The role of Domain Ontology in Semantic Data Integration

KEYWORDS: domain ontology,ontology construction, ontology spectrum, semantic data integration

Rohini R. Rao

Dept of M.C.A., M.I.T., Manipal

ABSTRACT *Data is available in different forms and in different locations. Adopting a semantic representation makes the data independent of its internal form and platform. The semantic representation of data enables data integration in any domain. The steps to be performed in order to construct an Ontology is discussed. An example which illustrates how semantic data integration can be achieved is discussed. Thus semantic data representation will help achieve Tim-Berners Lee's vision of a Semantic Web*

I. Introduction

In many domains such as business, manufacturing, health Care etc., effective information technology adoption is important. For example in a domain such as Health Care, the data is available in various Information Systems and in different data formats, and in various computers all over the world. A person could have many health problems for which the patient could consult a physician, specialist or health care worker. These health care providers could be in various hospitals, health care centers or in clinics. During every encounter with a health provider, information about the patient, such as health condition, laboratory & radiology reports, prescriptions etc., is recorded in electronic form. The data itself may be in flat files, relational format or some kind of semi-structured form on the internet. If each of this data is to exist in separate data silos and in different data formats, each will provide an incomplete view of the patient's condition. There is a need to provide an integrated view of this data to enable effective usage of this data. Doctors can make better health care decisions if they have an integrated view of all the relevant patient data.

Semantic Web technology helps to achieve this data integration [3]. In the first step, the data in various formats is mapped onto a data schema called as Ontology. Semantic Web Technologies like RDF Schema (RDFS) and OWL can be used to represent the Ontology or the data schema. The Ontology instances or the data itself can be represented using RDF and XML. This data representation makes the data independent of its internal form.

Ontology is a branch of metaphysics and it is the study of nature of existence. The most popular definition is that "Ontology is a formal, explicit specification of a shared conceptualization". Ontology is used often used to model domain knowledge. Domain Ontology is a description of a particular domain of discourse. It is described in terms of a finite list of predefined, reserved vocabulary of terms to define concepts (classes of objects) and the relationships between concepts for a specific domain [1,2]. The level of complexity of an Ontology that can be achieved is described through the Ontology spectrum.

II. The Ontology Spectrum

All the concepts in the ontology spectrum address issues in representing, classifying and disambiguating semantic content (as seen in fig 1). Ontology can range from the simple notion of a vocabulary, taxonomy to a thesaurus, conceptual Model and finally to a Logical Theory [2].The spectrum is able to classify the various concepts in terms of increasing semantic richness where the upper right half more appropriately represents Ontology.

At the lower end of the spectrum is Vocabulary, which is a collection of unambiguously defined terms used in communication which have a consistent meaning in all contexts. Taxonomy is a vocabulary in which terms are organized into a hierarchical manner into a tree structure. For semantic applications the information entities are classified into a form of a hierarchy according to the presumed relationships of the real-world entities that they represent. As one goes up the hierarchy towards the root the entities are more general and lower down the hierarchy they become more specialized.

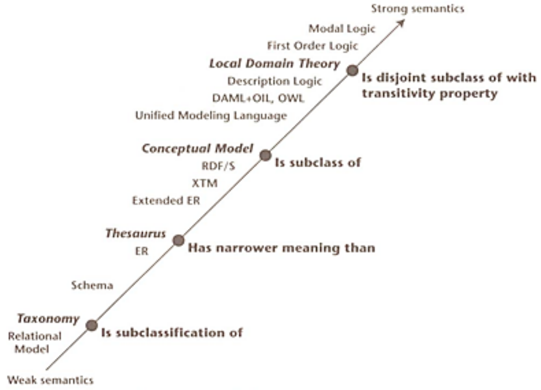

Fig 1: Ontology Spectrum [2]

This type of classification is also called generalization/specialization taxonomy. The common use of taxonomies is to browse or search for information. For example the Yahoo and Google taxonomies assist a user looking for content by categorizing that content as semantically as possible. The Thesaurus is a controlled vocabulary that is arranged in a known order and structured so that relationships like equivalence, homography, hierarchy and association are displayed clearly. The thesaurus is designed to support information retrieval by guiding both the person assigning Meta data and the searcher to choose the same terms for the same concept. A thesaurus ensures that concepts are described in a consistent manner; experienced users are able to easily refine their searches to locate the information they need and that users need not be familiar with technical or local terminology. One such Thesaurus is WordNet . A word in WordNet , has the following information associated with it: Synonyms to imply words that have the same meaning; Hypernyms to identify nodes that are in the parent of the taxonomy and hyponyms to indicate child nodes. Distinct words or phrases called terms, that are synonymous, are roughly at the same level of abstraction are grouped together into a synset. The synset acts as a concept in the knowledge representation and the term is a label or string representation for the underlying meaning. Concept has attributes, attribute values and relationships to other concepts that the concept participates in.

A Conceptual model is a model of a subject area called domain, which represents the primary entities, relationship between entities, attribute & attribute values, relationships and sometimes rules that associate all of them. Local Domain Theory is a model that needs to share knowledge in a specific subject area like medicine, social network etc. using ontologies. Ontology defines the common words and concepts (meanings) that are used to describe and represent an area of knowledge and so standardizes the meaning. One can express the semantics of a model to the highest degree possible with a logical theory which is an ontology that is directly semantically interpretable by software. Logical theories are built on axioms which are statements that are asserted to be true and inference rules that are used to prove theorems about the domain represented by the ontology as logical theory. The whole set of axioms, inference rules and theorems together constitute the logical theory.

Therefore Domain Ontology defined the common words and concepts used to describe and represent an area of knowledge and thus standardize the meanings. Ontology includes the following: Classes in the domain of interest, instances which are particular things, properties, relationship among instances, functions and processes involving instances, constraints on and rules involving instances. Ontology can also be classified as Upper and Lower ontology based on their scope [3]. Upper Ontology such as Dublin Core and Word Net, also known as foundation ontology is generally applicable across a wide range of domains. Lower Ontology such as the Gene Ontology, are also known as domain specific ontology. To achieve data integration, one needs to convert the data schema from its current form to an Ontology.

III. Ontology Construction

Ontology engineering studies the methods and methodologies for building ontology which are formal representations of a set of concepts within a domain and the relationships between those concepts [5,6]. Ontology can be created completely or learnt from textual data, semi structured data or even structured data such as relational databases.The development process is not a linear process and the steps will have to be iterated and backtracking to earlier steps may be necessary at any point in the process. In the first step, Ontology's scope can be defined by considering facts like domain for which it is to be used, the type of queries to be answered, further extensions anticipated etc. Nouns form the basis for class names and verbs form the basis for property names. After the identification of relevant terms, these terms must be organized in a taxonomic hierarchy in either top-down or bottom-up fashion. In the next step, the properties for classes are identified. The semantics of subclass requires that properties are attached to the highest class in the hierarchy to which they apply. The previously defined properties are enhances with facets such as Cardinality, Requiring values and Relational characteristics. If the data is in relational form, it is vital that the primary key of each table is converted into distinct URI or IRIs. After this step in the ontology construction process, it is possible to check the ontology for internal inconsistencies such as incompatible domain and range definitions for transitive, symmetric or inverse properties etc.

Finally the ontology is populated with instances; the diagrammatic representation can be seen in fig 2. This process of converting data into Ontology instances is called Ontology Mapping. The sharing and reuse of existing ontology increases the quality of the applications using them, as these applications become interoperable and are machine-processable [7]. Secondly, reuse, if performed in an efficient way, avoids the reimplementation of ontological components, which are already available on the Web. Furthermore, it can improve the quality of the reused ontology, as these are continuously revised and evaluated by various parties through reuse.

Gomez-Perez et al. present a survey of the most relevant methods, techniques and tools used for building ontology from text, machine readable dictionaries, knowledge bases, structured-data, semi-structured data and unstructured data [8]. Depending on the type of data the appropriate method,

IV. Semantic Data Integration

For instance, in a Health care domain, a patient may visit his family doctor periodically and his personal details and his allergy to penicillin is captured into a unstructured flat file like MS Word. Subsequently he visits the hospital for a blood test and has been diagnosed with malaria. This data is captured in relational form in the Hospital Information System. While the doctor decides the medication he must know about the patient's penicillin allergy to avoid an adverse drug reaction. This requires that the doctor sees an integrated view of all the patients data. To achieve this data from all the sources should be converted into Ontology Instances. Upon conversion, it is observed that both pieces of data is about the same Patient, because both ontology instances have the same URI (as seen in fig 2). So now the two representations can be merged and an integrated view of the data is provided to the doctor. Once the data is merged users can make queries on the whole data which ensures effective usage of the available patient data.

V. CONCLUSION

Ontology is used to model domain knowledge. Domain Ontology representation of data helps to achieve Semantic Data Integration for a specific domain. Through data integration, more information is available, which could lead to better decisions. The integrated data can be queried in an ad-hoc manner and it makes semantic searches possible. Ontology is a data representation form, which is formalized in a way that it further supports automatic information processing and reasoning. Ontology languages allow users to write explicit, formal conceptualizations of domain models.

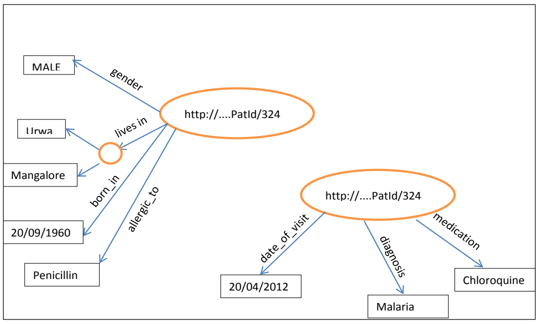

Fig 2: partial Patient information from two different Hospitals

"The first step is putting data on the Web in a form that machines can naturally understand, or converting it to that form. This creates what I call a Semantic Web—a web of data that can be processed directly or indirectly by machines."

-Tim Berners-Lee, Weaving the Web, Harper SanFrancisco, 1999 In conclusion, construction of Domain Ontology will help achieve Tim-Berners-Lee's vision of a Semantic Web – a web of data that can be processed directly or indirectly by machines.

REFERENCE [1] Grigoris Antoniou and Frank van Harmelen, "A Semantic Web Primer", | PHI Learning Private Limited, Second Edition , 2010. | [2] Michael C. Daconta, Leo J. Obrst and Kevin T Smith, "The Semantic Web A guide to the future of XML, Web Services and Knowledge Management", Wiley Publishing Inc, 2003. | [3] "W3 Consortium Semantic Web", internet : http://www.w3.org/standards/semanticweb/ [1st Mar 2013] | [4] "W3Schools Semantic Web", internet:http://www.w3schools.com/semweb/default.asp, [1st Mar 2013] | [5] "Ontology 101", | http://protege.stanford.edu/publications/ontology_development/ontology101-noy-mcguinness.html, [1st Mar 2013] | [6] Oscar Corcho, Mariano Fernandez-Lopez, Asuncion Gomez-Perez , "Methodologies, tools and languages for building ontologies. Where is their meeting point?", in Data &Knowledge Engineering vol. 46 pp.41–64, 2002. | [7] Elena Simperl, "Reusing ontologies on the Semantic Web: A feasibility study", in Data & Knowledge Engineering, vol. 68 issue 10, 2009. | [8] "Deliverable 1.5: A survey of ontology learning methods and techniques", Asunción Gómez-Pérez, David Manzano Macho, | http://www.stinnsbruck.at/fileadmin/documents/deliverables/Ontoweb/D1.5.pdf, [1st Mar 2013]