ROLE OF VOCABULARY FOR SEMANTIC INTEROPERABILITY IN ENABLING THE LINKED OPEN DATA PUBLISHING

Ahsan Morshed

Intelligent Sensing and System Laboratory (ICT)
Commonwealth Scientific and Industrial Research Organization (CSIRO), Hobart, TAS-7001, Australia
ahsan.morshed@csiro.au

ABSTRACT

In spite of the explosive growth of the Internet, information relevant to users often unavailable even when using the latest browsers. At the same time, there is an ever increasing number of documents that vary widely in content, format and quality. The documents often change in content and location because they do not belong to any kind of centralized control. On the other hand, there is a huge number of unknown users with extremely diverse needs, skills, education, and cultural and language backgrounds. One of the solutions to these problems might be to use standard terms with meaning, this can be termed as controlled vocabulary (CV). Though there is no specific notion of CV, we can define it as a set of concepts or preferred terms and existing relations among them. In this paper, we focus on the role of CV for publishing the web of Data on the web.

KEYWORDS

Ontology, Classification, Linked Open Data, Semantic Web & Vocabularies

1. INTRODUCTION

In spite of explosive growth of the Internet, information relevant to user is often unavailable even when using the latest browsers. At the same time, there is an ever increasing number of documents that vary widely in content, format and quality. The documents often change in content and location because they do not belong to any kind of centralized control. On the other hand, there is a huge number of unknown users with extremely diverse needs, skills, education, and cultural and language backgrounds. One of the solutions to these problems might be to use standard terms with meaning, this can be termed as controlled vocabulary (CV) [3, 4]. Though there is no specific notion of CV, we can define it as a set of concepts or preferred terms and existing relations among them. For example, thesauri, WordNet [22], MeSH [25], LCSH [23], all kinds of ontologies, etc. are sorts of CVs. These CVs are used to match purpose that make more flexible for information extraction. In a semantic or controlled vocabulary [15] a matching operator takes two-graph like structures, for instance ontologies or classifications and produces matching relationship among them. This semantic matching system is based on two key notions. One of them is the concept of nodes and the other is the concept of labels. In semantic matching, labels are written in natural language. These labels are disambiguated using a lexicon [29]. In this case, they are working as background knowledge. In this paper, we will see the contribution of CV for publishing the web of data purposes and review the main applications of controlled vocabularies.

DOI: 10.5121/ijdms.2012.4502

The rest of this paper is organized as follows: section two provides more details about classification of controlled vocabularies; section three presents the application of CVs in different aspects, section four presents semantic operability of controlled vocabularies for publishing linked open data and it also provides matching different matching techniques and tools and finally conclusion.

2. CLASSIFICATION OF CONTROLLED VOCABULARY

In our case, we can classify our controlled vocabularies based on nature, construction perspective and usage. These constructions are based on regions, countries, products, services, vertical markets, clients, customer alliances, structure subsidiaries histories and cultures etc. For instance, two words "*Center*" and "*Centre*" both have the same meaning but different spelling in different regions and cultures.

We can classify controlled vocabularies in the following way:

2.1 General controlled vocabulary:

This class of controlled vocabulary is mainly included in usage and existing relationships among the concepts and entities. For example, the most prominent representation of these vocabularies are Thesaurus, WordNet, Classification, Directories, Lightweight Ontologies [1], etc.

2.1.1 Thesaurus

A thesaurus [18, 49] can be defined as a controlled vocabulary that includes synonyms, hierarchies and associative relationships among terms to help users to find the information they need". For example, two users are looking for information on "*Automobile*". One may use the term "*Car*" while the other may use "Auto". Each of them queries the same information with different terms, but these terms belong to same concept. So, the success of finding relevant documents varies based on demand and context. To address the problem, thesauri map variations in terms (synonyms, abbreviations, acronyms and altered spelling) of a single preferred term for each concept. For document indexer, the thesauri provide the index term to be used to describe each concept. This enforces consistency of document indexing. For users of a Web site, the thesauri work in the background, mapping their keywords onto single preferred terms, so they can be presented with the complete set of relevant documents.

2.1.2. WordNet

A human compiled electronic dictionary is one kind of ontology that expresses meanings of bounded terms. It was developed by Prof. George Miller at Princeton University. It mainly builds up on a lexical knowledge base born from psycholinguistic research into the human lexicon. It has applications in different fields of research, sense disambiguation, semantic tagging and information retrieval [22].

2.1.3. EuroWordNet

This is an European project for WordNet. The aim of this project is to develop multilingual dictionaries with WordNet for several European languages. In this project based on WordNet, each individual net is linked to a central system which is called Inter-Lingual-Index. Each net is composed of about 30,000 synsets and 50,000 entries [8].

2.1.4. Dmoz

An open directory project is the most panoptic human edited directory of the Web. It is constructed and maintained by a vast, global community of volunteer editors. Web content is growing at staggering rates. Search engines are increasingly unable to provide useful results to search queries. The open directory provides a way to keep the Internet itself classified. It uses standard terms to tag the directories so that anyone can browse it [5].

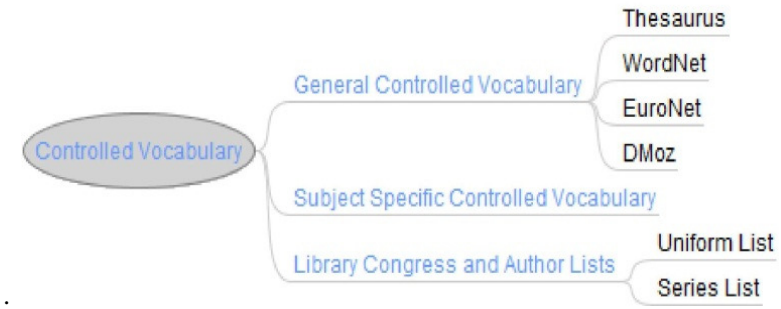

Figure 1. CVs

2.2 Subject specific controlled vocabulary (SSCV)

Construction of sentences, words and data are most of the time used in subject specific controlled vocabularies, for example languages to express chronology, hypothesis, comparison, etc. Typically an SSCV is expressed as key words, key phrases or classification codes that describe the theme of the resource. In the library sciences, due to the ever-increasing number of records, bibliographic systems are facing difficulties. Documents in the library system are heterogeneous: some of them provide few hints, some are disparate, while in others structural tags are sometimes not used properly, which results in inefficiency in extracting documents. However, controlled vocabularies which have traditionally been used in libraries, could serve as good-quality structures for subject browsing among entire documents. Subject heading systems and thesauri have traditionally been developed for subject indexing that would describe topics of the document more specifically structured [20].

Figure 2. Library of Congress Online Catalog

2.3 Library Congress and Authors List

The Semantic Web and library communities have both been working toward the same set of goals: naming concepts, naming entities and bringing different forms of those names together. The Semantic Web's efforts toward this end are relatively new, whereas libraries have been doing work in this area for hundreds of years. Vocabularies developed in libraries, particularly at the Library of Congress, are sophisticated and advanced in searching and representation. Libraries have a long-standing history of developing, implementing and providing tools and services that encourage the use of numerous controlled vocabularies. When the naming conventions are translated into Semantic Web technologies, they will help realize Berners-Lee's dream [23]. Furthermore, the roles of libraries in the Semantic Web are as follows:

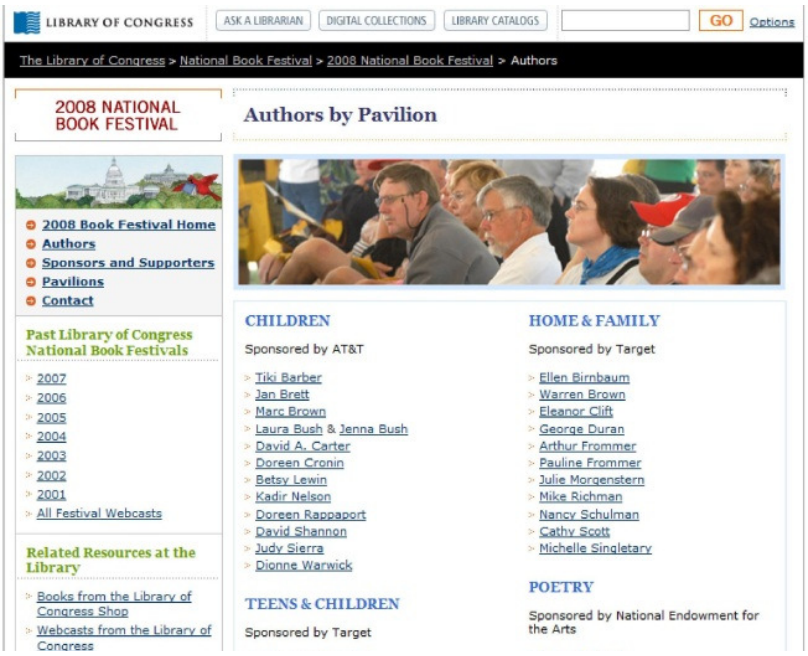

Figure 3. Library Congress Author List

- Exposing collections-use Semantic Web technologies to make content available
- Web'ifying,
- Thesaurus/Mappings/Services
- Sharing Learned.
- Persistence.

As all of the above roles are equally important, the intuition to move controlled vocabularies into a standard to which web services can gain easy access to information management. By conforming all these vocabularies to Semantic Web standards, such as controlled vocabularies will provide limitless opportunities to use them in different ways. This can make possible searching and browsing diverse records, verifying and identifying particular authors and browsing sets of topics related to a particular concept [20]. Authors List can be categorized into two ways:

2.3.1 Uniform List

This category [18] includes all universal names. For example, the "Bible", the "Gita", the "Quran", the "Tripod" and the "Lake of Garda" etc. This kind of series list of controlled vocabularies is included in different consecutive names. From a unique list it is easier to match the concepts they represent.

2.3.2 Series List

This category includes the series of same name with the different themes such as "Terminator-1", "Terminator-2", and"Terminator-3".

3. APPLICATIONS

3.1 Applications for managing controlled vocabularies

3.1.1 Traditional Controlled Vocabulary tools

The vocabulary which is used in legacy systems is called the traditional vocabulary. For example, the AGROVOC[9], a thesaurus is mainly in relational database format and is published on the website for browsing and navigating concepts and their relations. It was previously available only in four languages. Now it is available in 22 languages. Major drawbacks of traditional controlled vocabularies are that they were not well structured, they were only text format or SQL format, their relationships were not well defined, there was no semantics between the concepts and there was no Unified Resource Identifier (URI) for locating the concepts.

3.1.2 A Modern Controlled vocabulary collaborative management system

Modern controlled vocabularies[12]are one kind of lightweight ontologies with well defined multiple formats (SKOS, RDF, and OWL etc). In this vocabulary, each concept is assigned a URL. Using this URI, one can populate concept information and use this information for further research. One example of modern controlled vocabulary is AGROVOC VocBench. In VocBench, one can add or modify the concepts in distributed manner.

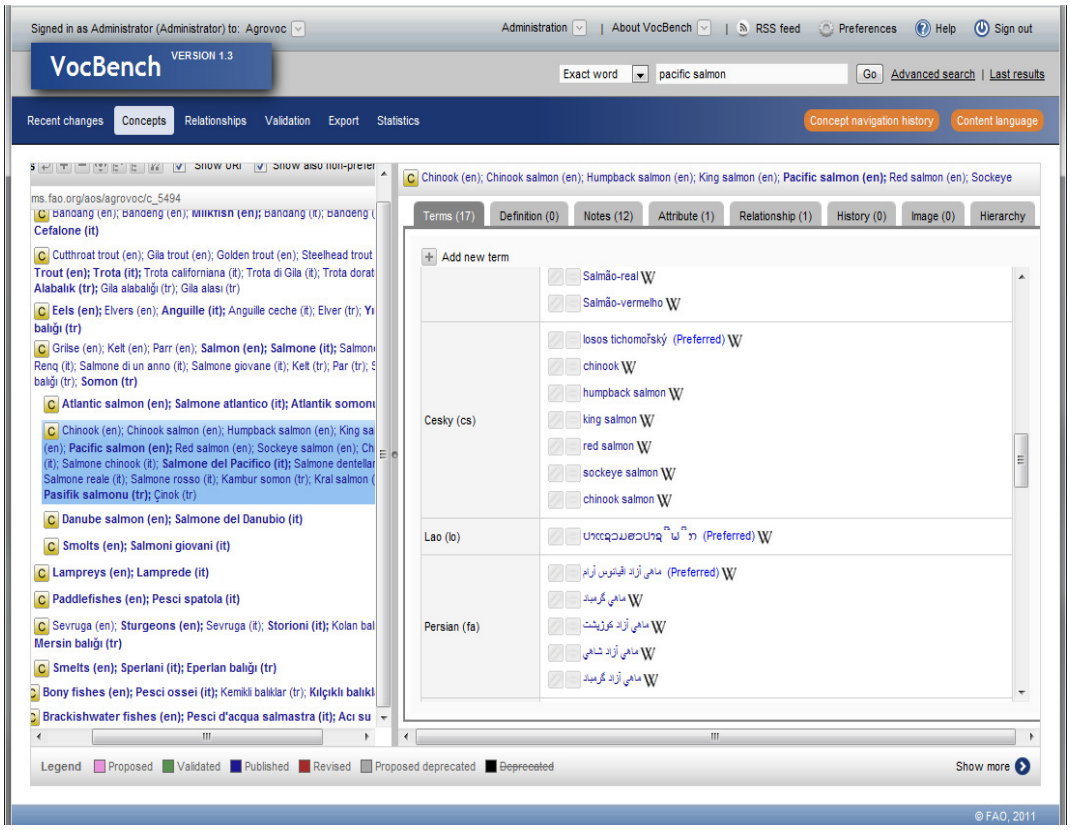

Figure 4. AGROVOC VocBench

3.2 Applications for exploiting controlled vocabularies

3.2.1 Background Knowledge

Controlled vocabularies are used in subject indexing schemes, subject headings, thesauri and taxonomies to provide a way to organize knowledge for subsequent retrieval [2,16].The Controlled vocabulary strategy assigns the use of predefined, authorized terms that have been preselected by the designer of the vocabulary. For easy accessing to the digital information and library catalogues, tags are carefully selected from the words and phrases in a controlled vocabulary. CV controls the use of synonyms (and near-synonyms) by establishing a single form of the term. This ensures that indexers apply the same terms to describe the same or similar concepts, thus reducing the probability that relevant resources will be missed during a user search. The biggest advantage to controlled vocabularies is that once you find the correct term, most of the information you need is grouped together in one place, saving you the time of having to search under all of the other synonyms for that term. In large organizations, controlled vocabularies may be introduced to improve inter-departmental communication. The use of controlled vocabularies ensures that everyone is using the same word to mean the same thing. This consistency of terms is one of the most important concepts in technical writing and knowledge management, where effort is expended to use the same word throughout a document organization instead of slightly different ones referring the same thing.

3.2.2 Document annotation

The objective of document annotation is to use appropriate terms so that machines can easily understand and correctly classify the documents, allowing the user easy access while searching or browsing. For example, Clusty[6], Vivisimo[7], Swoogle[45], etc. are classified documents under pre-defined keywords or terms so that one can go to specific locations to find the needed information[19]. Furthermore, document annotation is needed for building knowledge bases that will be used in the future Web and existing large sets of corporas. However, existing information retrieval systems use string matching techniques for full-text search or key phrase search. Thus, a major problem with these systems is overlapping the matching terms or matching results. To overcome these difficulties, more semantic information should be added to matching techniques. The present NLP (natural language processing) techniques cannot provide the complete solution. There is more work to be done. In additional, document annotation can help to improve the performance of information extraction.

3.2.3. Information retrieval and extraction

WordNet has been used as a comprehensive semantic lexicon in a module for full text message retrieval as a communication aid, in which queries are expanded through keyword design. In [16], automatic construction of thesauri, based on the occurrences determined by the automatic statistical identification of semantic relations is used for text categorization. English words can have different meanings or the same meaning with different structures or descriptions. For example, “center" and “centre" have the same meaning but different spelling for American and British English. Conversely, the same words can have different meanings, for example “bank" means “river side" or “financial institution". It is hard to classify documents or satisfy user queries according to the meaning of words. Text categorization is the process of categorizing the document under a specific class. WordNet lexical information builds a relation between sentences and coherent categories. Sebastiani [47] describes an algorithm for text categorization using WordNet.

3.2.4. Audio and Video retrieval

In the digital age, the most challenge is to handle the huge amount of hyper-media or non-textual information on the Web. For example, an YouTube [26], over 150,000 videos are uploaded and 100,000,000 queries are performed every day. In order to control these high volumes of hyper-

media information, information must be used and used in the right way. For instance, the multimedia miner [24] is a prototype to extract multimedia information and knowledge from the web to generate conceptual hierarchies for interactive information retrieval and build multi-dimensional cubes for multimedia data. Finally, WordNet or Thesaurus are used in query expansion for TV or radio programs to index the news automatically. It has some drawbacks; for instance, it is not domain specific and it is not possible to find relationships between terms with different partsof speech.

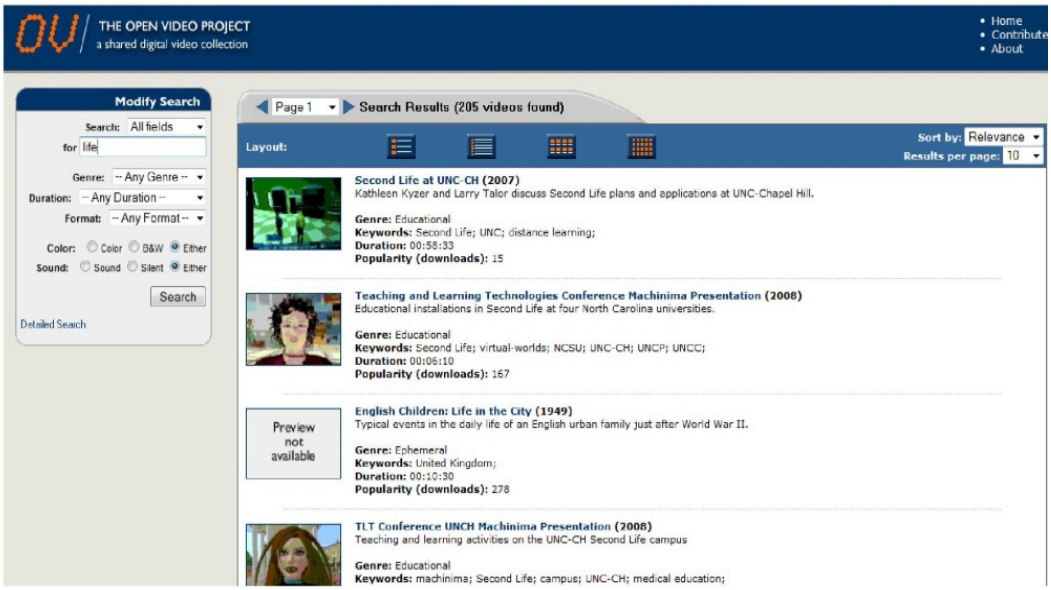

Figure 5. Video Indexing

3.2.5. Semantic interoperability, data exchange and integration

Controlled vocabularies are used in resolving semantic heterogeneity among data sources for data exchange and integration in different domain.

In bioinformatics domain [41], controlled vocabulary is used in ingrating molecular biological data for resolving different terms of the same thing and accessing data without know the structure and technical issues. In medical domain [39], different ontology alignment through controlled vocabularies is ued in semantic ingration of medica data. In geology and mining [39], controlled vocabulary is used for semantic interoperability of geodata from mining projects. For this purpose, concepts and their relationships are proposed in knowledge domain of mineral exploration for mining projects. In devolping controlled vocabulary, the used national standard of geosciences taxonomies and terminolgies. Further, controlled vocabulary is used resolving heterogeneity anong data sources from mining projects in integrating databases. Sharing data in hydrological domain [42], controlled vocabulary is also used. In [43], to annotate and integrate biological datasets, controlled vocabulary is used.

3.2.6 Managing information in social network

The endlessly growth of information resources on the web demands better classification. This classification is needed to browse web pages more smoothly. Previous orthodox information resources were not consistent because of changing static to dynamic pages on the Web. After changing those information resources to modern information resources, a more consistent to

categorization is needed. However, the problem was not only browsing the pages but also consisting of qualities of Web sites content. To overcome this problem, a change to apply online vocabulary resources is needed to help end users to find what they are looking for. Furthermore, social networking, linking data, Flickr[13], Google Maps[46] and intercompany collaboration, etc. brings have a common ground which further necessitates a controlled vocabulary.

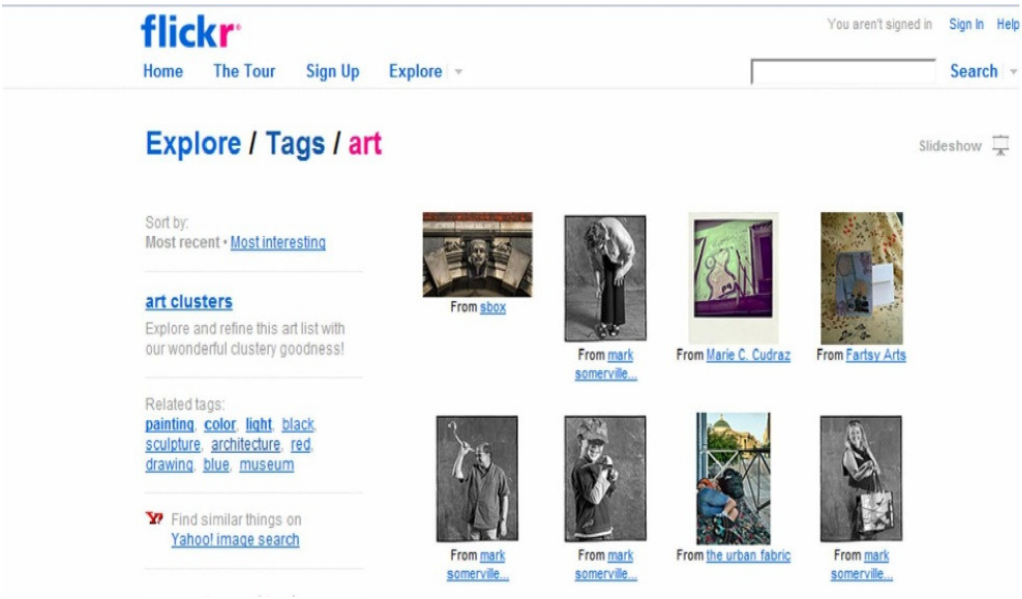

Figure 6. Controlled Vocabulary used as Tagging in Flickr

3.2.7 Controlled vocabulary in web intelligence and recommender systems

In personalized recommender system [44], controlled vocabualry is used in tagging of items. The item taxonomy is a set of controlled vocabulary terms.

4. SEMANTIC INTEROPERALABILITY OF CONTROLLED VOCABULARIES FOR PUBLISHING LINKED OPEN DATA

4.1 Controlled vocabularies in matching

People are breaking their legacy data silos and uploading the data on the web . To get the real value from these uploaded data , it is needed to connect them. It occurs the heterogenous issue. The Maching is the main factor for linking the data in distributed environment . According to Tim Berners-Lee, linking resource get the highest start for Linked Open Data principle's .

4.1. 1. Matching problem

The semantic heterogeneity is the big problem of matching the controlled vocabularies. In order to clarify our problem statement, let us proceed to match CVs. The CV stores concepts and relationships between these concepts. We write C_{cv} to denote the set of concepts stored in the CV database. We write c^i to denote a concept with ID i in the CV database (i.e., $c^i \in C^{cv}$). The

stored relations are specificity($c^i \subseteq$ cj) and disjoint relationship($c^i \cap$ cj). We write Ro to represent the ($\subseteq or \perp$) that holds between concept c^i and cj. The set Ro can be used to compute all the other possible relations that hold between concepts in CV. The set of concepts and the set of original relation in the CV can be represented in the form of a graph, whose nodes are concepts and which has two kind of edges .The first kind represent the specificity and it is shown as a directed edge. An edge directed from node i (concept c^i) and node j (concept cj) means that cj $\subseteq$ c^i. The second kind of relation is disjoint relation. Let a mapping element be a 4-tuple $\langle ID_{ij}, c_i, c_j, R \rangle$, where ID_{ij} = is unique identifier of the given element c^i = a set of concepts in CV1, cj = a set of concepts of concept CV2, R=relation which holds between concepts of vocabularies. The possible semantic relations are: equivalence ($\equiv$), more general ($\supseteq$), less general ($\subseteq$), and mismatch ($\perp$).

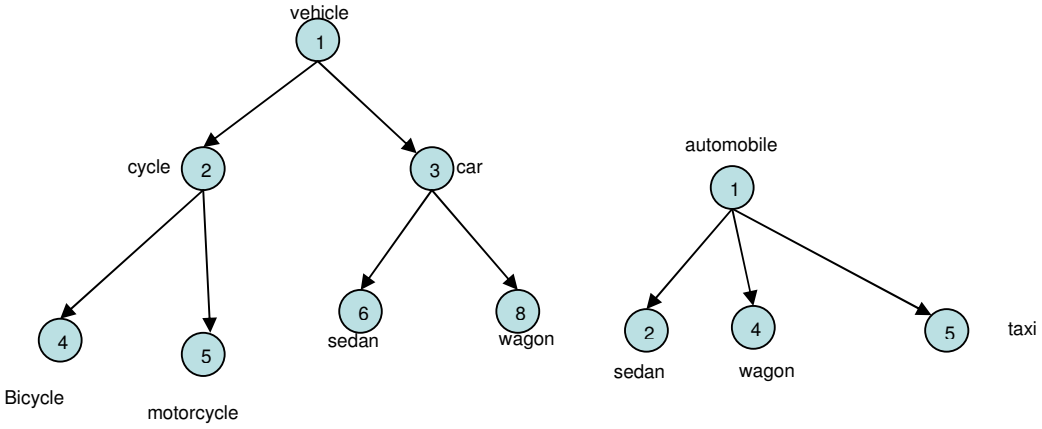

Figure 7. CV_1 and CV_2

For instance, we consider two concepts from CV1 and CV2. The two concepts respectively C^{car} and $C^{automobile}$ which represent concept label car and concept label automobile that mean car is a entity or thing in the real world, similarly automobile mean an entity in the real word. As we know that

$$C^{car} \equiv C_{automobile} \text{ if } C^{I}_{car} = C^{I}_{automobile}$$

Since there is no similarity between two concepts label then we cannot say that they are equivalent. Now, we check their synonym to find out if there is any similarity existing or not.

Synset of C^{car} : <car, auto, machine and>

Synset of $C_{automobile}$:< auto, automobile, motorcar>

Since they have common word "auto" then we can assume they have an existing relationship. However, it is not enough to draw the conclusion about similarity between two concepts only using synonym. We go through less general ($\subseteq$) and more general ($\supseteq$) relationship of concepts. For example, car is having two children

$$C_{sedan} \subseteq C^{car}$$

$C_{\text{wagon}} \subseteq C_{car}$

and automobile is having three children

$C_{sedan} \subseteq C_{automobile}$

$C_{\text{wagon}} \subseteq C_{automobile}$

$C_{taxi} \subseteq C_{automobile}$

Since they have two common children, we can assume they might same concept. For these reason we need to find out the parent of car and parent of automobile. For instance, these two concepts are having same parent vehicle. So, we can say that they are siblings.

$C_{vehicle} \supseteq C_{car}$

$C_{vehicle} \supseteq C_{automobile}$

The following assumption we know from CV database [3, 4] that

A word cannot exist in the database without at least one synset associated with it

A synset cannot exist in the database without at least one word associated with it.

According to this assume, we can say, a concept can represent one word or multiple words.

For instance, C_{car} can be represented word “car” and similarly, $C_{automobile}$ can be represented word “automobile”

$$W_{car}^{CV1} \neq W_{automobile}^{CV2}$$

These two words can compare only by syntactically [0,1]. Therefore, we can only use equivalent relation (≡) on it. The problem occurs due to different word form stores in different controlled vocabularies and there is no standard file or authenticate file to describe for forming of words. For instance,

If we have two words network and networking

W_{netork}^{CV1} $W_{networking}^{CV2}$

This above case, we can see that both words have a common 6 literals. In our case, we consider equivalent relation between words by given threshold in order to solve the problem. More precisely, we will give equivalent relation (≡) between words if their three literals are common. As result,

$W_{netork}^{CV1} \equiv W_{networking}^{CV2}$

Therefore, the only the equivalence relation could be used for words and synonyms. Furthermore, concepts are equivalent if they have the same concept label, i.e., they carry the same meaning in the real world for example if we say concept of car, it means a set of document which tell about the car [22]; otherwise, they are mismatched. Hence, equivalence is the strongest binding relation as the second entity is exactly the same as the first. On the other hand, “more general” and “less general” relations give us containment information with respect to the first entity, while the mismatch relation provides containment information with respect to the extension of the complement of the first entity.

There are some restrictions in case of mapping control vocabularies.

If $\langle c_i, c_j, \perp \rangle \in Ro(c_i, c_j \in C_{cv}, i \neq j)$ then $c^i \not\sqsubseteq cj$ and $cj \not\sqsubseteq ci$

If $\langle c_i, c_j, \perp \rangle \in Ro(c_i, c_j \in C_{cv}, i \neq j)$ then $\nexists c_k$ $(k \neq i, k \neq j)$ s.t. $c_k \subseteq c_i$ and $c_k \subseteq c_j$;

The first restriction specifies that there cannot be a "disjoint" relation between two concepts if one is more specific than other, the second restriction specifies that "mismatch" concept cannot have common descendent. Apart from these restrictions, we do not consider the "more general" or "less general" relation between words and synonyms.

4.1.2 Matching methods

4.1.2.1 String matching

These techniques are often used to match between two words from given entities. They consider strings as sequences of letters in an alphabet. They are typically based on the following intuition: the more similar the strings, the more likely they denote the same concepts. Some examples of string-based techniques which are extensively used in matching systems are *prefix*, *suffix*, *edit distance*, and *n-gram [27,35]*. For example*, we can consider a match between the words the "Hot" and "Hotel" according the prefix matching.*

4.1.2.2 Semantic matching

Semantic Matching is introduced in [14, 15, 27] and it does not consider straight string matching techniques for matching purpose. It takes two classifications and produces matches. This matching system based on two key notations; one is concept of node, concept of label. However, background knowledge is major factor for its functionalities. WordNet plays vital role for this purpose.

4.2. Vocabulary matching tools

There are several tools for matching purposes. We descirbe some of them in here.

4.2.1. FALCON-AO

Falcon [36] is a platform for Semantic Web applications that provides fundamental technology for finding, aligning and learning ontologies. Falcon-AO is an automatic ontology matching system that aids interoperability between ontologies. The Falcon-AO tool takes RDF /OWL as input and produces RDF as output. Furthermore, this tool includes LMO (linguistic matching for ontologies), GMO (graph matching for ontologies) and PBM (a partition-based matcher for large ontologies).

4.2.2 CTXMatch, S-match

Context Match (CTXmatch) and Semantic Matcher (S-match) [15, 37] is developed by the University of Trento. CtxMatch presents an approach to derive semantic relations between classes of two classification schemas, which are extracted from databases or ontologies. Based on the labels the system identifies equivalent entities. For this, it also makes use of synonyms defined in WordNet. Other element level matchers are also included. Through an SAT-solver the system identifies additional relations between the two schemas. The SAT-solver takes the structure of the schemas into account, especially the taxonomy and its inferred implications, e.g., the fact that any object in a class is also an element of all the superclasses there of. As a result, the system returns equivalence, subsumption, or mismatch between two classes. A recent version S-Match also provides explanations of the alignments.

4.2.3. Silk framework

The Silk frame work is a tool for matching the data from different Linked Open data source. It takes two RDF files (Resource Description Frame work) as input and generate the similarity matrics among the links by using the the string matching techniques. It uses the concept-to-concept matching approach [35].

4.3. Controlled facilitating the Linked Open Data

The key factor of semantic web is a web of data. These data need to be linked for the broader usages of semantic web community. "Linked Data is about using the Web to connect related data that wasn't previously linked, or using the Web to lower the barriers to linking data currently linked using other methods". More specifically, Wikipedia defines Linked Data [32] as "a term used to describe a recommended best practice for exposing, sharing, and connecting pieces of data, information, and knowledge on the Semantic Web using URIs and RDF." The controlled vocabularies play important role for this new dimension of datasharing arena. The most callenges are the data formats (i.e., XML, CVS, txt, etc) and licence policy . In order to publish the data, we need to make the datasources in RDF/XML format and the free licence policy so that anybody can use the published data in their applications.

For example, "AGROVOC thesaurus" is aligned with thirteen vocabularies, thesauri and ontologies in areas related to the domains it covers for joining the LOD. The Six of the linked resources are general in scope: the Library of Congress Subject Headings (LCSH), NAL Thesaurus, RAMEAU (Répertoire d'autorité-matière encyclopedique et alphabetique unifie), Eurovoc, DBpedia, and an experimental Linked Data version of the Dewey Decimal Classification. The remaining seven resources are specific to various domains: GEMET on the environment, STW for Economics, TheSoz is about social science and both GeoNames and the FAO Geopolitical Ontology cover countries and political regions. ASFA covers all aquatic science and the aptly named Biotechnology glossary covers biotechnology. These linked resources are mostly available as RDF/XML resources.

Vocabulary	Coverage	Lang used for link discovery	#matches
EUROVOC	General	EN	1,297
DDC	General	EN	409
LCSH	General	EN	1,093
NALT	Agriculture	EN	13,390
RAMEAU	General (cut on Agri.)	FR	686
DBpedia	General	EN	1,099
TheSoz	Social science	EN	846
STW	Economy	EN	1,136
FAO Geopol. Ontology	Geopolitical	EN	253
GEMET	Environment	EN	1,191
ASFA	Aquatic sciences	EN	1,812
Biotech	Biotechnology	EN	812
GeoNames	Gazeteer	EN	212

Table 1. Resources linked to AGROVOC.

The thesauri were considered in their entirety barring RAMEAU, for which only agriculture related concepts were considered (amounting to some 10% of its 150 000 concepts). Candidate mappings were found by applying string similarity matching algorithms to pairs of preferred labels [34] and by using the Ontology Alignment API [28] for managing the produced matches.

The common analysis language used was English in all cases except the AGROVOC - RAMEAU alignment for which French was used. Table 1 shows, for each resource linked to AGROVOC (column 1), its area of coverage (column 2), the language considered for mapping with AGROVOC (column 3), and the number of matches resulting from the evaluation (column 4).

Candidate links were presented to a domain expert for evaluation in the form of a spread sheet. Once validated the mappings were loaded in the same triple store where the linked data version of AGROVOC is stored. All resulting validated candidate matches were considered to be skos:exactMatch.

The objective when linking AGROVOC to other resources was to provide only main anchors, privileging accuracy over recall. This is why it only used exactMatch, found by means of string-similarity techniques as opposed to more sophisticated context-based approaches. Also, the One Sense per Domain hypothesis [34] supports the claim that in the case similar strings correspond to equivalent meanings. The use of more sophisticated approaches might have contributed to filtering out potential results more than widening their number (thus incrementing precision over recall), however this potential loss of precision was well compensated by the manual validation of candidate links by a domain expert [30]

.

In addition, these secure links from the AGROVOC LOD are used to facilitate the AGRIS[31] which is "a global public domain Database with 2830342 structured bibliographical records on agricultural science and technology. 79.78% of records are citations from scientific journals. The bibliographic references contain either links to the full text of the publication or additional information retrieved from related Internet resources" to network to join the LOD. The Agris Linked Open Data verison is called the "OpenAgris"[32]. The Open Agris uses the AGROVOC links to connect the Dbpedia and extract the information. All the process happens on the fly.

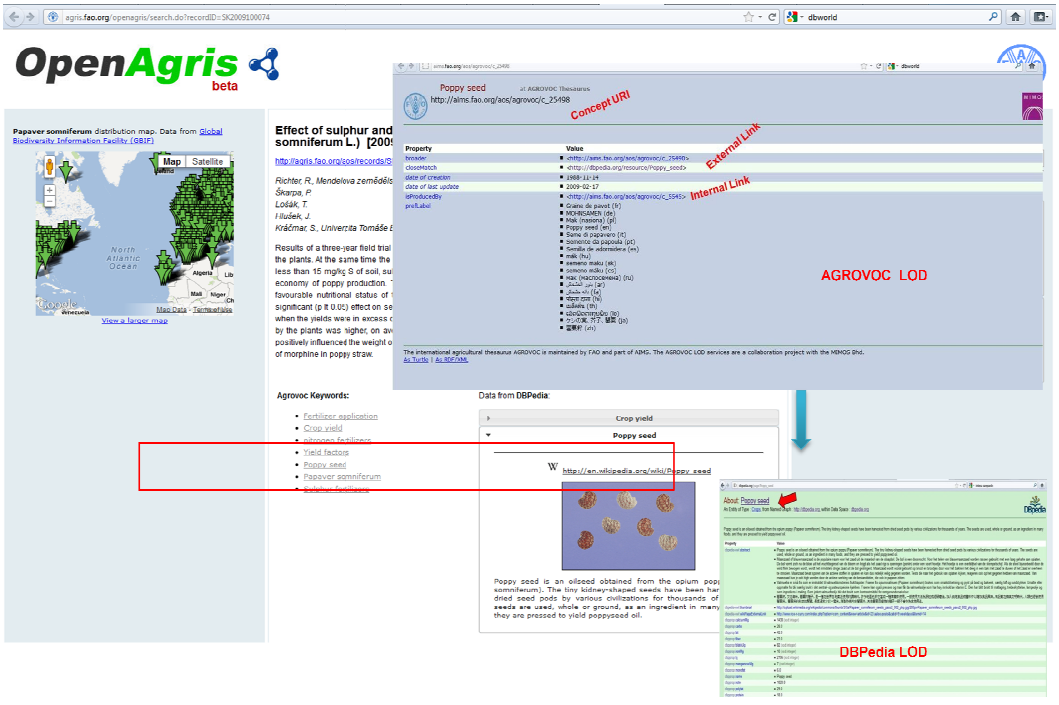

Figure 8. AGROVOC links use for the extracting the information

Finally, we learn a couple of lessons

- The AGROVOC can be a hub of linking data sources and use these links for extracting the information from different data providers
- The most important things that we can classify the information and search them easily by using the CVS.

5. CONCLUSION

Controlled vocabularies are playing vital role for information integration and information retrieval.It can be more useful as linking information, discovery knowledge, and knowledge base in the web. However, a complete universal controlled vocabulary is not yet to be done by any research. It is extremely necessary in the filed of information science, earth science, biological science, cyber science and medical science for common ground of vocabularies so that anyone can access information even he or she does not understand full of language. We have discussed pros and cons different kind of controlled vocabularies and mentioned some on going work on this domain.

ACKNOWLEDGEMENTS

The CSIRO Intelligent Sensing and Systems Laboratory and the Tasmanian node of the Australian Centre for Broadband Innovation are assisted by a grant from the Tasmanian Government which is administered by the Tasmanian Department of Economic Development, Tourism and the Arts. Author also would like to thanks Gudrun Johannsen, Johannes Keizer, and Prof. Fausto Giunchiglia

REFERENCES

[1] Zhu, H., & Madnick, S. (2006).A lightweight ontology approach to scalable interoperability. Working paper CISL, The Massachusetts Institute of Technology,Cambridge, MA ,USA, June 2006.
[2] Faatz, T. Kamps A. & Steinmetz, R.(2000) Background knowledge,indexing and matching interdependencies ofdocument management and ontology maintenance. In Proceedings of the First Workshop on Ontology Learning(OL-2000) in conjunction with the 14th European Conference on Arti_cial Intelligence (ECAI 2000), Berlin,.
[3] Morshed, A.(2009). Controlled vocabulary matching in distributed system. 26th British National Conference on Databases,UK, July 2009.
[4] Morshed.A(2010). Aligning Controlled vocabularies for enabling semantic matching in a distributed knowledge management system. Unpublished doctoral disseration, University of Trento, Trento, Italy.
[5] DMOZ. (2012).Open directory project.Retrieved Jan, 2012, from http://www.dmoz.org/.
[6] Clusty.(2012).MetaSearch Engine. Retrieved March, 2012, from http://clusty.com/.
[7] Vivisimo.(2012). MetaSearch Engine. Retrieved March,2012, from http://vivisimo.com/.
[8] EuroWordNet.(2012). Retrieved Jan,2012, from http://www.illc.uva.nl/EuroWordNet/.
[9] AGROVOC.(2012).AGROVOCThesaurus.RetrievedJan,2012,from http://aims.fao.org/standards/agrovoc/functionalities/search
[12] VocBench.(2012). Retrieved June, 2012, from http://aims.fao.org/tools/vocbench-2.
[13] Flickr.(2012). Retrieved March,2012, from http://www.flickr.com/.
[14] Giunchiglia, F &Shvaiko, P. (2003). Semantic matching. "Ontologies and Distributed Systems" workshop, IJCAI, 2003.
[15] Giunchiglia,F, Shvaiko P, & Yatskevich, M.(2004). S-match: An algorithm and an implementation of semantic matching. In Proceedings of ESWS'04, 2004.
[16] Shvaiko, P. ,Giunchiglia, F. & Yatskevich, M.(2006).Discovering missing background knowledge in onology matching. In 17th European Conference on Artificial Intelligance (ECAI 2006), volume 141, pages 382-386, 2006.

[17] Zhu, H & Madnick,S.(2006). A lightweight ontology approach to scalable interoperability. Working paper CISL, The Massachusetts Institute of Technology,Cambridge, MA ,USA, June 2006.
[18] Gilchrist, A & Aitchison, J& Bawden.(2006). Thesaurus construction and use:a practical manual. 4th ed., page 240, London, 2006. Aslib.
[19] Daphne, K & Sahami,M.(1997). Hierarchically classifying documents using very few words. In Douglas H. Fisher, editor, Proceedings of ICML-97, 14th International Conference on Machine Learning, pages 170–178, Nashville, US, 1997.Morgan Kaufmann Publishers, San Francisco, US.
[20] LCA.(2012).Library Congress Author List. Retrieved Feb ,2012 http://www.loc.gov/bookfest/2009/authors/.
[21] McGuinness, D.L, & Shvaiko, P & Giunchiglia, F & Pinheiro da Silva, P. (2004).Towards explaining semantic matching. In International Workshop on Description Logics at KR'04, 2004.
[22] Miller, G.(1998). WordNet: An electronic Lexical Database. MIT Press, 1998.
[23] LCSH.(2012).The Library of Congress Classification system.March,2012, Retrieved from http://www.loc.gov/catdir/cpso/lcco/lcco.html/.
[24] OVP.(2011). Open Video Project. Retrieved December,2011, from http://www.open-video.org/.
[25] MeSH.(2012)the National Library of Medicine's controlled vocabulary thesaurus. Retrieved April, 2012, from http://www.nlm.nih.gov/mesh/.
[26] YouTube.(2012). Retrieved May, 2012, from http://www.youtube.com/.
[27] Cohen, W.W & Ravikumar, P and Fienberg, S.E.(2003). A comparison of string distance metrics for name-matching tasks, in IJCAI-2003, 2003.
[28] Jérôme, D., Euzenat, J., Scharffe, F., & Cássia Trojahn dos Santos.(2011). The Alignment API 4.0. Semantic Web Journal, vol. 2, no. 1, pp. 3-10, 2011.
[29] Gale, W., Church, K, & Yarowsky, D. (1992)A Method for Disambiguating Word Senses in a Large Corpus. Computers and the Humanities, no. 26, pp. 415-439, 1992
[30] Morshed, A, &Caracciolo, C & Johannsen, G & Keizer, J.(2011). Thesaurus Alignment for Linked Data publishing. Ahsan Morshed, Caterina Caracciolo, Gudrun Johannsen and Johannes Keizer, DCMI International Conference on Dublin Core and Metadata Applications DC-2011
[31] Agris. (2012).Retrieved May, 2012, from http://agris.fao.org/knowledge-and-information-sharing-through-agris-network
[32] OpenAgris.(2012). Retrieved May, 2012, from http://agris.fao.org/news/openagris-journals-rdf-visual-reader
[33] LOD. (2012).Linked Open Data.Retrieved June, 2012, from http://linkeddata.org/
[34] Stoilos, G., Stamou, G., Kollias, S.(2005). A string metric for ontology alignment. In Proceedings of the 4th International Semantic Web Conference, pages 624–637, Berlin, Heidelberg. Springer-Verlag.
[35] Volz, J., Bizer, C., Gaedke M., Kobilarov, G. (2009) Silk – A Link Discovery Framework for the Web of Data . 2nd Workshop about Linked Data on the Web (LDOW2009), Madrid, Spain, April 2009.
[36] Gong Cheng ingsheng Jian, Wei Hu and Yuzhong Qu. Falcon-ao:Aligning ontologies with falcon. In In Proceedings of the International and Interdisciplinary Conference on Modeling and Using Context (CONTEXT), pages 85-91, 2005.
[37] Euzenate, J. & Shaviko, P.(Ed.)(2007). Ontology Matching. Springer, 1st edition, 2007.
[38] Bouquet P., Serafini, L., & Zanobini,S.(2003). Semantic coordination: a new approach and an application. In Proc. of the 2nd International Semantic Web Conference (ISWO'03). Sanibel Islands, Florida, USA, October 2003.
[39] Merabti, T., Soualmia, LF., Grosjean, J., Joubert, M. & Darmoni, SJ. (2003) Aligning Biomedical Terminologies in French: Towards Semantic Interoperability in Medical Applications. Chapitre 3, Medical Informatics, March, Pages 41-68, InTech, 2012 .
[40] Xiaogang Ma, Chonglong Wu, Emmanuel John M. Carranza, Ernst M. Schetselaar,Freek D. van der Meer, Gang Liu, Xinqing Wang, Xialin Zhang.(2010) Development of a controlled vocabulary for semantic interoperability of mineral exploration geodata for mining projects, Computer and Geosciences, Vol. 36, Issue. 12, pp. 1512-1522, 2010.
[41] SEMEDA.(2003). ontology based semantic integration of biological databases, Jacob Köhler, Stephan Philipp and Matthias Lange, Vol. 19 no. 18 2003, pages 2420–2427.
[42] Cuahsi.(2012). Retrieved 5 June, 2012 http://his.cuahsi.org/mastercvdata.html
[43] Srinubabu, G.(2011). Integration, Warehousing, and Analysis Strategies of Omics Data, Methods in Molecular Biology, 2011, Volume 719, Part 3, 399-414,

[44] Liang, H., Xu, Y. and Nayak, R. (2009) Personalized Recommender Systems Integrating Social Tags and Item Taxonomy. In: 2009 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops, September 15-18, 2009, Milano, Italy.
[45] Swoogle.(2012).Semantic Search Engine. Retrieved March, 2012, from http://swoogle.umbc.edu/.
[46] Map.(2012). GoogleMaps. Retrieved June, 2012, from https://maps.google.com.au/
[47] Sebastiani, F.(2002) Machine learning in automated text categorization. ACM Computing Surveys, 34(1):1–47, 2002.
[48] McCulloch, E.(2005). Digital direction thesauri:practical guidance for construction. Volume 54, 2005.
[49] Ibekwe-SanJuan,F.(2006) Construction and maintaining knowledge organization tools a symbolic appraoch. volume 62, 2006

Authors

Dr. Ahsan Morshed is Postdoctoral fellow at CSIRO. Before joining to CSIRO, he was an information management specialist at, FAO of UN, Rome, Italy. He is author of 19 publications and member of 4 scientific committees. He is a member of DC task group. His interest is in semantic web and Linked Open Data.