Fall 12-2011

Building Effective Ontology for Semantic Web: a Discussion Based on Practical Examples

Tianmu Zhang
UTK, tzhang5@utk.edu

Building effective ontology for Semantic web: a discussion based on practical examples

In partial fulfillment of the requirements
for IS 591 Independent Project/Research

Zhang, Tianmu
December, 2011

Instructed by:
Prof. Wilson, Bruce

School of Information Science, College of Information and Communication, UTK

Abstract

This paper serves as a report and summary of my independent study. The entire study aims to investigate semantic web and create useful ontology as a teaching and educational tool for others interested in learning more about Semantic web. This paper discussed several emerging issues about the semantic web and ontology building. This paper combines ontology implementation examples with research topics to identify current issues and potential solution in both application and theoretical level. It concludes that although semantic web and ontology technology are not mature enough currently, there is a clear tendency for them to be integrated into various applications to exert synergies.

Introduction

One of the main objects of semantic web is to enable a web of data to positively help us get what we want. In the traditional web model, the web server just returned the documents that are requested by the browser; and machines at both ends do not understand what the meanings of the documents they are transferring are. The foundation of the web, HTML, defines the syntax computer can understand, which is about how to display the documents to you. If we can get computer understand what's in the web pages, they can learn what we interested in, then computer can change from passively helping us to positively assisting us to retrieve what we want.

Semantic web supports methods that go beyond the traditional web application (both Web 1.0 and Web 2.0) in a way that it can facilitate machines to understand the meaning of information on the Internet. Ontology is a package of data together with their relationship, structure, and constrains. The most popular definition of ontology is that it is an explicit specification of the conceptualization of a domain [1]. Ontology makes information a meaningful knowledge which can not only convey semantic meanings but be interpreted and understand by machines as well. Similar to what we have in library field (such as controlled vocabulary and classification system), ontology can provides standard terms for annotating things and structured queries of entities. Although currently there are dissimilarities in different scientific areas in terms of the ontology language they use, it is certain that ontology is capable of unambiguously describing and uniquely identifying terms and concepts.

Diversity of ontology construction languages

In order to share common understanding of the structure of information for humans and software agents, we need to use well-structured format of information. These standardized formats are achieved by using semantic-enabled languages. Although there are many existing languages and data models, some of them being highly domain-specific, several functionalities are commonly designed towards the development and implementation of various languages. For example, semantic-enabled languages should be able to support at least one specific domain such as Open Biomedical Ontology (OBO), Gene Ontology (GO), Friend of a Friend (FOAF), etc. The diversity of ontology language not only lays in the specific scientific domains level, but also ontology construction format level as well. For example, we can use RDF/XML, Turtles, N triple N3, etc. to physically write ontology; although they use different syntax, they can generally achieve the same effectiveness and usefulness. Meanwhile, a good semantic-enabled language should support the compatibility of interoperating with language from other scientific domains.

Discussion of several issues with created ontology examples

This independent study creates three ontologies. They involve domains of biological classification ranking, social tagging and taxonomy, and geospatial datasets packaging. The following sections will examine each of them as illustrations in several research topics introduced respectively. Through the illustration of the underlining ontology, various functions and usages of ontology will be demonstrated.

Web semantic searching

For a long time we have experienced the strong power of various search engines, such as Google, Yahoo, Baidu, etc. These are all keyword search engines and are the most popular way of searching information on the web. However, we can expect much more capability, especially when we try to explore the potential of semantic web. Rather than using ranking algorithms such as Google's PageRank to predict relevancy, Semantic Search uses semantics of the language to reason and inference the most relevant results. Semantic search can not only improve search accuracy to generate more pertinent results, but also support complex queries involving inferencing and reasoning over complex data sets.

Using the organism classification ontology as an instance, the ontology describes the generic structural and hierarchical relationship among biological organism ranks. This ontology is aimed to illustrate how to use SPARQL Protocol and RDF Query Language (SPARQL) to implement relationship look-up and answer semantic questions. Semantic query language SPARQL makes it possible to extract new information from aggregation of inferred or deduced information. Software tools such as Protégé and Pellet can help to create new information from a composition of supplied raw RDF data and enhance information harvesting of content through their automatic reasoning systems. The bio-classification ontology example is used to illustrate basic semantic search that embody the meaning of queries and the available resources. For example, we can answer the question such as “what is the terminology and comments of the high bio-rank of the current levels?”

Figure 1 below displays the SPARQL query for answering above question and also the query results. In the first block of the query, all the prefix declarations specify the namespaces for all the properties that will be used in the below query block. We can see from the result that the middle column is the bio-classification type that is one level higher that the current selected one at the first column. Meanwhile, the right column is the comment for the bio-classification type on the middle column.

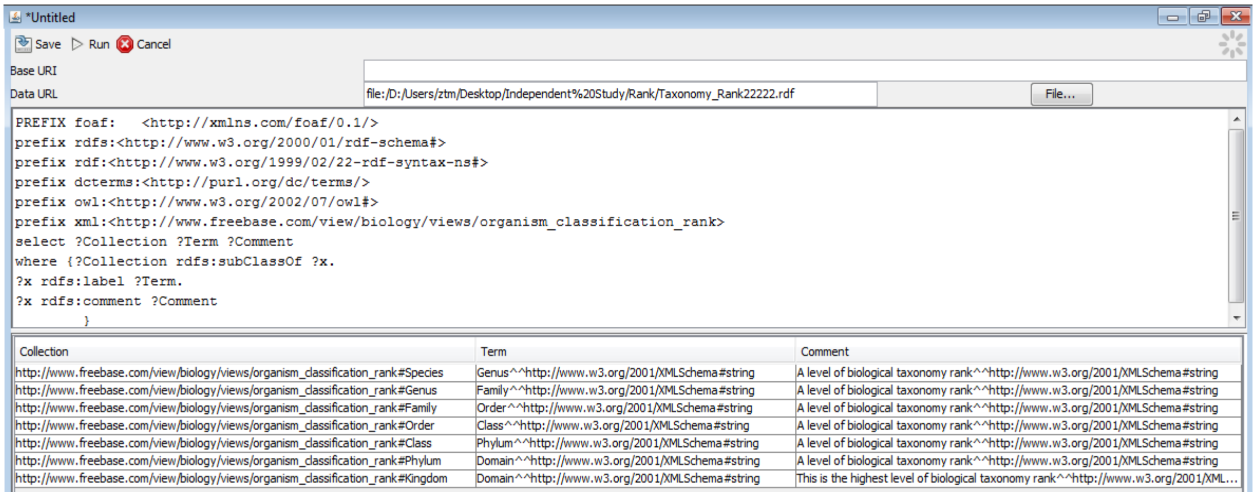


Figure 1. SPARQL query and results for Bio-Classification ontology

Full functioning semantic data searching is based on well-constructed individual ontologies and a complete network connection of these ontologies on the web. Currently, there exists many approaches that advocated by different organizations and institution. For example, there are semantic data retrieval and semantic document retrieval. A study introduces the notion of process-based semantic search, where semantics is exploited throughout all steps of the search process [2]. As aiming to achieve the maximum of web of data, they should all be designed to comply with the semantic search rationales discussed above.

Integrating FOAF framework into social networking

As a large interest of internet, Web 2.0 is currently dominating the web. Friend of a Friend (FOAF) is a framework launched by the FOAF project for representing information about people and their social connections in the form of machine-readable data on the web. Social networking data using the Friend of a Friend vocabulary is expending and will makes up a significant portion of all data on the Semantic Web [3]. Several researches address the advantages of FOAF in literatures. One study presented a survey of how FOAF was being used online and which parts of the FOAF vocabulary were utilized [4]. Another research uses learning techniques with FOAF data to infer characteristics of people in the network. The author create a set of rules based on his survey for adding properties to users found to be in a set of groups [5].To better take advantage of social tagging mechanisms, it could be good for FOAF ontology presenting semantic social networks in the form of named graph of entity, concept and instance associations, implementing the concept of ontology into the social dimension.

Figure 2 bellow shows the visual structure of a person's ontology in Protégé. Although the digital representation of this person is encapsulated as an object in the ontology, the structure of

RDF provide a standard method for exchanging information among applications. The diagram on the upper left corner is a diagram displaying all the classes and the hierarchy among them. The diagram on the upper right corner is the attributes of different classes; RDF triples that describe attributes and relationships are listed here as the objects of the specified properties. For example, Tianmu Zhang as a person has property workInfoHomepage as a predicate and the value “Illustration_ORE_Baseline.htm” as the object. The lower diagrams No.3 shows the hierarchy of relationships. Diagram No.4 gives us a vertical illustration of an instance property chain starts from “VirtualMe”.

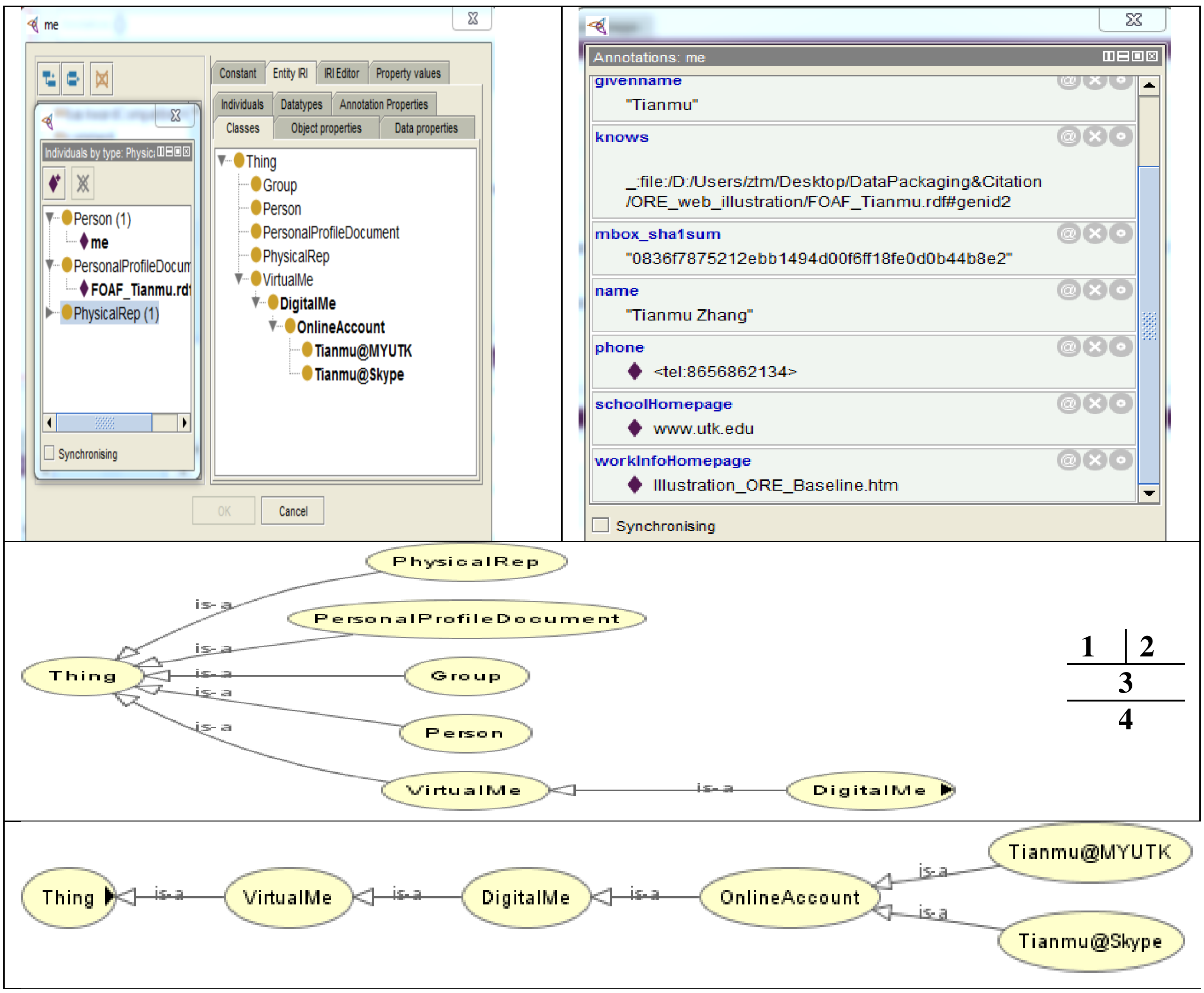


Figure 2. FOAF ontology in Protégé

Reuse ontology and scalability

To add the meaning and semantics to the data, common vocabularies are created within domain of interest. Relevant data concepts in the domain are captured along with their properties and objects and structured according to their relationship, which all together consist of ontology. The advantages of semantic ontology should include:

- Reusable
- Scalable

Reusability is a broadly recognized feature that a well constructed ontology should support; it could also include pattern, models and solution reusing. Reusing an ontology can help to provide a conceptualization of the domain, which reduced the effort of knowledge acquisition [6]. Scalability refers to the ability to extend the range and meaning of ontology. Instead of rewriting new entities and concepts to initiate new relationships, ontology can be easily added to the existing ontology.

As we have emphasized, achieving the Semantic Web functions requires us to create a lot of ontology or lined data, which is a tedious and costly challenging. Reuse existing ontology can reduce the cost of ontology engineering [7]. Another study points out that some principles about software design patterns for ontology engineering can be reused, adapted and extended for the construction of a more conceptualized ontology design patterns [8]. Building ontology involves several processes, such as determining scope, enumerating terms, defining categories, defining properties, defining Facets, and creating instances. A reuse mechanism can make all these process a prototype and expedite the creation of new ontology. At the same time, scalability can benefit a lot from reusing because reusing will lay a solid foundation for later integrating and adding on new ontology components.

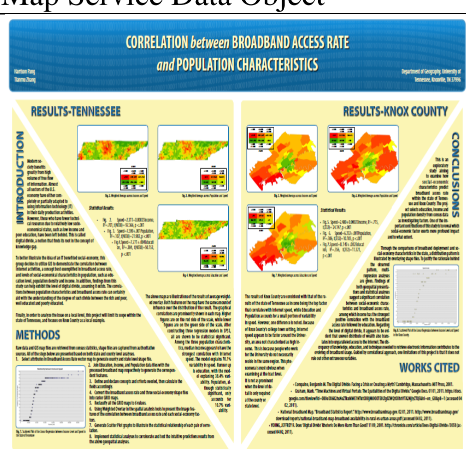
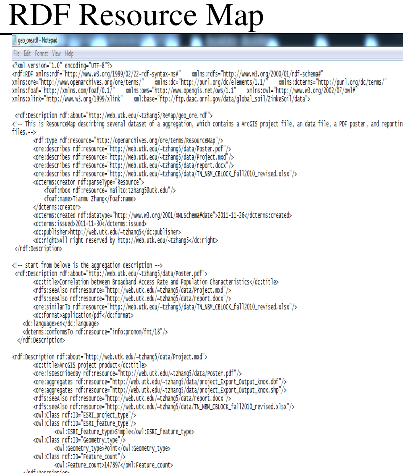
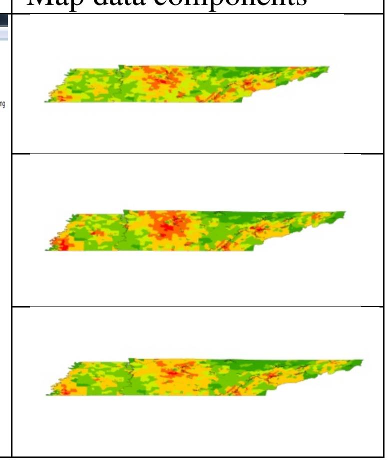
ORE data packaging and metadata at collection level

First of all, Archives Initiative Object Reuse and Exchange (OAI-ORE) [9] is a perfect framework for guiding how to reuse ontology. Moreover, ORE takes one step further and expends the notion of reusability into the data package or data collection level.

We have seen example of achieving semantic functions in a single ontology. However, in real world, many datasets are at collection level. For example, we might have a photo album in Flickr. The photo album is an aggregation of many photos. We often use the URL of one page to bookmark the whole aggregation. When we, human viewers, open a web page, we have the ability to distinguish all kinds of constituents of the page, relationship of buttons or links to external pages, and so on. However, without a standard description of the constituents and boundary of these aggregations, computers cannot unambiguously interpret them.

One of the important objects of this independent study is to explore the effectiveness of reusing collection-level ontology resource. Previously, Collection-level metadata is poorly understood and stored. Reasons are various on the system and operational level, such as information retrieval systems do not support the collection context harvest function, data packages themselves do not contain collection level metadata. Actually, the key point behind that is a lack of a standardized framework for participating parties to follow. Fortunately, ORE bridges this gap. It establish the rules by introducing the Resource Map (ReM), it has a machine-readable representation that provides details about the aggregation. We can assign HTTP URIs (URL) to both aggregation and ReM in order to make ORE work in HTTP-based web. So that whenever we want to retrieve an aggregation of resource, we use the corresponding HTTP URI to dereference it, server that receives this HTTP request will redirect the user to the Resource Map URL, which contains the ontological description of the aggregation and related information. Therefore, metadata at collection level can also be harvested and interpreted by systems.

Several research groups and institutions have taken the initiatives in this field. A few researches have developed models or software tools that are ORE compatible, for example, the Scientific Compound Object Publishing and Editing System (SCOPE) [10]. We can achieve HTTP-enabled ontology resource reuse through packaging datasets under the guideline of Open Linked Data and complying with the ORE standard. The independent study creates and demonstrates a reusable package of ontologically-structured scientific dataset. The ontology is a representation of a set of geospatial and social-economical data. Geo-ontology is just the common knowledge with the specific domain, geospatial field. This example of ontology reusing involves packaging science data, science metadata, companion files, and system metadata in a standardized the format conforming to OAI ORE. The Figure 3 below describes the constituents of the map service data object on the left column; the RDF resource map in the middle column is the essence of the whole package, it points out and describes all the components within the package; Right column is the ArcGIS shapefiles that consist of the constituents of data objects.

Map Service Data Object	RDF Resource Map	Map data components
		


```
geo_ore.rdf - Notepad
File Edit Format View Help
<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:ore="http://www.openarchives.org/ore/terms/"    xmlns:dc="http://purl.org/dc/elements/1.1/"    xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:foaf="http://xmlns.com/foaf/0.1/"    xmlns:ows="http://www.opengis.net/ows/1.1"    xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:xlink="http://www.w3.org/1999/xlink"    xml:base="ftp://ftp.daac.ornl.gov/data/global_soil/ZinkeSoil/data">

 <rdf:Description rdf:about="http://web.utk.edu/~tzhang5/ReMap/geo_ore.rdf">
<!-- This is ResourceMap descirbing several dataset of a aggregation, which contains a ArcGIS project file, an data file, a PDF poster, and reporting
files.-->
        <rdf:type rdf:resource="http://openarchives.org/ore/terms/ResourceMap"/>
        <ore:describes rdf:resource="http://web.utk.edu/~tzhang5/data/Poster.pdf"/>
        <ore:describes rdf:resource="http://web.utk.edu/~tzhang5/data/Project.mxd"/>
        <ore:describes rdf:resource="http://web.utk.edu/~tzhang5/data/report.docx"/>
        <ore:describes rdf:resource="http://web.utk.edu/~tzhang5/data/TN_NBM_CBLOCK_fall2010_revised.xlsx"/>
        <dcterms:creator rdf:parseType="Resource">
            <foaf:mbox rdf:resource="mailto:tzhang5@utk.edu"/>
            <foaf:name>Tianmu Zhang</foaf:name>
        </dcterms:creator>
        <dcterms:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">2011-11-26</dcterms:created>
        <dcterms:issued>2011-11-30</dcterms:issued>
        <dc:publisher>http://web.utk.edu/~tzhang5</dc:publisher>
        <dc:right>All right reserved by http://web.utk.edu/~tzhang5</dc:right>
 </rdf:Description>

<!-- start from belove is the aggregation description -->
 <rdf:Description rdf:about="http://web.utk.edu/~tzhang5/data/Poster.pdf">
        <dc:title>Correlation between Broadband Access Rate and Population Characteristics</dc:title>
        <rdfs:seeAlso rdf:resource="http://web.utk.edu/~tzhang5/data/Project.mxd"/>
        <rdfs:seeAlso rdf:resource="http://web.utk.edu/~tzhang5/data/report.docx"/>
        <ore:similarTo rdf:resource="http://web.utk.edu/~tzhang5/data/TN_NBM_CBLOCK_fall2010_revised.xlsx"/>
        <dc:format>application/pdf</dc:format>
    <dc:language>en</dc:language>
    <dcterms:conformsTo rdf:resource="info:pronom/fmt/18"/>
  </rdf:Description>

<rdf:Description rdf:about="http://web.utk.edu/~tzhang5/data/Project.mxd">
        <dc:title>ArcGIS project product</dc:title>
        <ore:isDescribedBy rdf:resource="http://web.utk.edu/~tzhang5/data/Poster.pdf"/>
        <ore:aggregates rdf:resource="http://web.utk.edu/~tzhang5/data/project_Export_Output_knox.dbf"/>
        <ore:aggregates rdf:resource="http://web.utk.edu/~tzhang5/data/project_Export_Output_knox.shp"/>
        <rdfs:seeAlso rdf:resource="http://web.utk.edu/~tzhang5/data/report.docx"/>
        <rdfs:seeAlso rdf:resource="http://web.utk.edu/~tzhang5/data/TN_NBM_CBLOCK_fall2010_revised.xlsx"/>
        <owl:Class rdf:ID="ESRI_project_type"/>
        <owl:Class rdf:ID="ESRI_feature_type"/>
                <owl:ESRI_feature_type>Simple</owl:ESRI_feature_type>
        <owl:Class rdf:ID="Geometry_type"/>
                <owl:Geometry_type>Point</owl:Geometry_type>
        <owl:Class rdf:ID="Feature_count"/>
                <owl:Feature_count>147897</owl:Feature_count>
    </rdf:Description>
```

Figure 3. Illustration of ORE ontology using Geo-spatial example

A potential solution to efficient searching

As survey shows, academic reading are rely less on browsing and more on searching, and the increase in the number of papers read by scientists per year but decrease in minutes spent reading each paper[11]. One essential difference between searching and browsing is that you have something specific in mind that you want when you browse while you might just causally look through materials generally when you are searching things. Ontology helps us to broaden the range of pertinent knowledge volume when we do searching and facilitates the semantic information seeking. It inspires ideas by and saves scientists a lot of time by returning to them the most possible relevant results they are caring about, which improves the efficiency of searching. When we get the result of a search we want, we no longer need another search if we want to find some concepts that related to the previous search results. Ontology provides one type of solution by pointing out the relevant and related objects, concepts, relationships to the searched items. Users can just click and retrieve the things they want. For example, if I want to search a term in a specific context in the Library of Congress (LC) authority, all that I am allowed to do is scamming through the lists and check the meaning of each potential vocabulary and then decide which one is the term I am really looking for. With the implementation of ontology, LC authority could give users a list of candidate concepts according to the relevance of the context the user given.

Limitation and conclusion

Although semantic web and ontology have all the aforementioned powerful functions and advantages, there are still some barriers that retard it from wide implementation. In order for machines to understand the semantics of information on the web, first of all, the web must contain enough well-structured data. Without a huge amount of semantic data as the root, we cannot really get sufficient information no matter how well these data are deduced, analyzed, or reasoned. One problem is even if a lot of people contributing their own piece of data into this web of data in the near future, it is unlikely to establish a mature mechanism to inspect, check, and validate these distributed datasets in a centralized way.

In could be possible that for some sophisticated implementation, we can answer question such as how does the weather affect the stock market. However, the purpose of semantic web is not towards data mining, all the well-structured data and relationships must be pre-coded when ontology was created so that the expected results can be generated. The data quality relies on the data contributors. As a result, rather than guaranteeing absolute correctness and accuracy, semantic web tends to provide a new method of digging data, without traditional statistical analyses or massive programming sources codes, so that non-technical individuals can easily get involved.

Reference

1. Gruber, T.R.: Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Guarino, N., Poli, R., eds.: Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, Kluwer Academic Publishers (1993)

2. T. Tran et al., SemSearchPro – Using semantics throughout the search process, Web Semantics: Sci. Serv. Agents WorldWide Web (2011), doi:10.1016/j.websem.2011.08.004

3. Golbeck, J. and Rothstein, M. Linking Social Networks on the Web with FOAF: A Semantic Web Case Study. In Proceedings of AAAI. 2008, 1138-1143.

4. L. Ding, L. Zhou, T. Finin, and A. Joshi. How the semantic web is being used: An analysis of foaf documents. In HICSS '05: Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4, Washington, DC, USA, 2005. IEEE Computer Society.

5. G. A. Grimnes, P. Edwards, and A. Preece. Learning meta-descriptions of the foaf network. In Proceedings of the International Semantic Web Conference, 2004.

6. D.N. Peralta, H.S.A.N.P. Pinto, N.J. Mamede, Reusing a time ontology, in: Proceedings of the First International Conference on Enterprise Information Systems ICEIS, 2003, pp. 121–128.

7. Simperl, E.P.B. Reusing ontologies on the Semantic Web: A feasibility study. In Proceedings of Data Knowl. Eng.. 2009, 905-925.

8. Gangemi, A., Gomez-Perez, A., Presutti, V., Suarez-Figueroa, M.C. Towards a Catalog of OWL-based Ontology Design Patterns. In proceedings of CAEPIA. 2007.

9. Lagoze, Carl, and Herbert Van de Sompel. Open Archives, "Open Archives Initiative." Last modified Oct 17 2008. Accessed September 28, 2011. http://www.openarchives.org/ore/1.0/primer.html.

10. Kwok Cheung, Jane Hunter, Anna Lashtabeg, John Drennan. "SCOPE: A Scientific Compound Object Publishing and Editing System." International Journal of Digital Curation 3, no. 2 (2008). http://www.ijdc.net/index.php/ijdc/article/view/84

11. C. Tenopir, D.W. King, D-Lib Mag. 14, issue 11/12 (2008).