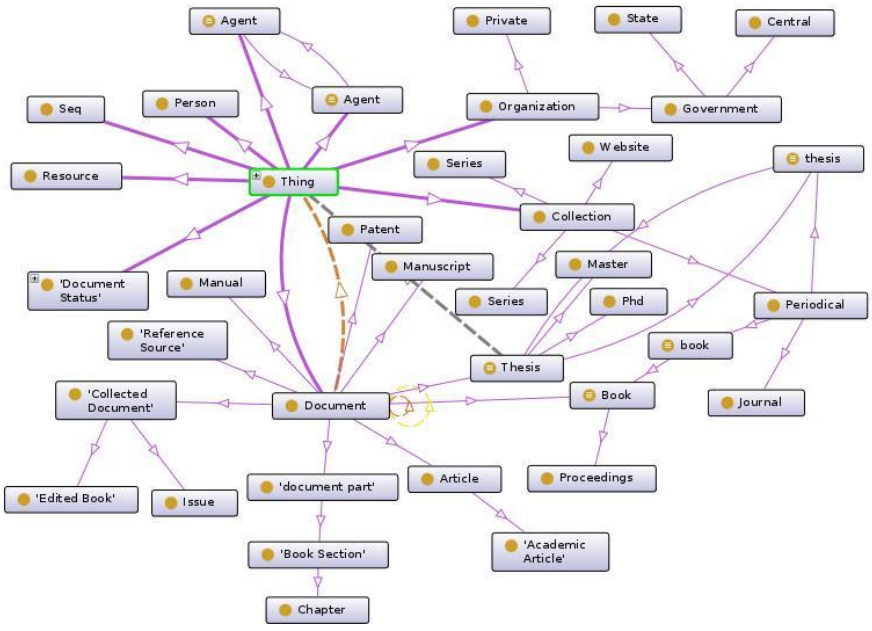


Fig. 1: Relation Name of the Class Publication

D. Formalization

The knowledge model is formalized by using Protégé the ontology editor. Protégé is an open source software developed by Stanford University which helps to convert our formal model into an OWL-DL. OWL supports graphical representation of a class hierarchy though OWLVIZ plug-in. This visualization function helps developers and users understand the structure of the ontology more easily than merely showing a text-based ontology structure. Bibliographic ontology consists of six classes.



E. Implementation

Ontology is a controlled vocabulary and a formal language that means that knowledge can be expressed in a way that is computer interpretable. Top-down approach has been chosen for definition of the most general concepts in the domain for hierarchy creation. The bibliographic ontology (bibont) describes bibliographic things. It provides a rich hierarchy of associative relations used to define a complex structure. The main components of the bibont are classes such as Collection, Person, Document, Agent, Resource and Organization - these classes are objects that abstract the real world. The second level in the ontology subclass represents more details of the superclass, such as Article, Periodical, Author and Private. Additionally, the bibont provides a rich hierarchy of associative relations defining a complex structure, as can be seen in figure 3.

In this research many classes and sub classes have been defined for the bibont. It is categorized by 6 levels. There are 148 classes, sub classes 84, axiom (entities) 387, object properties 43, data properties 16 providing the capability for rich semantic expression.

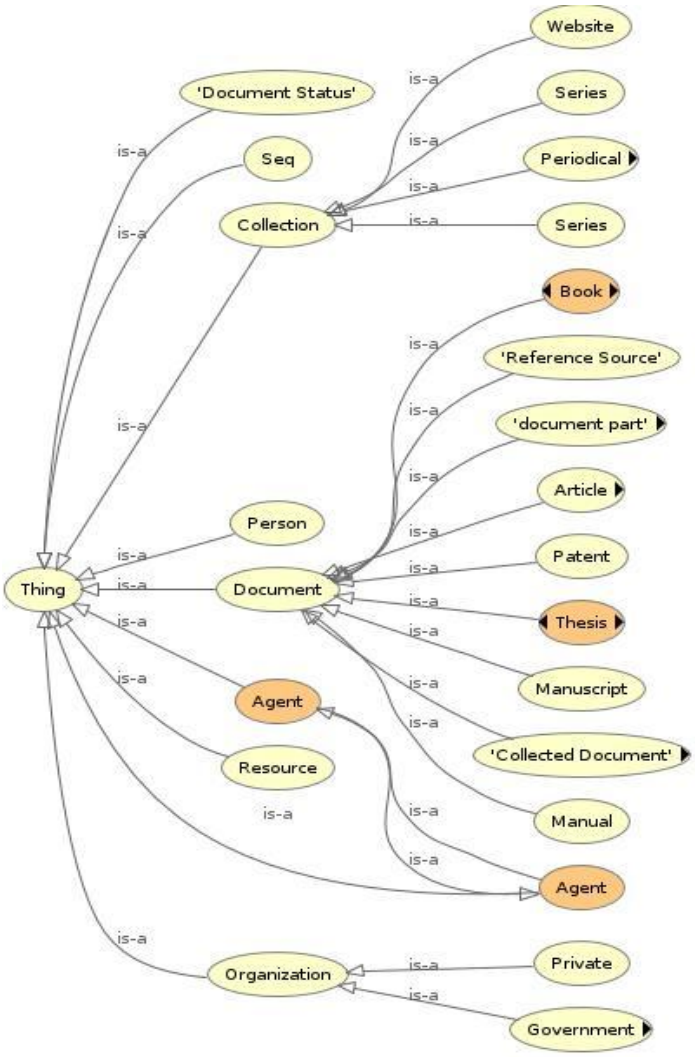


Fig. 3 Hierarchy of Associative Relations of Bibont

The second component is properties. It is also called relationships in OWL - between superclasses and subclasses; data properties and object properties are used. In bibont ontology two types of properties are used to link between two classes and linking between general classes and their instances. These relations have been used <Is-A>, <is Part Of>, and <has A>. These relations are created in the Object property tab, as shown in Figure 3. The object properties defined are: hasJournalName(Name of the journal in which the article has been published), citedBy(citations used by the other authors), presentedAt(Institution where the paper was presented), hasPublisher(name of the publisher), hasYear(year of publication), title (title of the article) and

URL (URL of the article). And also the data properties such as date, doi(date of Issue), eISSN(Electronic ISSN number), isbn, name, number of pages, volume etc. adding constraints to defining classes.

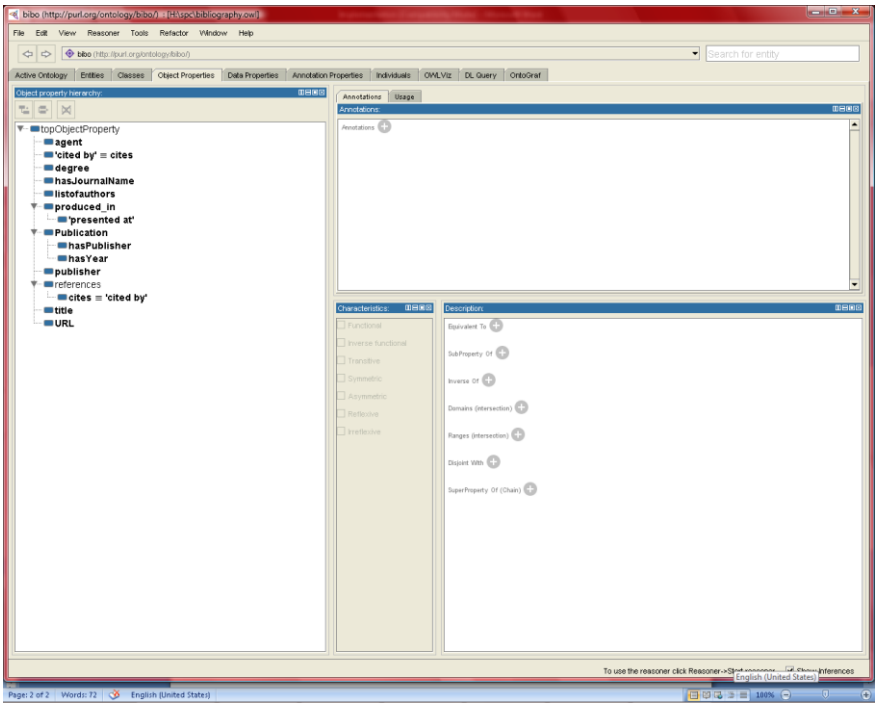


Fig. 4 Object Properties

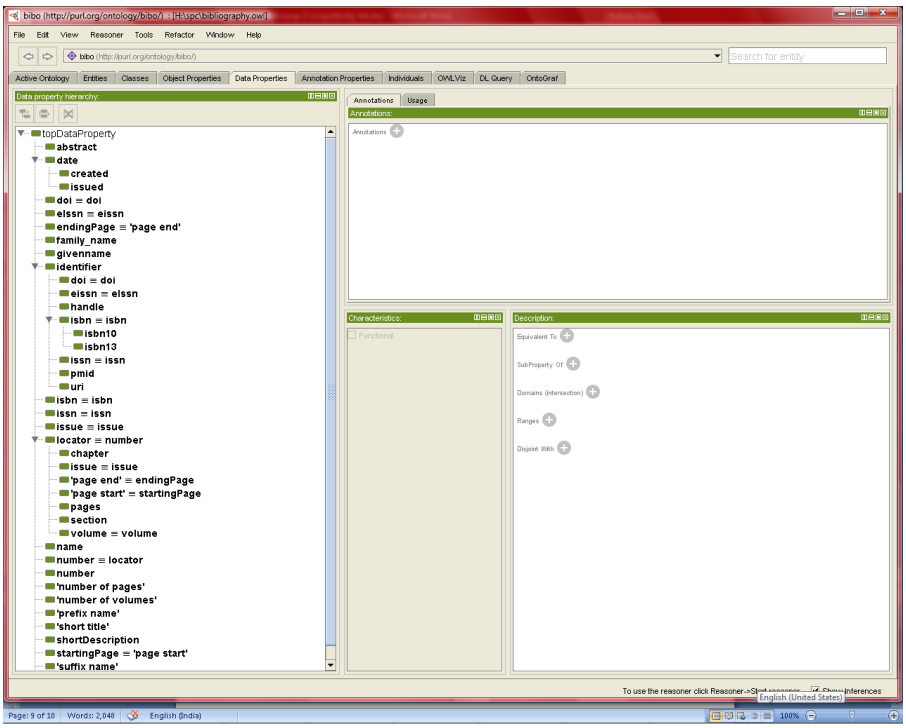


Fig. 5 Data Properties

The usage of classes and individuals in bibont can be checked through using the usage class feature in Protégé, for example there have been fifteen usages of the class Collection in bibont. Four usages are of subclass and 11 usages are of Domain. See Figure 6

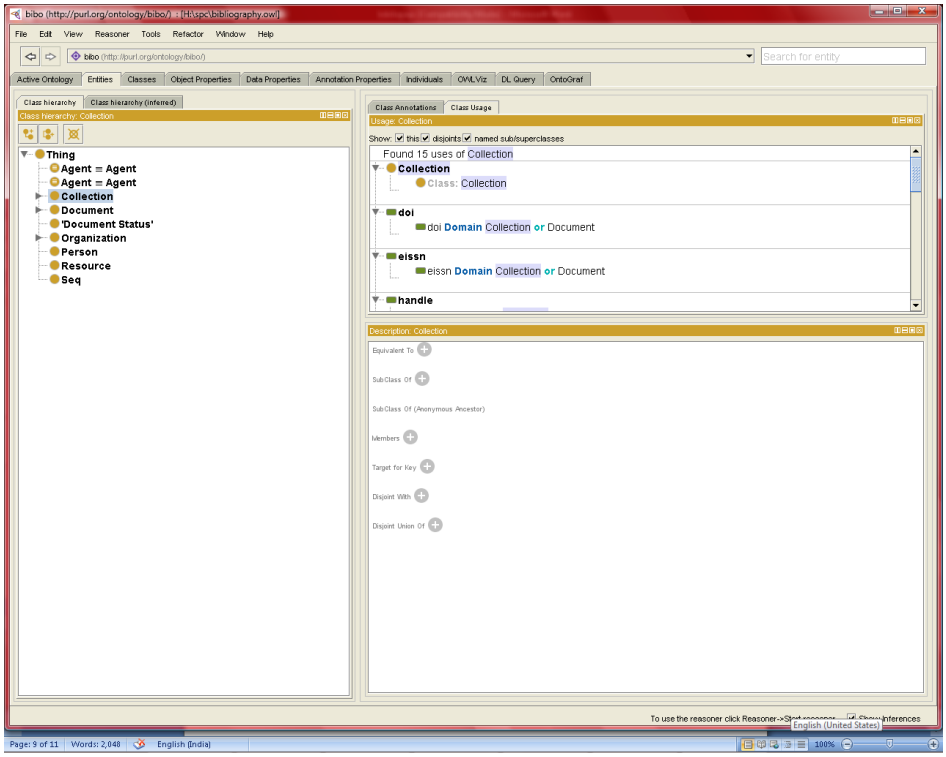


Fig. 6 Class Usage in Bibont

F. Data Structure

Experimental data is composed of 250 journal articles from FOSTIS food science database. We have used protege to add the values to the bibont. In this, the individuals has to be added in such a way that no linking conflict should arise. An individual will have many attributes and properties which describes its identity. All these are incorporated while adding the individuals. The example below shows bibliography information for individual article of the journal articles.

TI: Dietary fiber in the prevention and treatment of metabolic syndrome: A review.

AU: Aleixandre, A.; Miguel, M.

PY: 2008

SO: CRC Critical Reviews in Food Science and Nutrition, 48(10) 905-912

AB: This review contains several ideas about the possible benefits of dietary fiber intake in patients with metabolic syndrome. The principal beneficial effects of a fiber-rich diet in these patients are: prevention of obesity, improved glucose levels, and control of the profile of blood lipids. We now also know that dietary fiber may favor the control of arterial blood pressure. Animal experiments have also shown the benefit of different types of fiber on these variables. Of particular relevance are the studies using obese Zucker rats, which present similar anomalies to those seen in patients with metabolic syndrome. There is therefore a growing interest in discovering new sources of natural fiber. Some of these different kinds of fiber may then be used as functional ingredients to obtain foods with properties that are beneficial to health.

G. Evaluation

An evaluation could look at the terminology used from a technical or expert perspective, to ensure terms are defined accurately and the correct methods have been used during the development. Ontology evaluation should also ensure that we avoid concepts duplication, excessiveness and inconsistent relationships to make our understanding simpler. The evaluation can be done at the developing process and running time. At the development process the interim evaluation is applied to ensure its completeness and consistency and to improve it. Ontology consistency can be checked by using a reasoner such as FaCT++

Pellet, or Hermit. The reasoner is one of the main advantages of using a logic-based language such as OWL–DL, which is supported by Protégé4. The reasoner can be used at the point of developing ontology, publishing time, and run time in applications as a querying mechanism.

In Bibont the consistency was checked through the reasoner. Fact++ has been used for reasoning the ontology. The mistakes committed during ontology construction are spotted out by the reasoner. The problem that is faced when the ontology was developed is, due to a wrong setting of property characteristics, the reasoner displays error messages like inconsistent ontologies. The error has been occurred, if a class is incorrectly classified it will appear in red in a root class called Nothing. Reasoning capabilities are exploited to detect logical inconsistencies within the ontology. The consistency checks can help in an adequate manner while constructing the ontologies.

IV. CONCLUSION

In this paper we presented the development of the bibliographic ontology (bibont) which provides a backbone to represent, collect, share and allow inference from Food science database of FOSTIS in a logical way. The advent of the internet has carved numerous pathways for the users to find and discover information. The searches made by the users have evolved over a period of time from browsing just bibliographic data to find a print publication that will enable them to order from a publisher or a research library to simplifying the search results to open a full text journal article and whatever else will enable them to answer their queries. Thus the conventional databases need to excogitate new ways to address such concerns. In order to overcome this problem and make the bibliographic data openly available, we adopted semantic web technologies/linked data publishing to satisfy users' basic information needs and the services to meet their scientifically meaningful information needs. The user's requisite mainly spreads to cover semantic search and recovery using the facts stored in the knowledge database and, while the knowledge representation exhibit the usage of the ontology to uncover the hidden knowledge for the academic community. Describing bibont in a unified model helps to improve information accessibility through greater semantic interoperability of information. It also makes it possible to build a scholarly semantic web by integrating data through relationships with other scientific data on the web thus creating added information.

References

1. P. C. Weinstein, "Ontology-Based Metadata: Transforming the MARC Legacy," in Proc. of the Third ACM International Conference on Digital Libraries, 1998, p. 254-263. Retrieved January 23, 2015, from Citeseer Web site http://citeseer.ist.psu.edu/212498.html.
2. C. Lagoze and J. Hunter, "The ABC Ontology and Model," Journal of Digital Information, vol.2 no.2, article no. 77, pp. 160-176, 2001. Retrieved January 23, 2015, from http://jodi.ecs.soton.ac.uk/Articles/v02/i02/Lagoze/lagoze-final.pdf.
3. Y. Sure and V. Iosif, "First Results of a Semantic Web Technologies Evaluation," Common Industry Program at the federated event: ODBASE'02 Ontologies, Databases and Applied Semantics, 2002, California, Irvine.
4. H.M. Muller, E. K. Eimear and W. S. Paul, "Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature," PLoS Biology, vol.2 no.2, pp. 1984-1998, 2004.
5. AGRIS AP: The AGRIS Application Profile for the International Information System on Agricultural Sciences and Technology Guidelines on Best Practices for Information Object Description, 2005. Retrieved February 20, 2015, from http://www.fao.org/docrep/008/ae909e/ae909e00.htm
6. H.H. Kim, "ONTOWEB: Implementing an Ontology-Based Web Retrieval System," Journal of the American Society for Information Science and Technology, vol.56 no.11, pp.1167–1176, 2005.
7. J. Lim, A.Yoon, C.H. Sun, "OntoSNP: Ontology Driven Knowledgebase for SNP," International Conference on Hybrid Information Technology (ICHIT'06), 2006.
8. M. Sini, G. Salokhe, C. Pardy, J. Albert, J. Keizer and S. Katz, "Ontology-based Navigation of Bibliographic Metadata: Example from the Food, Nutrition and Agriculture Journal," ARD Prasad & Devika P. Madalli Eds.: ICSD-2007, pp. 64-76, 2007.
9. Y. H. Noh, "A Study on Constructing the Ontology of LIS Journal," Journal of the Korean Society for information Management, vol.28 no.2, pp. 177-193, 2011.
10. R.Costa, "Capturing Knowledge Representations Using Semantic Relationships: An Ontology-based approach," SEMAPRO 2012: The Sixth International Conference on Advances in Semantic Processing, pp. 75-81, 2012.
11. J. A. Reyes-Ortiz, A. L. Jiménez, J. Cater, C. A. Meléndez, P. B. Márquez, M. García, "Ontology-based Knowledge Representation for Supporting Medical Decisions," Research in Computing Science, vol.68, pp. 127–136, 2013.
12. J. Hladka, J. Mynarz, J. and V. Sklenak, "Experience with transformation of bibliographic data into linked data," Journal of Systems Integration, vol.3 no.1, pp. 54 – 62, 2012. Available at: http://www.si-journal.org.

13. F. S. Ahlam and J. Lu, "Building Information Science ontology (OIS) with Methontology and Protégé," Journal of Internet Technology and Secured Transactions (JITST), vol.1 iss.3/4, pp. 100-109, 2012.

14. L. Gomez-Perez and Juristo, "METHONTOLOGY:from ontological art towards ontological engineering,". In AAAI Symposium on Ontological Engineering, Stanford, pp.33-40, 1997.

AUTHOR(S) PROFILE

Ms T. Padmavathi is presently working as Senior Technical Officer in Central Food Technological Research Institute, FOSTIS/Library. She holds B.Sc, MLISc and has more than 20 years of experience. Her areas of interest include: Semantic web technologies, digital libraries, library automation and digitisation

Dr M. Krishnamurthy is presently working as Asst. Professor, Documentation Research and Training Centre, Indian statistical Institute, Bangalore. He has teaching experience of 10 years and has published over 50 papers in various journals, and conferences. He is a full bright fellow and has to his credit many awards. He has also visited many countries worldwide for delivering lectures and attended conferences

