Available online at www.isas.org.in/jisas
Journal of the Indian Society of Agricultural Statistics 74(1) 2020 77–84

Ontology Learning Algorithm for Development of Ontologies from Taxonomic Text and USDA Soil Taxonomy Ontology

Chandan Kumar Deb[1], Sudeep Marwaha[1] and R.N. Pandey[2]

[1]ICAR-Indian Agricultural Statistics Research Institute, New Delhi
[2]ICAR-Indian Agricultural Research Institute, New Delhi

Received 20 May 2017; Revised 20 June 2019; Accepted 29 June 2019

SUMMARY

Web based software use ontologies to keep the web one step ahead from the conventional web. Ontology based software architecture makes the platform suitable to work together for human as well as machine. By means of Ontology the unstructured knowledge is easily converted into the structured one. Soil Taxonomy Ontology developed for USDA soil taxonomy by Das (2010) and Das *et al.* (2012) for soil orders available in India is available to only sub group level. In this work developed knowledgebase is used to develop web based software with N-tier architecture and the ontology has been extended up to family and series level. It also covers all the twelve order of USDA Soil Taxonomy and provides the state wise series description of the soil. Search module of the software provides the exclusive search of the soil taxonomy and the edit module provide the facility to add, delete and edit facility of the ontology information. Additionally we have developed an algorithm for automated ontology learning from the taxonomic texts with a case study of soil taxonomy.

Keywords: Semantic web, Ontology, N-tier architecture, Ontology learning.

1. INTRODUCTION

Ontology is the heart of the semantic web and also acts in synergy with software agent and semantic web [Berners-Lee *et al.,* 2001]. Ontology helps in many ways to better describe the information of the web and their internal relationships. The main building block of the ontology is the statements. Statement defines concepts, relationships and imposes constraints to the concepts. Conceptually, this is very similar to the database schema or an object oriented class diagram. Across applications, communication can easily be achieved with the help of inbuilt ontology in the application. Although the scratch building of ontology is a very difficult task but once it is built; it can easily be extended and reused extensively.

The classic examples of developed ontologies are Gene Ontology [Gene Ontology Consortium (2000)] and Plant Ontology [Plant Ontology Consortium (2002)]. AmiGO functions as Browsing and searching tool for retrieving the data in Gene Ontology. Taxonomy has a great correspondence with the ontology. A methodology for conversion of taxonomies to ontology was proposed by Bedi and Marwaha (2004). The proposed methodology is tested and implemented for a pilot soil ontology using the IEEE standard Web Ontology Language (OWL) and protégé 2.1 OWL plug-in. OWL is the W3C recommendation for describing Ontology [Dean *et al.*, 2003]. Ontology-based intelligent retrieval system for soil knowledge [Ming *et al.*, 2009] is a system which searches documents related to soils by using soil domain ontology. This system retrieves information like "Relationship between Laterite soil and air pollution". Ontology Based Expert System [Bedi and Marwaha, 2005 and Marwaha, 2008] provides facilities to diagnose diseases and identify insects.

Soil Taxonomy Ontology has been built [Das (2010) and Das *et al.* (2012)] for USDA soil Taxonomy

Corresponding author: Chandan Kumar Deb
E-mail address: chandan.iasri@gmail.com

[USDA, NRCS (2010)] based on soil morphology that can be observed and measured in the field. In his work, a detailed study of the USDA soil taxonomy can be done by a given query interface, but his work didn't cover the taxonomy in totality. It only covered the seven orders among the twelve order of the USDA Soil Taxonomy. It also didn't cover the hierarchy up to the family and series level of the soil taxonomy. The developed Soil Taxonomy Ontology contains the information up to the sub group levels i.e. the family and series of the taxonomic hierarchy of the developed seven orders of soil ontology.

Under the present research work the Soil Taxonomy Ontology has been extended in two ways. Firstly, the existing seven orders are extended to the series level and secondly, the remaining five orders of the soil are also added to the taxonomy up to the subgroup level. All the user privileges remains intact in this research work and additionally two module namely state wise series description module and information edit module has been incorporated to the system. To overcome the shortcomings of the manual ontology building we have developed an algorithm for automated Ontology Learning with a case study of Soil Ontology.

2. MATERIAL AND METHODS

The Soil Taxonomy Ontology is a web based software with N-tier architecture. Entire application is developed based on two main tasks. First task is development of the user interface or the front end and the second task is the development of the back end which consists of database and knowledge base. These two ends of the software are bridged up by different java API (application programming interface).

2.1 Architecture of the Software

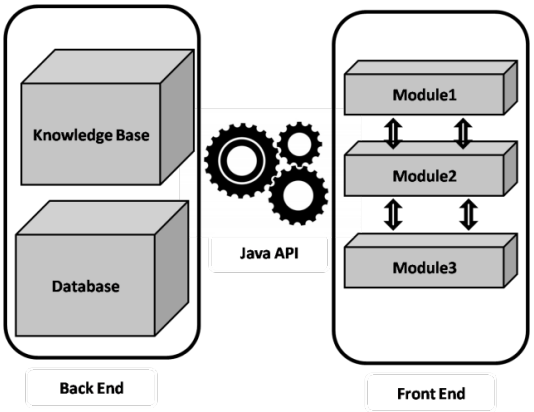

Fig. 1. Architecture of the Software

2.2 Schematic Process Flow of Development of the Software

Total process of the software development is described by the schematic diagram given in the Fig. 2. Here we have used the USDA soil taxonomy as an information source for the development of the ontology. This is fed into the process flow of the development. In this process flow the identification of ontological building block i.e. class, instance, properties etc has been identified. After that the population of the ontology is done and the user interface is developed.

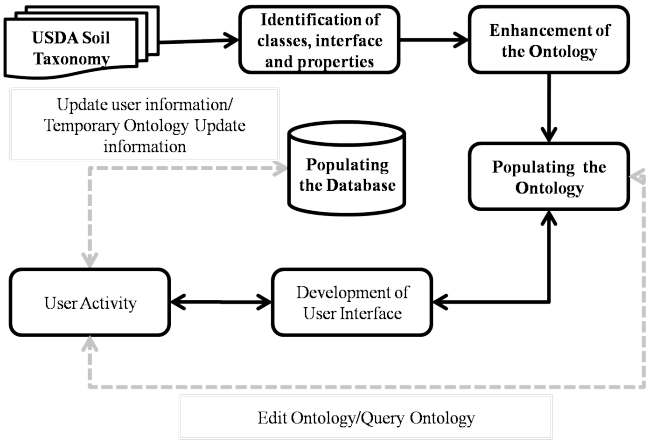

Fig. 2. Process flow of the development of soil taxonomy ontology

2.3 Identification of ontology Class, Individuals and Properties

2.3.1 Class Identification

Identification of class is the most important task of any ontology building. The names of the Orders, the formative element in the Order name, used as an identifier at lower categorical levels, derivation or source of the formative element and the mnemonicon for each Order. Each Suborder name consists of two syllables. The first is suggestive of the class (*i.e.* Suborder), and the second name of the Order is as reflected by the formative element (*e.g.* oll for Mollisol). Likewise, the names of Great Group are coined by prefixing one additional prefix (formative element) to the appropriate Suborder name. Subgroup names consist of the name of the appropriate Great group modified by one more adjective. Families, in this category, the intent has been to group the soils within a subgroup having similar physical and chemical properties that affect their responses to management and manipulation for use. For Series, the local name is used for the class name.

Examples:

Suppose the series ***Zaheerabad*** has the classification ***Clayey skeletal kaolinitic isohyperthermic Kandic Paleustalfs***

alfs is used for order Alfisols, **ustalfs** is used for the suborder ustalfs, **Paleustalfs** is used for the Great group, **Kandic Paleustalfs** is used for the sub group. So the family ***Clayey skeletal kaolinitic isohyperthermic Kandic Paleustalfs*** is the sub class of **Kandic Paleustalfs** class and the series ***Zaheerabad*** is the sub class of this family. In this manner a particular family or series are added to this ontology [Fig. 3.2].

In Soil Ontology the hierarchy is:

Order (Alfisols) Sub order (Ustalfs) Great group (Paleustalfs) Sub group (*e.g. Kandic Paleust*alfs) Family (*e.g. Clayey skeletal kaolinitic isohyperthermic Kandic Paleustalfs*) Series (*e.g. Zaheerabad*)

2.3.2 Individual Identification

According to the object oriented programming concepts, Individuals are the physical existence of the class. Like class identification for the ontology development, individual identification of every class is very important. For taxonomic class of the soil ontology we have used the same name as the class name. In case of property class like Basic_Property_Alfisols, we have created the property as individuals for order Alfisols.

In the same manner each and every classes are populated with different individuals.

2.3.3 Property Identification

Constructs and populates of Ontology are very much dependent on the property identification. Property is a very important component, because the two related class can only be joined by the property. On the basis of the related class, the property has been identified.

2.4 Tools and Technologies for Soil Ontology

The software development process is dependent in many ways on technology which is used for the development of the software. This software is developed using Java technology and total development process has been done on the Integrated Development Environment (IDE) Netbeans 6.9. All the development has been done through JDK 1.7 and additionally some API has been used to deal with the ontology.

In JEE 2.0 the web interface has been developed. The front end was developed by HTML, CSS and JavaScript. Apart from the core java class some of the programming is done through the JSP pages.

In the back end of the software has been divided into two parts. First part is the Database and Second part is the Knowledgebase. Behind the database development we used Microsoft SQL Server 2008

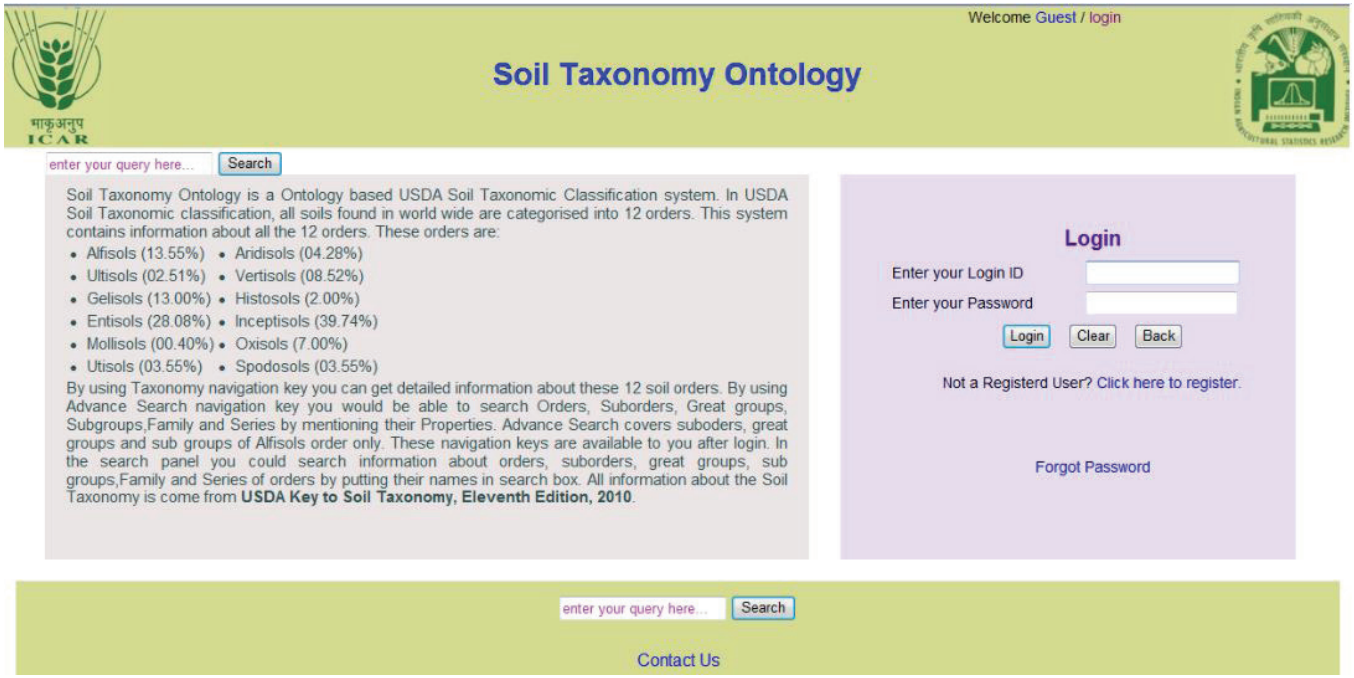

Fig. 3. Home Page of the Soil Taxonomy Ontology Software

and for the knowledgebase we used Protégé 3.4.6. Interaction of the front end with the back end has been done through the different java API. To connect to the database we have used the conventional JDBC Bridge but the connection as well as the interaction with the knowledgebase has been done through main java API namely; JENA, OWLSyntax, and ProtegeOWL. The second bridging process is literally known as the Semantic Web Framework Layer.

3. RESULTS AND DISCUSSION

3.1 Results of populated enhanced ontology

After the identification of class, interface and the properties of the domain i.e. USDA soil taxonomy, it is time to populate the Ontology with the real information. Some of the results are described below.

In this research work, we have worked on manual ontology building and also suggested an approach to make the process of ontology building automated. Although this work is done for 5 Orders, 20 Suborders, approx. 138 Great Group and approx. 793 Sub Groups.

Population of the ontology and proper tuning of soil taxonomy is one of the main objectives of this research work. The knowledge base of the Soil Taxonomy Ontology has been enhanced in many aspects. First, we have extended the existing seven orders up to the series level. Second, the population of the ontology up to the series level is done. Some of results of the populated ontology have been listed below.

i) The class has been extended for the world wide soil taxonomy.

ii) Property classes and their subclasses in soil ontology

iii) Property classes of family and series

iv) Classes for holding the geographical description of Series

v) Restriction applied to *has Basic Property* property

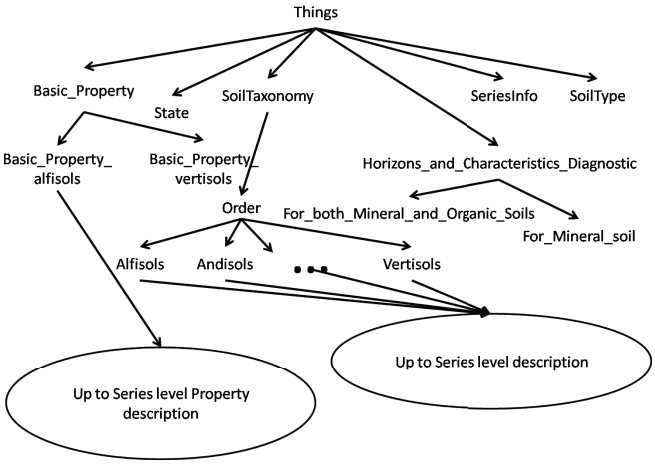

Fig. 3.1. Soil Ontology extended up to soil series

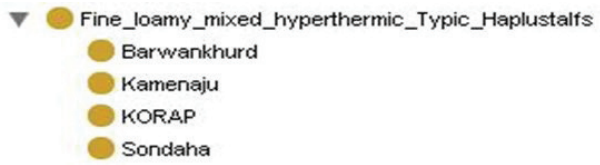

Fig. 3.2. Soil Ontology extended up to soil series in protégé

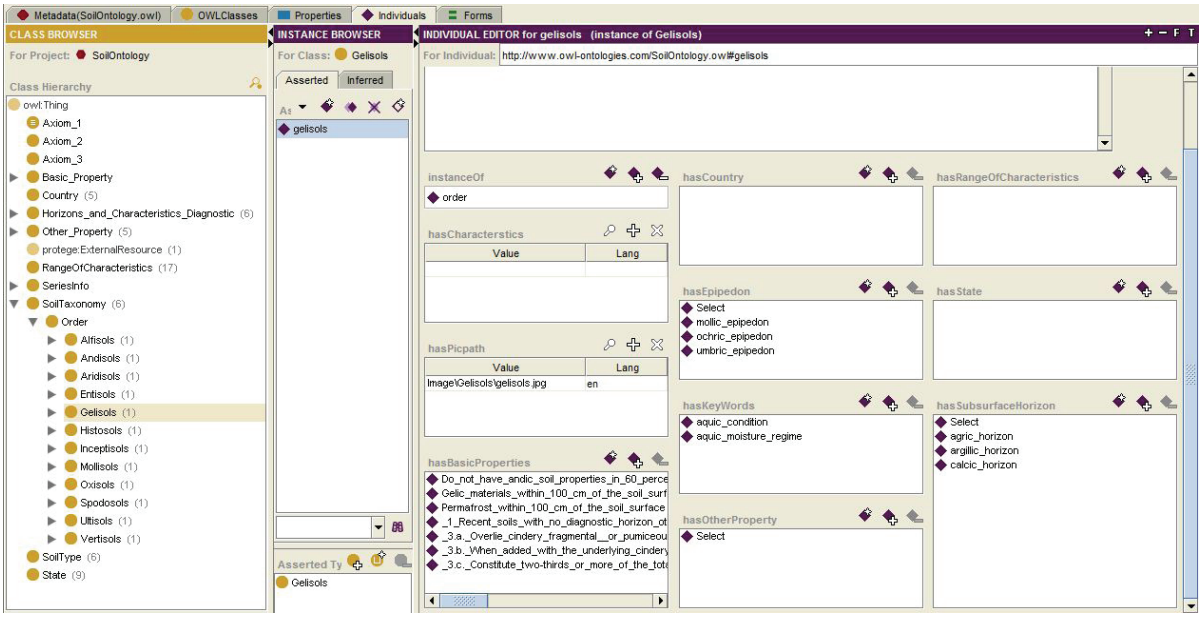

Fig. 3.3. Gelisols class with its individual gelisols with its properties and their corresponding values in Protégé OWL Plug-in

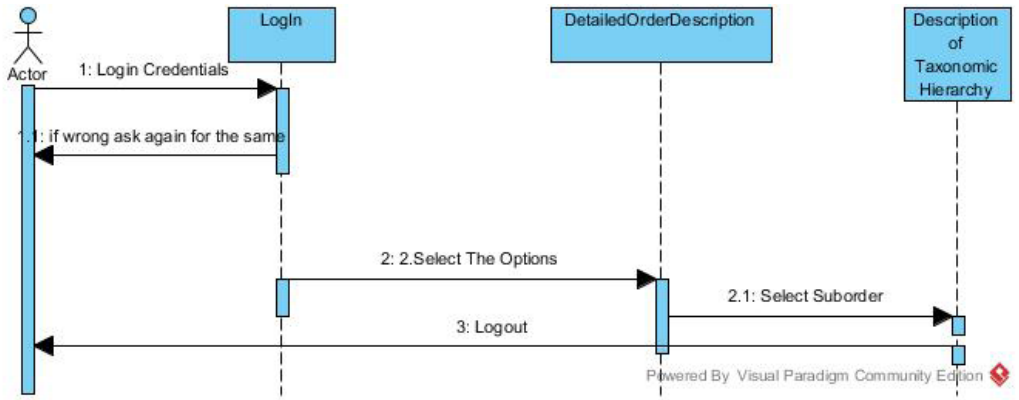

Fig. 3.1. Sequence diagram of Study of USDA Soil Taxonomy

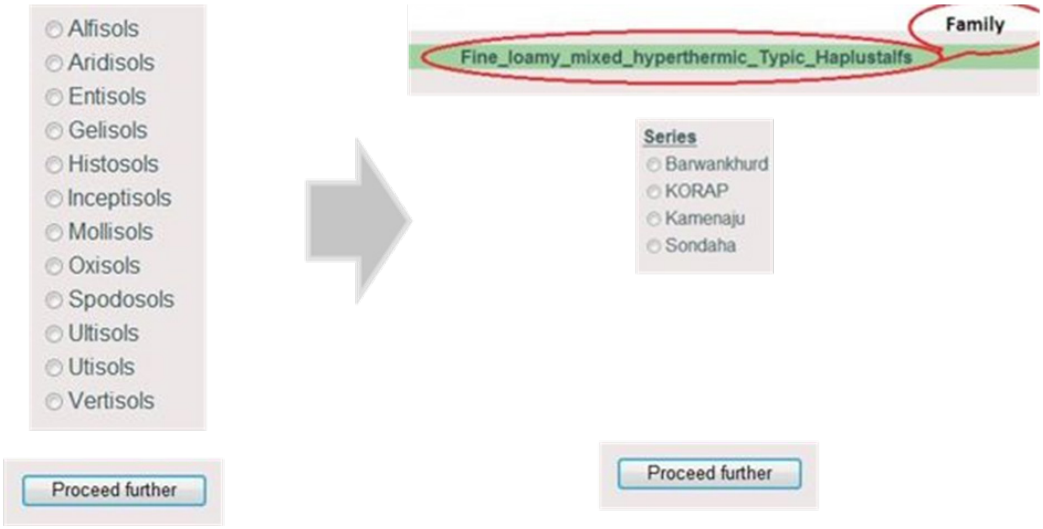

Fig. 3.4 Output of the study of Soil Taxonomy

3.2 Software facilitated the study of USDA soil taxonomy

The software provides the facility to a detailed study of the USDA soil taxonomy. It starts from the order and in a step by step manner it gives a detailed information of the selected category. In every step, it gives a details of the selected class and enlists the associated subclass. Fig 3.4 is the sequence diagram describing the steps of a detailed study of the soil taxonomy.

3.3 State wise series description of Indian soil series

For proper agricultural practice, the local information of the soil is very important. In the USDA Soil Taxonomy the series is the lowest hierarchy which is strongly coupled with the local soil description. One of the principle focus of this research work is to provide the series description of the soil. The software provides series information in two ways- firstly the taxonomic description which is available in "*Taxonomy*" tab of the software and secondly the state wise series description in "*Series*" tab of the software.

3.4 Classify newly found soil into proper hierarchy

Another powerful module of Soil Taxonomy Ontology software is totally deedicated to the searching of an existing hirarchy of the soil taxonomy. Fig. 3.5 depicted the activity behind the search of the system. The search can be done through the simple search module or the advanced search module. The simple search is done by using the key words and the taxonomic term. The advanced search module is a relatively sophisticated one than the simple search module. The advanced search is done on the basis of specific information for any hierarchy (Order, suborder etc.) of the soil taxonomy. Both the search result produce a proper hierarchical format for proper understanding of the taxonomy.

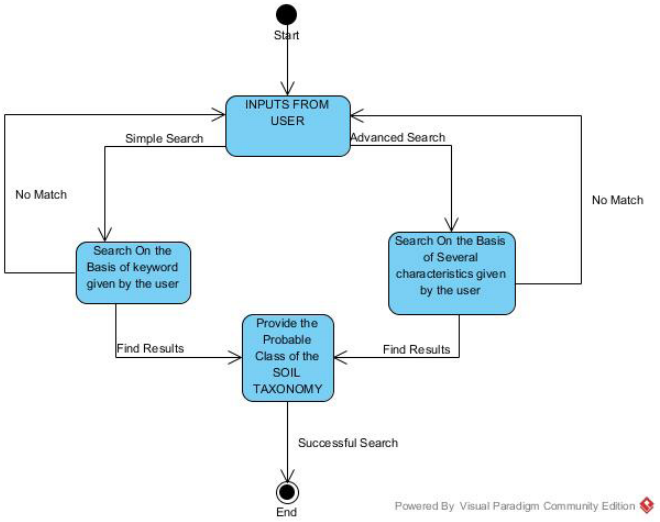

Fig. 3.5: Activity Diagram of Search Module

3.5 Software administrative functionality and ontology edit module

Web based software must have some administrative facility to combat many problems which appears after a long term use of the software. Cleaning of the garbage data, tuning of the software's necessary data also comes under the administrative functionality. This software has sign up facility to the user. Soil Taxonomy Ontology has three types of users. The administrator, domain experts and the general user are the users kind.

Among the three types, general user can only retrieve the information, use the simple and advanced search module and have the facility to classify newly found soil.

Like General user all the privileges are also available for the advanced user i.e The Domain Experts and the Administrator. Additionally they are involved in the Edit Ontology Module. The Domain Experts make the changes in the ontology. The change can be done in two ways first the new information which is already present in the ontology or edit the information which is already present in the ontology.

The Administrator is like super user of the system. Administrator can approve or disapprove the changes made by the domain experts.

4. AUTOMATED ONTOLOGY LEARNING ALGORITHM – CASE STUDY SOIL ONTOLOGY

As we have previously mentioned ontology building from the scratch is not only a difficult job but also it is very time consuming. This is very obvious that the taxonomic text is more structured than the plain text. We used this criterion to make the ontology development automated. We propose a methodology for ontology building through Natural Language Processing (NLP). We used the frame work of ontology learning proposed by Deb *et*. al. 2015.

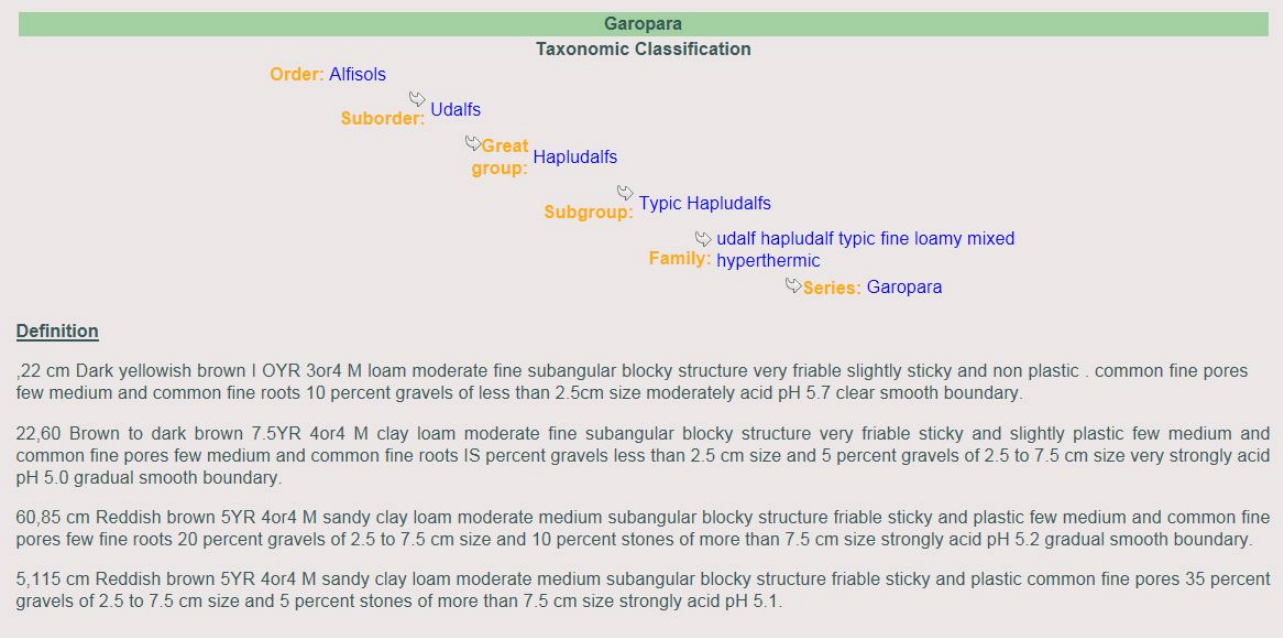

Fig. 3.6: Details study of Garopara Series by search

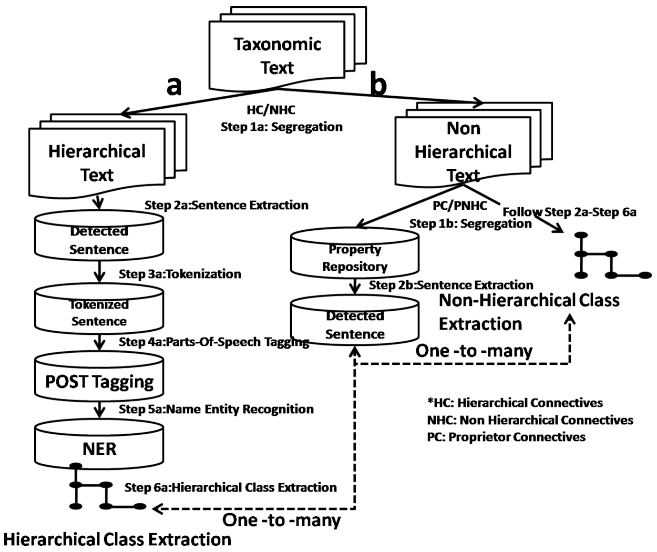

Fig. 4.1 Describes how the process of ontology learning may proceed

Step 1a and Step 1b: Segregation: This is the first step of ontology learning process. In this step total text or the part of the text which is under processing is divided into two parts i.e. the text contains the taxonomy and the text does not contains the taxonomy. The text is segregated on the basis of connectives present in the sentence. Connectives is a set that contains the key words that established connection between the objects available in taxonomic text.

e.g. "Udalfs are the Alfisols." The sentence contains two object for ontology i.e "Udalfs" and "Alfisols". "are the" established the linguistic connection between them. So for this text "are the" is one of the connectives.

Step 2a and Step 2b: Sentence Detection: Sentence detection is the next step of the ontology learning algorithm. Segment the total text into sentences for further task of NLP.

Step 3a: Tokenization: The sentence is further subdivided into words and single symbol called tokenization.

Step 4a: Parts-Of-Speech Tagging: In this task of NLP we find the proper noun for the identification of the taxonomic class of the taxonomic text.

Step 5a: Name Entity Recognition: After Step 4a Name Entity Recognition is very important. For detection of name the corpus can be built on the basis of the corresponding domain.

Step 6a: Hierarchical Class Recognition: In this step we have extracted the is-a relationship or the parent child relationship of the extracted name

First of all we provide the taxonomic text to the natural language processor. On the basis of connectives it will segregate the sentence into two parts hierarchical and non hierarchical text. Here the term connectives means a special set of words which describes the parent-child relationship in the text. In case of taxonomic text it is more prominent and easily usable. The above described separation of the sentence group makes the task of taxonomic relation extraction very easy.

After separation of the hierarchical and non hierarchical text the non hierarchical text further segregated on the basis another connectives i.e. proprietor connectives. This connective is used for the separation of properties for a particular class which is extracted from the hierarchical text or non hierarchical text.

On one hand we have the extracted taxonomic class and subclass and on the other hand we have extracted the properties. Both the extracted things are attached through some semi automated process (Manual relation extraction and association rule extraction) and the Ontological structure will easily be built.

5. CONCLUSION

In this work we have developed web based software with N-tier architecture with Soil Ontology as a knowledgebase. The software provides the facility for a detailed study of the USDA Soil Taxonomy up to the series level. The developed software has the state wise series description facility which can be used for the local soil description that can easily be used in the agricultural practice. It also provides the editing facility of the existing ontology. We have also developed an algorithm for automated ontology learning.

ACKNOWLDGEMENTS

The first author gratefully acknowledges the INSPIRE Fellowship provided by Department of Science and Technology, New Delhi.

REFERENCES

Berners-Lee T., Hendler J. and Lassila O. (2001). The Semantic Web. *Scientific American,* **284(5)**, 34-43.

Bedi, P., Marwaha, S. (2004). Designing Ontologies from Traditional Taxonomies, *In the Proceedings of International Conference on Cognitive Science*, Allahabad, India.

Bedi, P. and Marwaha, S. (2005). Framework for Ontology Based Expert Systems: Disease and Pests Identification in Crops-A case study, in Proceedings of the International Conference of Artificial Intelligence. (ICAI-05): 256-259.

Clark, K. (2008). *SPARQL Protocol for RDF*, *W3C Recommendation*, http://www.w3.org/TR/rdf-sparql-protocol/.

Das, M. (2010). Building and querying soil ontology for agriculture. Unpublished M.Sc. Thesis., IARI, New Delhi.

Das, M., Malhotra, P.K., Marwaha, Sudeep and Pandey, R.N. (2012). Building and Querying Soil Ontology. *J. Indian. Soc. Ag. Stat.* **66(3)**, 459-464.

Dean, M., Connolly, D., Harmelen, F.V., Hendler, J., Horrocks, I., McGuinness, D.L., Patel-Schneider, P.F. and Stein, L.A. (2003). *OWL Web Ontology Language Reference, W3C Working Draft.* http://www.w3.org/TR/2003/WD-owl-ref-20030331.

Deb, C.K., Marwaha, S., Arora, A., & Das, M. (2018). A framework for ontology learning from taxonomic data. In Big Data Analytics (pp. 29-37). Springer, Singapore.

Gene Ontology Consortium. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25-29.

Maji, A.K., Baruah U., Dubey P.N., Butte P.S., Verma T.P, Shilu K. and Angami V. (2004a). Soil Series of Nagaland, NBSS Publ. No.109. NBSS & LUP, Nagpur, pp. 126.

Maji, A.K., Sarkar D., Baruah U., Bhaskar B.P., Sen T.K, Vadivelu S. and Gajbhiye K.S. (2004b). Soil Series of Assam, NBSS Publ. No.101. NBSS & LUP, Nagpur, pp. 229.

Maji, A.K., Sarkar D., Baruah U., Sen T.K and Patil V.P. (2006). Soil Series of Manipur, NBSS Publ. No.134. NBSS & LUP, Nagpur, pp. 53.

Marwaha, S. (2008). Temporal Extension to Ontologies for Semantic Web enabled Systems. Unpublished Ph.D. Thesis, University of Delhi.

Ming, Z., Qingling, Z., Dong, T., Ping, Q., and Xiaoshuan, Z. (2009). Ontology-based intelligent retrieval system for soil knowledge. *WSEAS Trans. Info. Sci. and App. 6, 7*. http://portal.acm.org/citation.cfm?id=1639432.

NBSS & LUP Staff (2006). NBSS Publ. No.136. Soil Series of Kerala, NBSS & LUP, Nagpur, pp. 136.

Plant Ontology Consortium. (2002). Plant Ontology Consortium. http://www.plantontology.org/.

Sahoo, A.K, Sarkar, D. and Gajbhiye, K.S. (2004). Soil Series of Bihar, Publ. No.98 NBSS & LUP, Nagpur, pp. 289.

Sarkar, D., Gajbhiye, K.S, Sahoo A.K. and Sah, K.D. (2005). Soil Series of Orissa, NBSS Publ. No.119 NBSS & LUP, Nagpur, pp. 254.

Sarkar, D., Dubey P.N., Ray S.K., Baruah U., Sehgal, J. and Bhattacharyya, T. (2004). Soil Series of Tripura, NBSS Publ. No.111 NBSS & LUP, Nagpur, pp. 115.

Singh, R.S., Baruah U., Sarkar, D, Butte P.S., Gajbhiye, K.S. (2005). Soil Series of Meghalaya, NBSS Publ. No.121 NBSS & LUP, Nagpur, pp. 86.

Singh, R.S., Singh, S.K., Jain, B.L., Shyampura, R.L. and Gajbhiye, K.S. (2002). Soil Series of Rajasthan, NBSS Publ. No.95 NBSS & LUP, Nagpur, pp. 364.

Smith, M., Welty, Chris and McGuinness, Deborah L. (2004). *OWL Web Ontology* Language Guide (W3C Recommendation). http://www.w3.org/TR/owl-guide-20040210/.

USDA, NRCS. (2010). *Keys to Soil Taxonomy*, Eleventh Edition.