The Development of Ontology from Multiple Databases

Shahreen Kasim, Nurul Aswa Omar, Mohd Farhan Md Fudzee, Azizul Azhar Ramli, Mohamad Aizi Salamat, Hairulnizam Mahdin

Software and Multimedia Centre, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia

Corresponding author: shahreen@uthm.edu.my

Abstract. The area of halal industry is the fastest growing global business across the world. The halal food industry is thus crucial for Muslims all over the world as it serves to ensure them that the food items they consume daily are syariah compliant. Currently, ontology has been widely used in computer sciences area such as web on the heterogeneous information processing, semantic web, and information retrieval. However, ontology has still not been used widely in the halal industry. Today, Muslim community still have problem to verify halal status for products in the market especially foods consisting of E number. This research tried to solve problem in validating the halal status from various halal sources. There are various chemical ontology from multilple databases found to help this ontology development. The E numbers in this chemical ontology are codes for chemicals that can be used as food additives. With this E numbers ontology, Muslim community could identify and verify the halal status effectively for halal products in the market.

1. Introduction

In general, ontology is a representation of knowledge. Ontology is an explicit formal specification of the terms in domain and relation among them [1], [2]. Ontology in computer sciences and information sciences, defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties), and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application [3]. Nowadays, many areas in computer sciences use ontology such as knowledge engineering, software reuse, digital libraries, web on the heterogeneous information processing, information retrieval and semantic web [4]. In this study ontology will use as a data in computation similarity measure. We choose the ontologies because ontologies offer a structured and unambiguous representation of knowledge in the form of conceptualizations interconnected by means of semantic pointers [5].

However, the ontology in semantic web area has still not been used widely particularly in the halal industry. The halal industry has been the fastest growing global business in Malaysia. Today, Muslim community still have problem to verify halal status for halal products in the market especially in foods consisting of E number [6]. Most of the information available on the internet simply displays a list of companies and list of products with the identification of their halal status. Nevertheless, there are also some that display status of materials used in the food on the internet but the status is always in conflict or not the same as other websites. Besides, halal certification logo is lack of security which makes it easier for this logo to be copied [7].

Due to this problem, varieties of food products that have halal status have been doubted. Therefore, to solve this problem, there is the need to establish a method by which the user can check the status of food and know the food source especially those oriented from E number. In this paper, ontology will apply at E numbers in which E numbers are codes for chemicals that can be used as food additives. As told, E number is chemical nature and to establish a database on the E number consists of many different types of resources. Due to the variety of sources used, ontology mapping technique was used to combine all types of databases. Section research approach will describe the data collection phase and in the following section results and discussion will describe early results in the developing of the ontology.

2. Research Approach

In this study, we used chemical data where E number represents chemical entities of the food. Section A, B and C below will describe on this data and the process in detail.

A. E number as a Chemical Entities

E number is no longer a foreign ingredient in the food world. E number is used as food additives for flavor enhancers, stabilizers food, colours, preservatives, antioxidants and antibiotics. E number is made up of chemical substances which are permitted to be used in food. Nevertheless, the halal status is always at doubt, particularly for Muslim consumers. Various databases and chemical ontology have been created but its function does not describe the E number. Various chemical ontology and databases found to help this ontology construction, among which are Chemical Entities of Biological Interest (ChEBI), the National Center for Biomedical ontology (NCBO), Comparative Toxicogenomics Database (CTD), PubChem Bioassay Database (PubChem) and Human Metabolome Database (HMDB). Due to different sources of information to build E number, ontology was used to complete this study. Although various chemical databases have been found to help establish this ontology, not all the data will be used to build the E number ontology. Next, *Section B* will describe the process of collecting the E number.

B. Collection of E number

There are four websites which we collected the E number; http://www.guidedways.com [8], http://www.muslimtents.com [9], http://www.muslimconsumergroup.com/enumbers_list.html [10] and http://special.worldofislam.info/Food/numbers.html [10]. These E numbers are accompanied by their halal status.

Currently, on these websites, there are 516 E numbers which represents the E number from E100 to E1599. The E number was classified into 9 groups which are E100-E199 (colours), E200-E299 (preservatives), E300-E399 (antioxidants & acidity regulators), E400-E499 (thickeners, stabilizers & emulsifiers), E500-E599 (pH regulators & anti-caking agents), E600-E699 (flavour enhancers), E700-E799 (antibiotics), E900-E999 (miscellaneous) and E1100-E1599 (additional chemicals). The purpose of taking E number from different websites is due to the differences of halal status shown in each website. For instance, E101 representing Riboflavin (Vitamin B2) has different status between guidedways.com and muslimtents.com website. Section C below will describe the methods to filter these data.

C. Filtering data that related with E number

There were 2 databases or ontology used in this study, but all the data requires data filtering in advance to avoid the new created database full of unwanted data.

(a) ChEBI - Chemical Entities of Biological Interest

ChEBI is a freely available dictionary of 'small molecular entities' and ChEBI incorporate an ontological classification, whereby the relationships between compounds, groups or classes of compounds and their parents, children and or siblings are specified [12]. There are six types of formats which can be selected for downloading ChEBI data which include the SDF file, OWL file, OBO file, Flat file/tab delimited, Oracle binary table dumps and Generic SQL (Structured Query Language) table dumps. In this study, the data was downloaded in the form of flat file which is easier and various spreadsheets tools available to import this into a relational database. The files were stored in the same structure as the relational database. As shown in Fig. 1 after downloading all the ChEBI data, the data was then separated according to necessary data and data that should not be used. The following was the method used to filter ChEBI data. Once the ChEBI data was downloaded using the flat file format, data inspection was conducted to identify the contents of each table and important keywords in the database. Here we notice that chebi_id and compound_id are the identifications which were used to ensure that they have relationships between tables. Each chemical has a chemical structure of its own id for example Riboflavin has CHEBI: 17015 and has compound_id 8843. Therefore, in order to identify any information in this database, we need to know chebi_id and compound_id for each material. The identification of chebi_id can be done using the search function in www.ebi.ac.uk by typing the name of ingredients. After the chebi_id is identified, it could be a reference to a table using the keyword chebi_id and separate the information into a new table with information about the E number only. Similarly, compound_id uses the same process as it is used in chebi_id but the difference is, it is used in a table that uses compound_id.

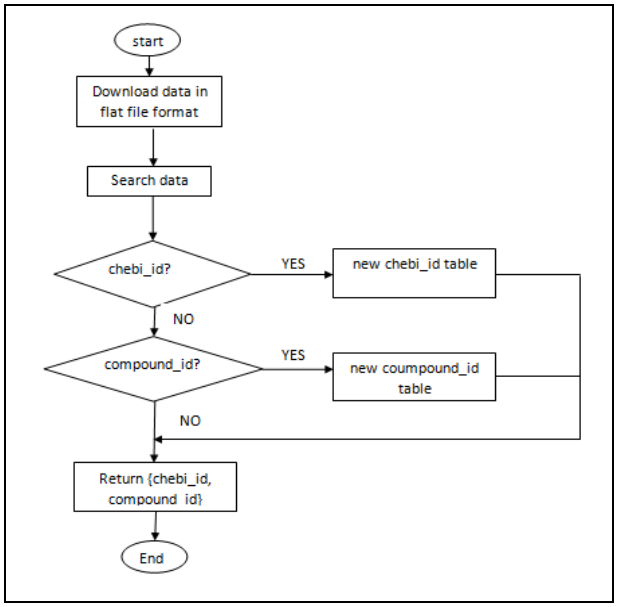

Fig. 1 Flowchart of Filtering ChEBI Data.

(b) PubChem - PubChem Compound Database

The PubChem Compound Database provides information on the biological activities of small molecules. PubChem is organized as three linked databases; PubChem Substance, PubChem Compound, and PubChem BioAssay. PubChem's chemical structure records links to other Entrez databases providing information on biological properties; PubMed scientific literature and NCBI's protein 3D structure resource [13]. A part of the E number is from PubChem database. Data from these PubChem can be downloaded for adding information to build the E number ontology and these data can be downloaded in a variety of formats; Abstract Syntax Notation One (ASN.1), Extensible Markup Language (XML) and Standard Delay Format (SDF). The PubChem offers users to download

data individually. This is easier to filter data by not having to download all the data in PubChem database but only download the required data. Besides, PubChem also provides a chemical structure search which user can use the names of the PubChem chemical or PubChem chemical id to find the desired information. Each chemical data in PubChem have CID code, for example; CID 6093240 representing Pigment Rubine (E180). After searching the required chemical data, information about that chemical will appear and it can be downloaded in a variety of formats. In this study, we downloaded these data in XML format. Fig. 2 below is a flow process to download the PubChem data.

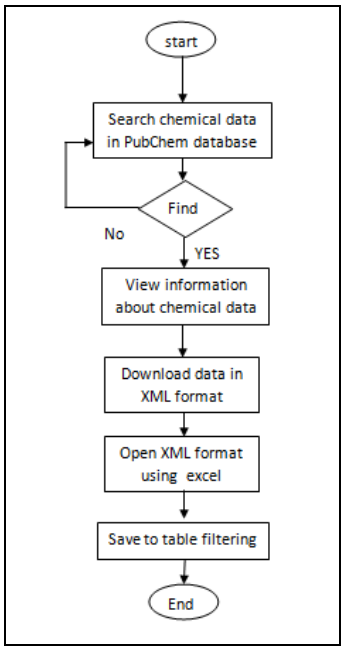

Fig. 2 Flow process to download PubChem data.

3. Results and Discussion

To develop E number ontology, the selection of ontology editors is important. Among ontology editor that can be used to build ontology are Protégé, OILEd, KAON (**Ka**rlsruhe **on**tology), TopBraid and Apollo [14]. In this study, we chose to use Protégé as ontology editor. Protégé is a free open source ontology editor and knowledge-base framework. It is based on Java, is extensible and provides a plug and play environment that makes it a flexible base for rapid prototyping. The protégé platform supports modeling ontologies via a web or a desktop client. Protégé editor can be developed in a variety of formats including OWL, RDF(S), and XML Schema [15].Ontology development can be started after the data is collected. Data are classified into several parts, according to the E number ingredients that have been set. The example visualization classes E number ingredients in the form of the Protégé editor as shown in Fig. 7. As has been described earlier, the E number was classified into 9 groups. Fig. 4(a) shows that each group represents a specific function such as E100-E199 represents the colours. Each representation still has other subclasses. Refer to Fig. 4(b), E100-E199 is classified into 7 groups which are E100-E109 (yellow), E110-E119 (orange), E120-E129 (red), E130-E139 (blue & violet), E140-E149 (green), E150-E159 (brown &black), E160-199 (gold & others). Each group of color has individual. Individual is the ground level components of ontology or specifying the actual value of specific instances of the class [16]. Fig. 4(c) shows colour group for E100-E109 (yellow) in which have 9 individuals; E100 (Curcumin), E101 (Riboflavin), E101a (Riboflavin-5'-Phosphate), E102 (Tartrazine), E103 (Alkannin), E104 (Quinoline Yellow WS), E105 (Fast Yellow AB), E106 (Riboflavin-5-Sodium Phosphate), E107 (Yellow 2G).

In addition, ontology has annotation. Annotation for ontology is a vocabulary for performing several types of annotation such as comment, entities annotation, textual annotation, notes and example. Annotation in ontology such images and audio [17] also in biomedical research [18] have an impact to each of research ground. Annotation is the process of assigning E number terms and their synonyms. Synonyms are words with the same or similar meaning. For E number ontology, we use code like E101 but at the same time this E number is named as Curcumin. It also has synonyms such as "kacha haldi", "natural yellow 3", "turmeric" and "turmeric yellow". In the protégé editor, synonyms for every E number are placed in the (same individual as) partition. Examples of annotation for this protégé are shown in Fig. 8 below.

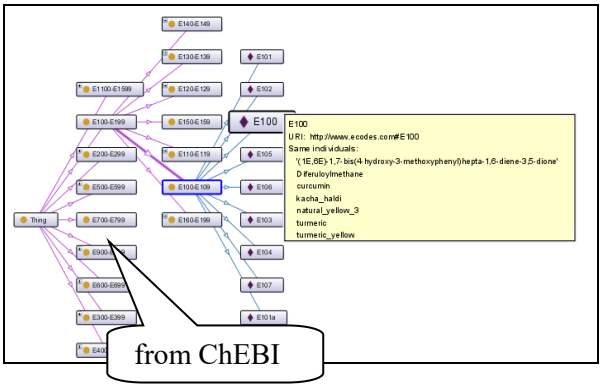

Fig. 3 The example of annotation visualization for instances

4. Conclusion

In conclusion, ontology development based on the domain E number is an attempt to help the Muslim community verify halal status information accurately. The development of E number ontology is based on the information of halal status from various websites. Various chemical ontology and databases are also involved in the construction of this ontology E number. This is because E numbers are codes for chemicals that can be used as food additives. With this E number ontology, it would be much easier for the Muslim community to identify and verify the halal status for halal products in the market.

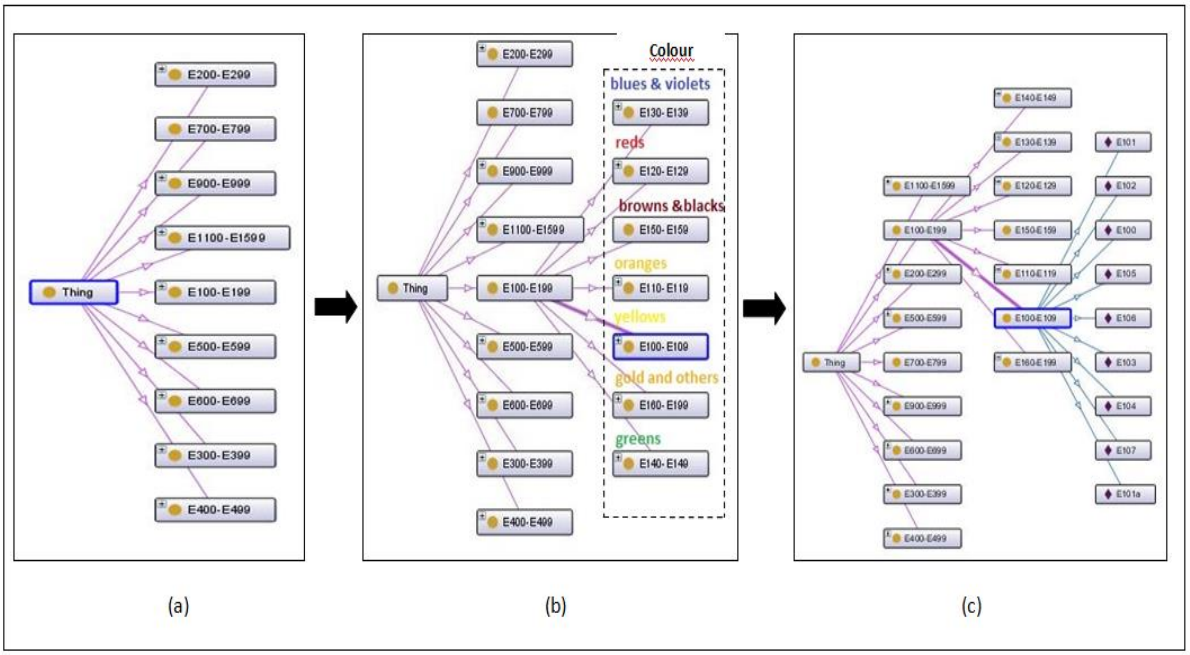

Fig. 4 The example visualization for E number ontology of (a) 9 classified E number (b) colour group E number (c) individual for colour group

References

[1] N. F. Noy and D. L. McGuinness. (March 2001). Ontology Development 101: A Guide to Creating Your First Ontology. *Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford MedicalInformatics TechnicalReport.*[Online]. Available:http://protege.stanford.edu/publications/ontology_development/ontology101noy-mcguinness.html.

[2] Gruber.T.R, "A Translation Approach to Portable Ontology Specification," *Knowledge Acquisition*, vol. 5,pp. 199-220, 1993.

[3] Gruber.T.R, "Ontology," in *Encyclopedia of Database Systems*, L. Liu, and M. T. Ozsu,1st ed, Springer-Verlag, 2009.

[4] Q. Wei, "Development and Application of Knowledge Engineering Based on Ontology," in *Third International Conference on Knowledge Discovery and Data Mining*, Phuket Thailand, 2010, pp. 518-521.

[5] D. Sánchez, M. Batet, D. Isern, and A. Valls, "Ontology-based semantic similarity: A new feature-based approach," *Expert Syst. Appl.*, vol. 39, no. 9, pp. 7718–7728, Jul. 2012.

[6] M. Kassim, Yahaya, C.K.H.C.K., Zaharuddin, M.H.M, Z.A. Bakar, "A prototype of halal product recognition system" in *International Conference on Computer & Information Science (ICCIS)*, 2012, pp.990-994.

[7] S. Zailani, Z. Arrifin, N. A. Wahid, R.Othman, and Y. Fernando, "Halal traceability and Halal tracking systems in strengthening halal food supply chain for food industry in Malaysia (A Review)", *J. Food Technology*, vol. 8, pp. 74-81, 2010.

[8] Guidedways Technology. Available: http://www.guidedways.com.

[9] Muslim Hosting Community. Available: http://www.muslimtents.com.

[10] Muslim Consumer Group. Available: http://www.muslimconsumergroup.com/enumbers_list.html.

[11] Food Ingredients Number. Available: http://special.worldofislam.info/Food/numbers.html.

[12] Chemical entities of Biological Interest (ChEBI). Available: http://www.ebi.ac.uk/chebi.
[13] PubChem Compound. Available: http://www.ncbi.nlm.nih.gov.
[14] S. Saad, N. Salim, H. Zainal, and Z. Muda, "A process for building domain ontology: an experience in developing Solat ontology" in *International Conference on Electrical Engineering and Informatics*, Bandung Indonesia, 2011, pp. 1-5.
[15] Stanford medical informatics. Available: http://protege.stanford.edu.
[16] N. N. Mohammad, "Product Structure Ontology To Support Semantic Search In Manufacturing Requirements Management," M.S.thesis, Faculty Computer Science and Information System., UTM., Johor, Malaysia, 2010.
[17] P. Ciccarese, M. Ocana, LJG. Castro, S. Das, and T. Clark. (May 2011). An Open Annotation Ontology for Science on Web 3.0. *J. Biomed Semantics*.[Online]. vol. 2. (Suppl 2):S4. Available:http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3102893/.
[18] S. Kasim, S. Deris, R.M. Othman, "A new computational framework for gene expression clustering", Lecture Notes in Computer Science 6440 LNAI (Part 1), pp. 603-610, 2010.