Symbiosis of thesaurus, domain expert and reference sources in designing a framework for the construction of a multilingual ontology for Islamic Portal

Juhana Salim, Siti Farhana Mohamad Hashim, and Shahrul Azman Mohamad Noah

Knowledge Technology Research Group, Centre for Artificial Intelligent Technology,
Universiti Kebangsaan Malaysia, Selangor, Malaysia
js@ftsm.ukm.my, fana3687@yahoo.co.uk, samn@ftsm.ukm.my

Abstract: This paper discusses the conceptual and vocabulary problems users faced when searching the web and subsequently shows how a well structured thesaurus can be used as knowledge base for an interface that can assist user with search topic clarification. Recent research justified that thesaurus is useful in building ontology to help standardize terminologies and more importantly help to save time in building ontology that fully depend on domain expert. Several initiatives has been undertaken in web'ifying thesaurus with the idea of converting thesaurus or controlled vocabularies to semantic web standards such as Web Ontology (OWL). The aim of this research is to explore how thesaurus can be integrated into ontology development. Our research method emphasizes on semantic web technologies and the development of ontology using thesaurus, domain expert and reference sources such as Index Islamicus, encyclopedia, biographies etc. as the basis for implementing novel mechanism for retrieving Islamic web in 3 different languages simultaneously. As a result, we developed a framework for the development of Islamic ontology that advocates the symbiosis of thesaurus, domain expert and several reference sources in Islam as the basis for the construction of a multilingual ontology for our Islamic portal.

Keywords: thesaurus, multilingual ontology, Islamic retrieval system, ontology development

1. Introduction

In recent years, there were a variety of difficulties that users have to face in searching for information through Internet. Past research had shown that in performing searches, users use query terms that consist of "everyday language, technical terms (with or without knowledge of underlying concepts) and various explanatory model, all influenced by psychosocial and cultural variations." [1].Very often users failed to retrieve relevant information that meet their needs and this could be due to the conceptual and vocabulary problems that users faced when searching the web. The first part of this paper elucidates queries illustrating these problems and how a well structured thesaurus can be used as knowledge base for an interface that can assist user with search topic clarification. We then explain how thesaurus helps users to identify the right appropriate concepts to use when searching for information through Internet. However, if we want to create a knowledge-rich description of, for example an (image of an) art object, medical, business, crime etc. such as required by the "semantic web", thesauri turn out to provide only part of the knowledge needed. The second part of this paper discussed various controlled vocabularies, classification schemes and thesauri that can serve as some building blocks of the semantic web. These vocabularies have been developed over decades and can be used in the development of robust web services and Semantic Web technologies. Thus the third part of this paper, addresses initiatives undertaken to 'web'ifying' thesaurus with the idea of converting thesaurus/controlled vocabularies to semantic web standards. Since the aim of this research is to develop ontology using symbiotic approach involving thesaurus, then fourth and fifth part of this paper discuss the concept of ontology and thesaurus. Part six analysed the

Received: April 11th, 2011. Accepted: December 15th, 2011

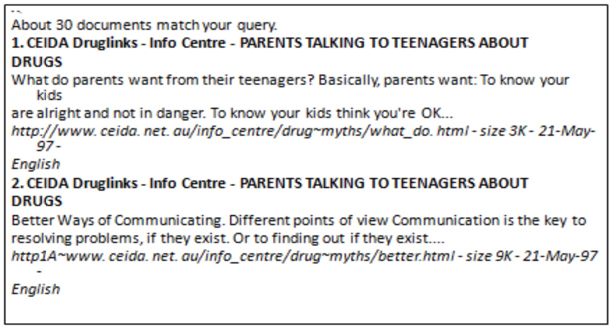
variety of projects that have used thesaurus and summarised the steps needed in order to create domain ontology. Based on the analysis and synthesis of the past related projects, in part seven, we explained the symbiotic method that we used in our construction of multilingual domain ontology related to Islamic Act of Worship. Part eight provides the conclusion.

2. Conceptual and Vocabulary Problems When Searching the Web.

In accessing information through the web, there are several problems that we often face. Today's search tools perform rather poorly in the sense that information access is mostly based on keyword searching or even mere browsing of topic areas. This unfocused approach often leads to undesired results. The first example illustrates a query (Query 1.1) whereby a user wants to find information on 'Drug use' by teenagers and the result that did not retrieve relevant documents. The second example (Query 1.2) shows the synonyms expansion for teenage where the use of thesaurus would help user in synonym expansion, thus improving search strategy.

Example 1:

Query 1.1. teenage* AND drug* (AltaVista)



About 30 documents match your query.

1. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS

What do parents want from their teenagers? Basically, parents want: To know your kids
are alright and not in danger. To know your kids think you're OK...

http://www. ceida. net. au/info_centre/drug~myths/what_do. html - size 3K - 21-May-97 -

English

2. CEIDA Druglinks - Info Centre - PARENTS TALKING TO TEENAGERS ABOUT DRUGS

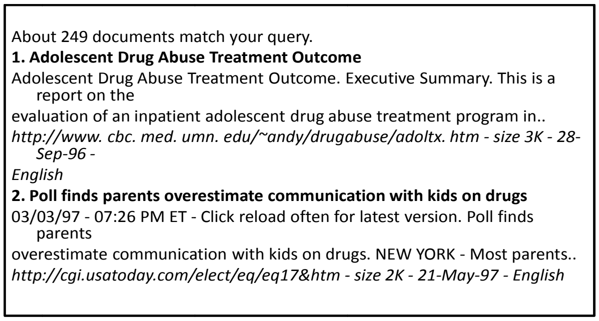
Better Ways of Communicating. Different points of view Communication is the key to
resolving problems, if they exist. Or to finding out if they exist....

http1A~www. ceida. net. au/info_centre/drug~myths/better.html - size 9K - 21-May-97 -

English

Example 2:

Query 1.2. Synonym expansion of teenager
(teenage* OR teen OR teens OR youth OR adolescent* OR kid* OR "high school")AND drug*



About 249 documents match your query.

1. Adolescent Drug Abuse Treatment Outcome

Adolescent Drug Abuse Treatment Outcome. Executive Summary. This is a report on the
evaluation of an inpatient adolescent drug abuse treatment program in..

http://www. cbc. med. umn. edu/~andy/drugabuse/adoltx. htm - size 3K - 28-Sep-96 -

English

2. Poll finds parents overestimate communication with kids on drugs

03/03/97 - 07:26 PM ET - Click reload often for latest version. Poll finds parents
overestimate communication with kids on drugs. NEW YORK - Most parents..

http://cgi.usatoday.com/elect/eq/eq17&htm - size 2K - 21-May-97 - English

Similarly in accessing Islamic information through the Net, there are several problems that users face. One might, for example, want to find out which organization established the Halal Food. A simple search for 'Halal Food' might result in a huge list of documents containing these words, but actually none of them containing the desired result: i.e. IFANCA or Islam Food and Nutrition Council of America that lists organizations that established Halal Food. The problem become even worse, if the result searched only appears in a foreign language.

This example shows how ontologies can help to improve the management of information. Semantically annotated documents, i.e. documents which are indexed with ontological terms and concepts instead of simple keywords, provide several advantages. Numerous ontology that have been constructed can assist user with search topic clarification. In relation to ontology construction, thesaurus can be used as knowledge base for an interface that can assist user with search topic clarification.

3. Thesaurus as Knowledge Base

The third part of this paper shows how well structured thesaurus can be used as knowledge base for an interface that can assist user with search topic clarification. By referring to the Art and Architecture Thesaurus, a user who seeks to search for information that will give a better understanding of 'drawings of mural in 'public places', can use 'public art' as suggested in the Art and Architecture Thesaurus. A thesaurus is a structure that manages the complexities of terminology and provides conceptual relationships, ideally through an embedded classification/ontology. A thesaurus may specify descriptors authorized for indexing and searching. These descriptors form a controlled vocabulary (authority list, index language). The following list provides additional thesaurus and classification schemes illustrating different functions:

(i). **HS Harmonized Commodity Description and Coding System World Customs**
Info: http://pacific.commerce.ubc.ca/trade/HS.html

(ii). **NAICS North American Industrial Classification System**
Info:www.census.gov/epcd/www/naics.html, www.naics.com

(iii). **ICD-10 The International Statistical Classification of Diseases and Related Health Problems, tenth revision.**
Info:www.who.int/whosis/icd10/index.html,www.cdc.gov/nchs/about/major/dvs/icd10des.htm

(iv). **CPT Physicians' Current Procedural Terminology. CPT 2003.**
Info:http://www.amaassn.org/ama/pub/category/3113.html, listing of codes https://webstore.ama-assn.org/index.jhtml

However, if we want to create a knowledge-rich description of for example an (image of an) art object, medical, business, crime etc. such as required by the "semantic web", thesauri turn out to provide only part of the knowledge needed. Zeng et al. [2] identified the factor which caused search results that are not relevant is because the terms and concepts used, often do not accurately reflect users' information needs and therefore do not constitute effective queries. Problems with specific domain vocabularies may occur at various kinds namely lexical form and mismatches such as misspellings, semantic misunderstanding and misleading mental models. Various controlled vocabularies, classification schemes and thesauri can serve as some building blocks of the semantic web. These vocabularies have been developed over decades and can be used in the development of robust web services and Semantic Web technologies. According to Harper and Tillett [3] several initial collaboration between semantic Web, Library and metadata communities are creating partnerships to complete work in this area. The Semantic Web and library communities have both been working on naming concepts, entities and bringing different forms of those manes together. In fact, libraries communities have been working on naming entities referred to as identification of subject headings, for more than hundreds of years. There are sophisticated and advanced tools and controlled vocabularies like Library Subject Headings and Classification, Medical Subject Headings developed by the library of Congress. When these tools are referred and applied and translated into semantic web technologies will help realize Berner's Lee's vision [4] : "*I have a dream for the Web [in which computers] become capable of analyzing all the data on the Web – the content, links, and transactions between people and computers. A Semantic Web, which should make this possible,* *has yet to emerge, but when it does, the day-to-day mechanisms of trade, bureaucracy and our daily lives will be handled by machines talking to machines. The intelligent agents' people have touted for ages will finally materialize*." The Semantic Web Technologies is frequently described in terms of Semantic web stack which are dependent on layers below it (Figure 1).

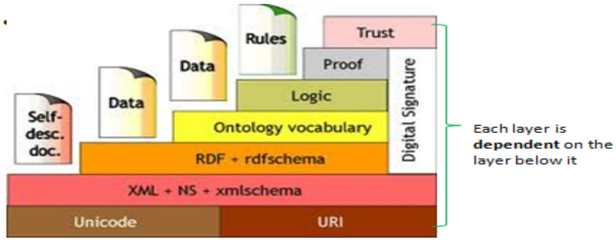


Figure 1. Semantic Web Technology
Source: Antoniou, G. and Harmelen, F.V. [5]

Development on various levels depicted has been in the making for quite a long time. Berners- Lee identified the slow development is due to the query and rules work that was held up because people did not want to start on it until the ontology work had finished. Thus, initiatives can be undertaken by web'ifying thesaurus with the idea of converting thesaurus, controlled vocabularies such as Library of Congress Subject Headings, Search Lists in Dewey decimal classification system to semantic web standards such as Web Ontology (OWL). Such efforts which is outside the scope of W3C, will provide limitless potential especially the integration of research functionality such as searching and browsing diverse resources, verification of a resource's author, browsing related to a particular concept into online reference sources, library catalogues to authoring tools [3]. Amongst the Subject and Genre Vocabularies for the Semantic Web includes:

(i). Alcohol and Other Drug Thesaurus (AOD Thesaurus: US Nat. Inst. Of Alcohol Abuse and Alcoholism
(ii). Medical Subject Heading (MeSH) and Unified Medical Language System (UMLS): US National Library of Medicine
(iii). Art and Architecture Thesaurus (AAT): Getty Foundation
(iv). Dewey Decimal Classification : US Library of Congress and OCLC/Forest Press
(v). Library of Congress Subject Heading: : US Library of Congress
(vi). WordNet: Pronceton University. George Miller
(vii). CYC Ontology :CYC Corporation

These vocabularies present tremendous potential like improving access to web resources and Semantic web data, enhanced network applications and improve search engine result. Amongst the variety of project that help to bring more of the tools that libraries developed into semantic web and web services involves the symbiosis of thesaurus and ontology.

4. What is Ontology?

Ontology can act as backbone in information retrieval system where it can search information that fulfill the user's need and can provide related information to the user's search domain. This can expand the user's knowledge in searching related information. The word ontology was taken from Philosophy, where it means a systematic explanation of being. In 2000, the word ontology itself became a relevant word for the Knowledge Engineering community. The earliest definition of ontology was given by Neches and colleagues [6], who defined ontology as follows: "ontology defines the basic terms and relations to define the extensions to the vocabulary". A few years later, Gruber [7] defined ontology as "an explicit specification of a conceptualization". Explicit means that the type of concept used, and the constraints on their use are explicitly defined. Based on Gruber's definition, many definitions of ontology are proposed. Borst modified slightly Gruber's definition to: "Ontologies are defined as formal specification of a shared conceptualization". Formal refers to the fact that the ontology should be machine-readable and shared reflects the notion that an ontology captures consensual knowledge, that is, it is not private of some individual, but accepted by a group". Then Gruber's and Borst's definition have been merged and explained by Struder and colleagues [8] as follows: "conceptualization refers to an abstract model of some phenomenon in the world by having identified the relevant concepts of that phenomenon.

Ontology in the context of Agricultural Ontology Service [9] is a system of terms, the definition of these terms and the specification of relationships between the terms. This definition extends the approach of classical thesauri by providing the opportunity for creating an infinite number of different semantic relationships. There are several advantages in using ontology for information retrieval. First, there will be no changes in content of documents. Second, the process of annotating the document is performed by ontological tools which are based on a specific domain. It means that the semantic meanings and interpretations of keywords are bound to that domain and this will result in a more efficient retrieval process. Third, document specific representations no longer affect the search. This is important in the case of multilingual representations because there is the need to produce the same results, no matter which language is used for retrieval. It can be done by mapping the keywords in several languages to the same concept in an ontology thus giving the same meaning to those keywords.

Furthermore, based on the ontology provided, user can get the websites that relate to the user's query. It means that user can retrieve websites that are outside the given keyword. An example is the word that has the same meaning (synonym) or refers to the same concept like in the case for 'kahwin', 'nikah' and 'marriage'. Ontology can handle the keyword like concept in database by interpreting semantic relations between keywords or within table in the field of database. Ontology is also suitable for sharing information in a distributed environment.

5. What is Thesaurus?

Thesaurus is a book that lists words in groups that have similar meanings [10] According to Xing et al., [11] a standard thesaurus contains terms and simple semantic relationships such as classification and hierarchical relationships in which it is common to the ontology. This statement is supported by the research conducted by Xing et al. [11] which states that the ontology contains information about concepts, semantic relationship between concept, instances and axioms of a domain. The presence of semantic relationships in the thesaurus makes it as a reason why it is often associated with the development ontology. According to Chang et al. [12], they found similarities between the ontology and thesaurus. Then similarities are:

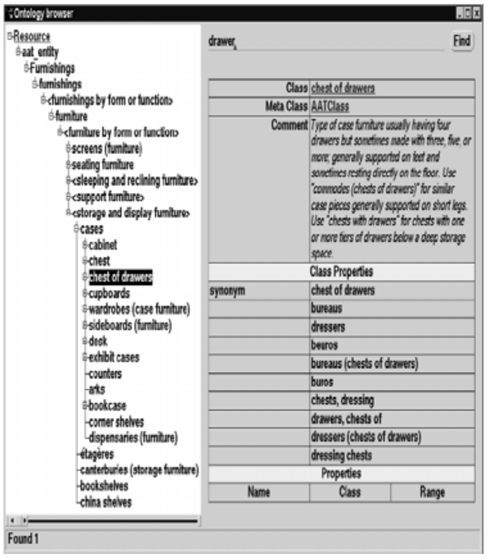
(i). Describe the specific domain of knowledge;
(ii). Contains terms and relations between terms;
(iii). Can be used by people in the information management application to catalog and retrieve information;
(iv). Should be maintained and reviewed on an ongoing basis.

However, Chang et al. [12] also found that thesaurus provides a rough relationship without a clear explanation. It can be seen through the difference between a thesaurus and ontology. For the thesaurus it has a limited number of relationships such as hierarchical which refers to Broad term (BT) and narrow term (NT) and non-hierarchical which refers to related term (RT). In the case of ontology, the attributes and relations are not limited to the ones used in a thesaurus. All relationships to clarify the domain associated with the term can be used. These relations which are used to distinguish between the concepts are important in order to enable computers to intelligently search knowledge.

6. Thesaurus as a tool in ontology development

In early 2000, there are some initiatives to use thesaurus as a tool in ontology development because according to the research conducted by Xing et al. [11], thesaurus can provide standardization of term. Besides that, thesaurus is helpful to formalize the domain concepts and establish a scientific hierarchy Xing et al.. Some of the researchers who used thesaurus in their ontology development include Weilinga et al. [13], Chang et al. [12], Lauser et al. [14] and Xing et al. [11]

Weilinga et al. [13]used Art and Architecture Thesaurus in order to build Antique Furniture ontology. Firstly, they treated the main terms as concept names in the knowledge base. The full AAT hierarchy was converted into a hierarchy of concepts, where each concept has a label slot corresponding with the main term in AAT and a synonyms slot where alternate terms are represented. The knowledge base is represented in RDFS by constructing an RDFS browser to inspect and browse the hierarchy (Figure 2)



```
Resource
  aat_entity
    Furnishings
      furnishings
        <furnishings by form or function>
          furniture
            <furniture by form or function>
              screens (furniture)
              seating furniture
              <sleeping and reclining furniture>
              <support furniture>
              <storage and display furniture>
                cases
                  cabinet
                  chest
                  chest of drawers
                  cupboards
                  wardrobes (case furniture)
                  sideboards (furniture)
                  desk
                  exhibit cases
                  counters
                  arks
                  bookcase
                  corner shelves
                  dispensaries (furniture)
                étagères
                canterburies (storage furniture)
                bookshelves
                china shelves
```

Class	chest of drawers
Meta Class	AATClass
Comment	*Type of case furniture usually having four drawers but sometimes made with three, five, or more; generally supported on feet and sometimes resting directly on the floor. Use "commodes (chests of drawers)" for similar case pieces generally supported on short legs. Use "chests with drawers" for chests with one or more tiers of drawers below a deep storage space.*
Class Properties	
synonym	chest of drawers
	bureaus
	dressers
	beuros
	bureaus (chests of drawers)
	buros
	chests, dressing
	drawers, chests of
	dressers (chests of drawers)
	dressing chests

Properties

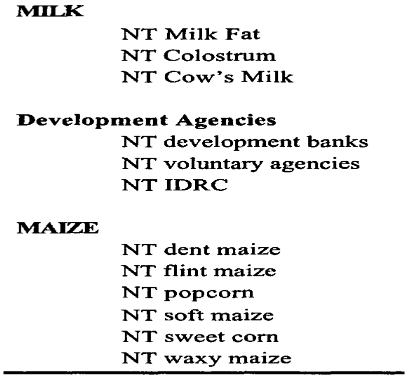
Name	Class	Range

Figure 2. Part of AAT hierarchies in RDFS Browser
Source: Weilinga et al. [13]

For the second step, they augment a number of concepts with additional slots and fillers. For example, concepts representing a style or period were augmented with slots time period from, time period to, general style and region. The values for these slots were partly derived using explicit tables of periods, and partly by using the intermediate concepts in AAT. Lastly, they add knowledge about the relation between possible values of fields and nodes in the knowledge base by using WordNet and special purpose documents.

Besides Weilinga et al. [13], Chang et al. [12] also discussed the relationship between thesaurus and ontology. As a starting point, Chang explained the hierarchical relationships in the thesaurus as referring to BT/NT (Broad Term/ Narrow Term), RT (Related Terms) and UF (Use For)/USE and how these relations can be converted to property, synonyms and partitive relation. To develop the ontology, Chang used a tool named KAON. KAON is a kind of open source tools that manages the ontology infrastructure for business applications. In KAON, RT is considered as property while preferential relation (UF/USE) is treated as synonym in the ontology. Other than that, BT/NT is regarded as a partitive relation.

In the AGROVOC, Lauser et al. [14], Broad Term (BT) and Narrower Term (NT) refer to a form of hierarchical relationship. However, the semantic relationship is not clearly stated as shown in Figure 3.



MILK

NT Milk Fat
NT Colostrum
NT Cow's Milk

Development Agencies

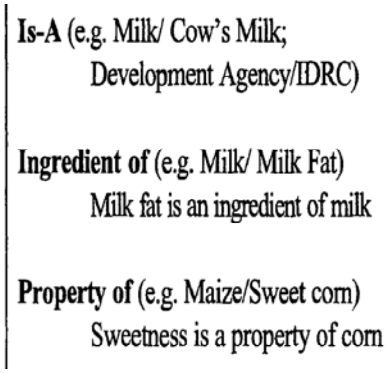
NT development banks
NT voluntary agencies
NT IDRC

MAIZE

NT dent maize
NT flint maize
NT popcorn
NT soft maize
NT sweet corn
NT waxy maize

Figure 3. Example of NT in AGROVOC
Source: Lauser et al. [14]

Based on Figure 3 the relation of BT/NT in AGROVOC thesaurus Lauser et al. can be interpreted as shown in Figure 4.



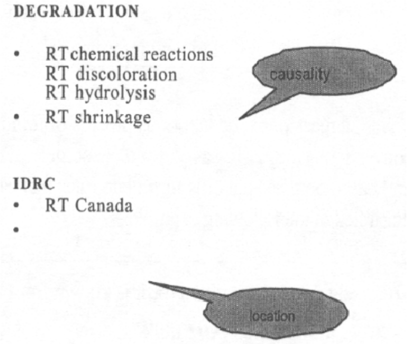
Is-A (e.g. Milk/ Cow's Milk;
Development Agency/IDRC)

Ingredient of (e.g. Milk/ Milk Fat)
Milk fat is an ingredient of milk

Property of (e.g. Maize/Sweet corn)
Sweetness is a property of corn

Figure 4. Interpretation of BT/NT in AGROVOC
Source: Lauser et al. [14]

RT in AGROVOC which represents the similarity relationship (associative relation) involving the semantics appears to be unclear. RT in AGROVOC can be referred to as cause, agency or instrumentality, the features of an object of the action or disciplinary process, the sequence of time or space and antonyms. Examples of RT in AGROVOC can be seen in figure 5.



DEGRADATION

- RTchemical reactions
  RT discoloration
  RT hydrolysis
- RT shrinkage

IDRC

- RT Canada
-

Figure 5. Example of RT in AGROVOC
Source: Lauser et al. [14]

There are several steps taken to re-engineering AGROVOC which involves building ontology. The first step involved analyzing the existing relationship to establish the semantic relationships more clearly. For example, the relationship BT/NT can be converted to “is-a” form relationship by default and can be re-interpreted to relate to others with as much as possible when needed. RT relationships can be re-interpreted back to more specific relationships, such as “produces”, “used by” and “made for”. Apart from analyzing the relationship among them, there are steps being taken to determine composite concept in basic concept which can be represented clearly. For example, ‘perishable product’ can be represented as product by having perishable as attribute.

Thesaurus has been applied in the development of ontology in research conducted by Xing et al. [11]. In their study, they developed domain ontologies related to China travel using the Classified Chinese Thesaurus. According to Xing et al. [11], there are four key elements in the development of ontology which are Terms, Hierarchies, Semantic Network and Ability Reasoning. However, the thesaurus participation in the development of ontology only involved two elements which are ‘Terms’ and ‘Hierarchies’.

The ‘Terms’ refers to the basic concepts of a domain. It is also supported by the definition of ontology provided by Swartout et al. [15] that is the ontology is a hierarchically structured set of terms to describe a domain that can be used as the basis for a knowledge base. To ensure that the terms used meets the standard, Xing et al. suggest using thesaurus. However the use of thesaurus cannot explain the specific field more clearly. According to domain experts, they need to make the addition of new terms to clarify some important concepts that does not exist in the thesaurus. As a solution, they defined two kinds of terms which are conceptual terms and abstract terms. Conceptual term refers to the conceptual classes that are taken from thesaurus while abstract term refers to the abstract classes provided by domain experts. For example, the term “Xiang Dong Temple” and “Dai Luo Ding” are terms that are important in the tourism domain, but the term is not found in the thesaurus. At the same time, the term translated from thesaurus does not necessarily satisfy the field of ontology that is to be developed. In this situation, changes must be made based on the developed domain ontology.

Hierarchies is a relationship between concepts such as “is-a”, “kind-of” and “part-of”. Hierarchies also refer to the inheriting relationship and if they are in different classes, it refers to the combination of relationship such as: intersection; union; inverse set and complementary set to other classes. In this way, the term (concept) can be connected as hierarchy as shown in Figure 6 which shows a hierarchy of “landscape”.

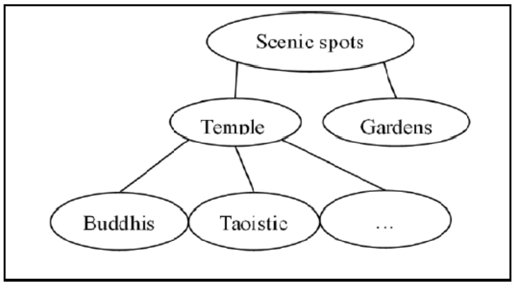


Figure 6. Part of “landscape” hierarchy
Source: Xing et al. [11]

The terms which is in Figure 7, can be formalized as:
Is-Kind-Of (Scenic spots, Temple)
Is-Kind-Of (Scenic spots, Gardens)
Is-Kind-Of (Buddhist Temple, Temple)
Is-Kind-Of (Temple, Taoistic Temple)

There are two types of changes in the hierarchical thesaurus; changing the hierarchy according to domain knowledge and dividing new hierarchy according to domain knowledge in which the thesaurus did not divide specifically (Figure 7). After undergoing several changes, it can be used in the ontology.

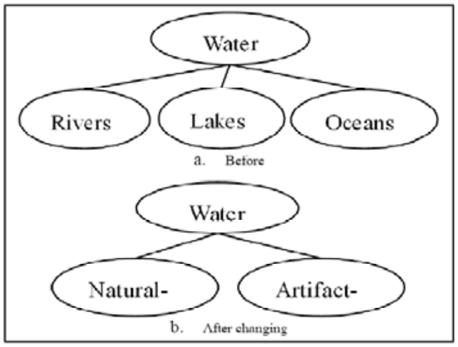


Figure 7. Changes in the hierarchy
Source: Xing et al. [11]

From the analysis on the researches that have used thesaurus in ontology development it can be synthesized that they have similar approach in changing the term relations in thesaurus such as UF, BT, NT and RT in the form of ontology. The difference we can see is how the thesaurus is applied in the development of ontology. Basically, Xing et al. [11] developed ontology manually while Weilinga et al. [13] developed ontology automatically. As an implication to this research, thesaurus can be implemented in the development of domain ontology. According to Xing et al. [11], the ontology development using thesaurus has an advantage whereby an ontology built by thesaurus has a very good extendibility which can avoid wide change when people want to update the ontology later.

7. Methodology

This research emphasizes on semantic web technologies, ontology and thesaurus such as Library Congress Subject Heading and other reference sources such as Index Islamicus, thesaurus, encyclopedia, biographies etc. as the basis for implementing novel mechanism for retrieving Islamic web in 3 different languages simultaneously.

In this research, existing knowledge sources such as documents, reports, etc. are mapped into the domain ontology and semantically enriched. This semantically enriched information enables better knowledge indexing and searching process and implicitly a better management of knowledge. An ontology based system will be used not only to improve precision but also search time. Due to these reasons, ontology based approach will be the core technology for the development of a framework for building multilingual domain ontologies for Islamic Portal. The framework can be viewed graphically in Figure 8

The first approach to develop the Islamic ontology is by using extraction of web content. As this research presents our ongoing work in establishing multilingual domain ontology for Islamic portal that we had developed in the first phase, we therefore use Islamic Extraction and Retrieval System (I-ES™) to extract the content of authoritative web pages. Before the extraction of the web pages, we need to identify which web pages are authoritative web pages. So, there are several steps in identifying the authoritative web pages. In this research, the researcher has a meeting with domain expert from the Faculty of Islamic Studies in Universiti Kebangsaan Malaysia. The purpose of this meeting is to get her opinion on the criteria of the authoritative Islamic web pages need to have. According to Evfi [16], some of the criteria for an authoritative Islamic web pages includes: the information must be clarified clearly and accurately, must be relevant and logic to the subject and must be based on Al-Quran dan As-Sunnah.

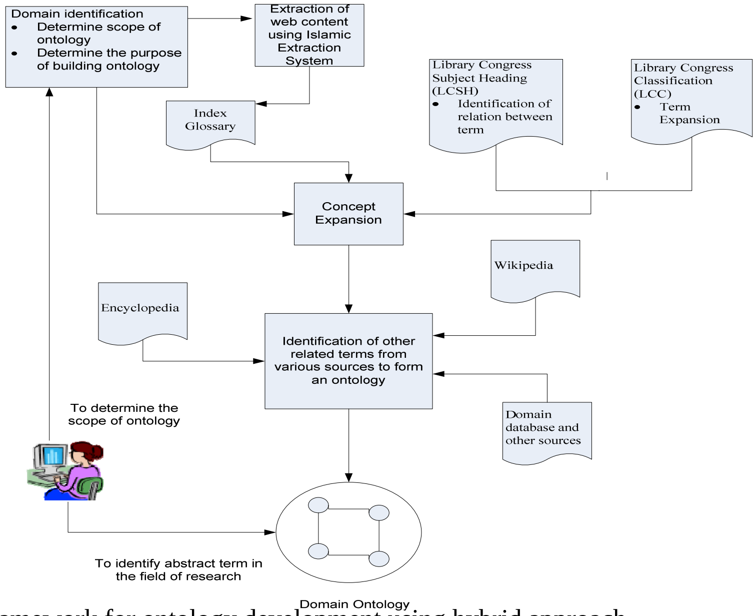
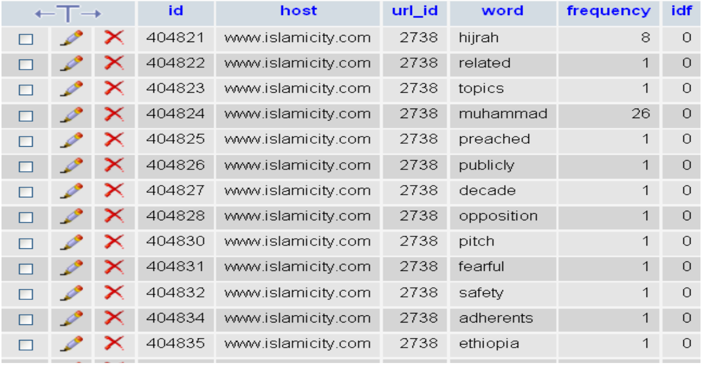

Domain identification
Determine scope of ontology
Determine the purpose of building ontology
Extraction of web content using Islamic Extraction System
Index Glossary
Library Congress Subject Heading (LCSH)
Identification of relation between term
Library Congress Classification (LCC)
Term Expansion
Concept Expansion
Wikipedia
Encyclopedia
Identification of other related terms from various sources to form an ontology
Domain database and other sources
To determine the scope of ontology
To identify abstract term in the field of research
Domain Ontology

Figure 8. A framework for ontology development using hybrid approach

After choosing the authoritative Islamic web pages, there are several stages involved in our ontology development method employed. Firstly, the system extracts the words contained in the websites and remove the stopwords in order to enable word index to be generated (Figure 9).

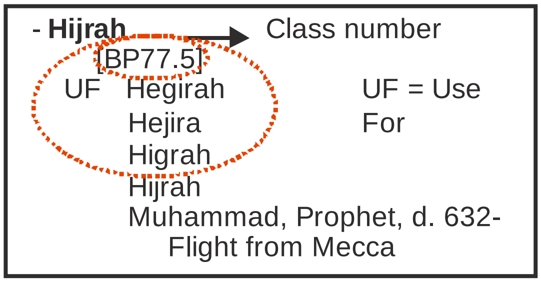


←T→			id	host	url_id	word	frequency	idf
☐	✎	✕	404821	www.islamicity.com	2738	hijrah	8	0
☐	✎	✕	404822	www.islamicity.com	2738	related	1	0
☐	✎	✕	404823	www.islamicity.com	2738	topics	1	0
☐	✎	✕	404824	www.islamicity.com	2738	muhammad	26	0
☐	✎	✕	404825	www.islamicity.com	2738	preached	1	0
☐	✎	✕	404826	www.islamicity.com	2738	publicly	1	0
☐	✎	✕	404827	www.islamicity.com	2738	decade	1	0
☐	✎	✕	404828	www.islamicity.com	2738	opposition	1	0
☐	✎	✕	404830	www.islamicity.com	2738	pitch	1	0
☐	✎	✕	404831	www.islamicity.com	2738	fearful	1	0
☐	✎	✕	404832	www.islamicity.com	2738	safety	1	0
☐	✎	✕	404834	www.islamicity.com	2738	adherents	1	0
☐	✎	✕	404835	www.islamicity.com	2738	ethiopia	1	0

Figure 9. List of word index

The word index is further expanded using thesaurus and other reference sources such as Library of Congress Subject Heading (LCSH) and Library of Congress Classification (LCC) BP Schedules. For example, for the word ‘hijrah’ obtained from the extraction process, the second step involves using established thesaurus. In this research we use LCSH. Figure 10 shows an example of the phrase ‘hijrah’ in LCSH.

All terms of the thesaurus are converted to classes (concepts). The Broader and Narrower Term relationships are used to form the hierarchical class-subclass structure, which constitutes the basic taxonomy of the ontology. Next, the Related Term and Use is represented as properties from these classes thus creating an initial set of non-hierarchical relation. Figure 11 shows an example of ontology developed after referring to LCSH. In this example, the terms that contain in ‘Use’ can be converted as synonym to the term ‘hijrah.’



- **Hijrah** → Class number
[BP77.5]
UF Hegirah   UF = Use For
Hejira
Higrah
Hijrah
Muhammad, Prophet, d. 632-
Flight from Mecca

Figure 10 Word ‘hijrah’ from LCSH

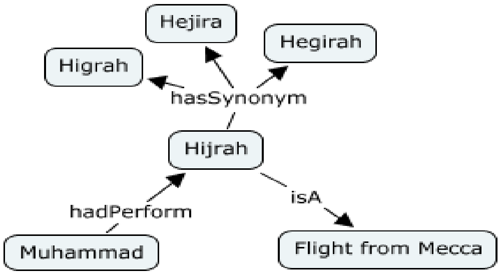
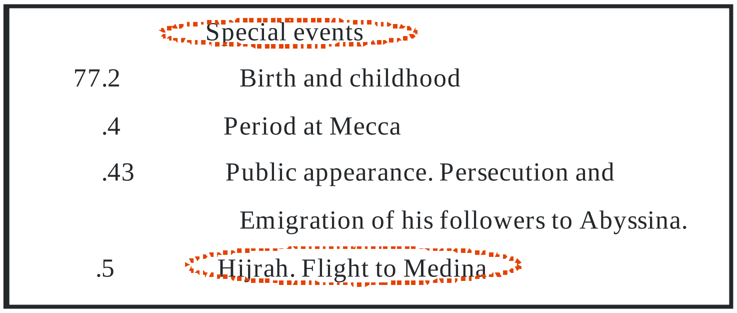


Figure 11 Output after referring to LCSH

In the third stage, we referred to the Library of Congress Classification (LCC) specifically class BP 77.5 (Figure 11) to identify and add more relations/properties to further develop our domain ontology.



Special events

77.2	Birth and childhood
.4	Period at Mecca
.43	Public appearance. Persecution and Emigration of his followers to Abyssina.
.5	Hijrah. Flight to Medina

Figure 11. Class BP 77.5 in LCC

Other reference source such as Index Islamicus is used to further expand our domain ontology. Figure 12 shows the page after referring to LCC and Index Islamicus.

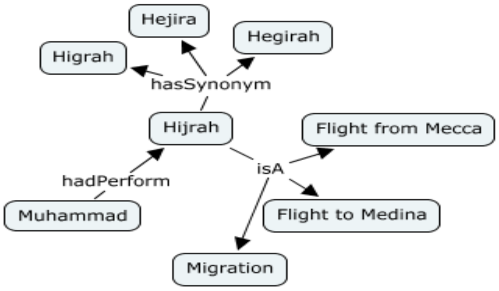
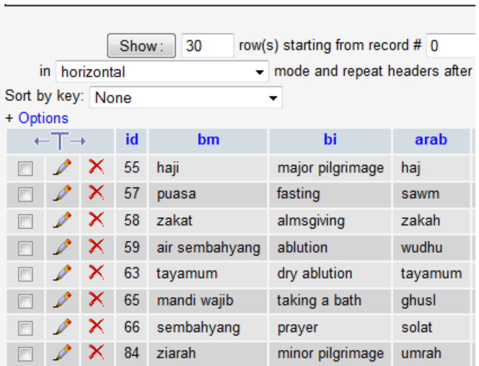


Figure 12. Output after referring to LCC and Index Islamicus

After adding more properties to the concepts using Index Islamicus, we used other reference sources such as encyclopaedias, handbook, almanac etc. To further expand our domain ontology, we extended to other knowledge acquisition processes which involve meetings with an expert to look for abstract concepts (concepts which do not cover by thesaurus and other sources).

As this portal to be developed is a multilingual portal involving Malay, English and Arabic, we translated the word index that we have generated using Google Translate and Kamus Al-Irshad Empat Bahasa [17] which is a Malay, English, Arabic, Urdu/Hindi dictionary. We then used an online transliteration tool, Transliterating Arabic to English in One Step (http://stevemorse.org/arabic/ara2eng.html) to transliterate Arabic characters to Roman characters and stored the list of transliteration term in a dictionary as shown in Figure 13.


Show : 30 row(s) starting from record # 0

in horizontal mode and repeat headers after

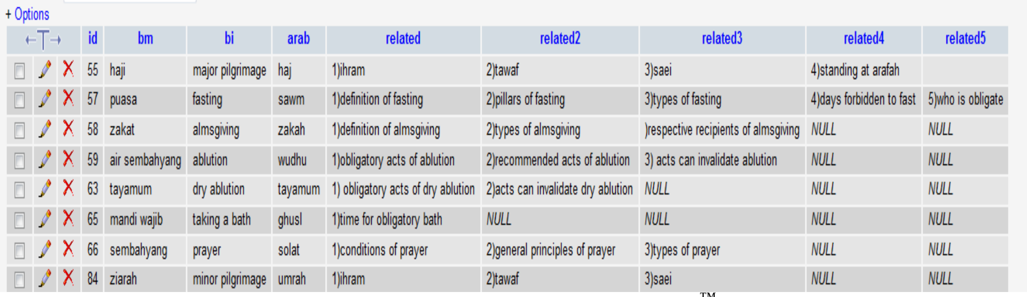
Sort by key: None

+ Options

←T→			id	bm	bi	arab
□	✎	✗	55	haji	major pilgrimage	haj
□	✎	✗	57	puasa	fasting	sawm
□	✎	✗	58	zakat	almsgiving	zakah
□	✎	✗	59	air sembahyang	ablution	wudhu
□	✎	✗	63	tayamum	dry ablution	tayamum
□	✎	✗	65	mandi wajib	taking a bath	ghusl
□	✎	✗	66	sembahyang	prayer	solat
□	✎	✗	84	ziarah	minor pilgrimage	umrah

Figure 13. List of transliteration term

In modeling the ontology, firstly, we defined class, attributes and data values. After that, we transformed class, attributes and data value using the ontology editor, Top Braid Composer. At this stage, the ontology is developed. Lastly, before we embedded the newly developed ontology into I-ES™, we transformed this ontology modeling into relational database in I-ES™ as shown in Figure 14.


+ Options

←T→			id	bm	bi	arab	related	related2	related3	related4	related5
□	✎	✗	55	haji	major pilgrimage	haj	1)ihram	2)tawaf	3)saei	4)standing at arafah	
□	✎	✗	57	puasa	fasting	sawm	1)definition of fasting	2)pillars of fasting	3)types of fasting	4)days forbidden to fast	5)who is obligate
□	✎	✗	58	zakat	almsgiving	zakah	1)definition of almsgiving	2)types of almsgiving	)respective recipients of almsgiving	NULL	NULL
□	✎	✗	59	air sembahyang	ablution	wudhu	1)obligatory acts of ablution	2)recommended acts of ablution	3) acts can invalidate ablution	NULL	NULL
□	✎	✗	63	tayamum	dry ablution	tayamum	1) obligatory acts of dry ablution	2)acts can invalidate dry ablution	NULL	NULL	NULL
□	✎	✗	65	mandi wajib	taking a bath	ghusl	1)time for obligatory bath	NULL	NULL	NULL	NULL
□	✎	✗	66	sembahyang	prayer	solat	1)conditions of prayer	2)general principles of prayer	3)types of prayer	NULL	NULL
□	✎	✗	84	ziarah	minor pilgrimage	umrah	1)ihram	2)tawaf	3)saei	NULL	NULL

Figure 14. Relational database in I-ES™

Conclusion

The goal of the Semantic Web initiative is to annotate large amounts of information resources with knowledge rich metadata. Based on past research, building ontologies for large domain such as agriculture, medicine, occupation, education or arts by fully dependence on domain expert is a costly affair and time consuming. However, many domain thesauri have been built can be a basis for the construction of an ontology, but thesaurus does not cover for an abstract concepts. For a thesaurus, it should satisfy in number of criteria: it should have a strict subclass, superclass hierarchical structure; it should be base on unique concepts rather than on natural language terms; it should be representable in a format that is compliance with emerging web standards and in ontology construction, additional knowledge should be added to the basic hierarchical structure of concepts derived from the thesaurus. Based on past research about thesaurus in ontology development and various existing methodology for building ontology, we can see the symbiosis of thesaurus and domain expert in ontology development. The outcome of this research is a framework for building multilingual domain ontology for Islamic Portal which applies a new approach in creating ontology. The next phase in this research would be to implement the ontology which had been developed to improve retrieval in our Islamic Extraction and Retrieval System and ultimately to assist users with search topic clarification. Thus, the symbiosis of thesaurus, domain expert and reference sources aims to help users to expand their query using the related terms given.

Acknowledgement

The authors would like to thank the domain expert in this research, Prof Dr. Jawiah Dakir from the Faculty of Islamic Studies in Universiti Kebangsaan Malaysia, who assisted in adding more attributes and relationships to the ontology constructed.

References

[1] Q. Zeng and T. Tse, “Exploration and Developing Consumer Health Vocabularies,” Journal of the American Medical Informatics Association, vol. 1, pp. 24-29, 2006.

[2] Q. Zeng, S. Goryachen, T. Tse, Alla Keselman and Aziz Boxwala. “Estimating Consumer Familiarity with Health Terminology: A Context-based Approach,” Journal of the American Medical Informatics Association, vol. 3, pp. 349-356, 2008.

[3] C. A. Harper, & B. B. Tillett, “Library of Congress controlled vocabularies and their application to the Semantic Web,” Cataloging & Classification Quarterly, vol. 43(3), pp. 47-68, 2007.

[4] T. Berners-Lee, F. Mark, “Weaving the Web,” San Francisco: HarperOne, 2007.

[5] G., Antoniou, and F.V.Harmelen, “A Semantic Web Primer,” Massachusetts: The MIT Press, 2004.

[6] R. Neches, R.E. Fikes, T. Finin, TR Gruber, T. Senator and W.R. Swartout. Enabling Technology for knowledge sharing. AI Magazine 12(3): 36–56, 1991.

[7] T.R. Gruber .A Translation Approach to Portable Ontology Specification. Knowledge Acquisition, pp 199-220, 1993.

[8] R. Struder, R. Benjamins and D. Fensel. Knowledge Engineering: Principles and Methods. Data and Knowledge Engineering 25: 161-197, 1998.

[9] B. Lauser, T. Wildemann, A. Poulus, F. Fisseha, J. Keizer and S. Katz. A Comprehensive Framework for Building Multilingual Domain Ontologies: Creating a Prototype Biosecurity Ontology. Proc. Int. Conf. On Dublin Core and Metadata for e-Communities, pp 113-123, 2002

[10] Oxford Advanced Learners Dictionary. http://www.oxfordadvancedlearnersdictionary.com/, 2010

[11] Xing Xin, Li Ru, Liu KaiYing, “Building Ontology Base on Thesaurus,” 2nd International Conference on Biomedical Engineering and Informatics, pp. 1-4, 2009.

[12] Chang Chun and Lu Wenlin, “From agricultural thesaurus to ontology,” 5th AOS Workshop. pp. 1-4, 2001.

[13] B. J. Wielinga, A. Th. Schreiber, J.Wielemaker, J.A.C. Sandberg, “From Thesaurus to Ontology,” K_CAP’01, pp. 194-201, 2001.

[14] B. Lauser, M. Sini, Anita Liang, J. Keizer, S. Katz, “From AGROVOC to the Agricultural Ontology Service/ Concept Server. An OWL model for creating Ontologies in the agricultural domain,” DCMI '06 Proceedings of the 2006 International Conference on Dublin Core and Metadata Applications: Metadata for Knowledge and Learning, pp. 68-77, 2006.

[15] B. Swartout, P. Ramesh, K. Knight, T. Russ. Toward Distributed Use of Large-Scale Ontologies. AAAI Symposium on Ontological Engineering, Stanford, pp 605-621, 1997.

[16] Evfi Mahdiyah, “Sistem Pengekstrakan Maklumat Islam,” Tesis Sarjana Sains Maklumat, Universiti Kebangsaan Malaysia, Selangor, Malaysia, 2009.

[17] Kamarul Ariffin Ithnan, “Kamus Al-Irshad Empat Bahasa,” Darul Nu’man, Kuala Lumpur, Malaysia, 2007.

Juhana Salim is a professor at the Faculty of Information Science and Technology**,** PhD (Infomration Science - Universiti Kebangsaan Malaysia), Masters of Science in Librarianship (Western Michigan University, Kalamazoo). Bachelor of Arts (Western Michigan University, Kalamazoo).

Siti Farhana Mohamad Hashim is a postgraduate student, B.Sc Information Science (UKM), Member of Persatuan Persatuan Capaian Maklumat dan Pengurusan Pengetahuan(PECAMP).

Shahrul Azman Mohamad Noah is a professor at the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, and currently heads the Knowledge Technology research group. He received his MSc and PhD in Information Studies from the University of Sheffiled, UK in 1994 and 1998 respectively. His research interests include semantic technology, information retrieval and ontology. Prof. Noah is a member of the IEEE Computer Society.