A Novel Semi-Automated Ontology Construction Framework (SOCF) for Psoriasis Detection: Pioneering the Psoriasis Risk Assessment Remedy (PRAR) Database

SIVASANKARI. S.[1], Gracia Jacob SHOMONA[2]

[1] Anna University, SSN College of Engineering, Department of CSE,
Kalavakkam, Chennai-603110, India,
kmiruthu@gmail.com

[2] SSN College of Engineering, Department of CSE, Kalavakkam, Chennai-603110, India,
shomonagj@ssn.edu.in

Abstract: Ontology provides an organizational framework of concepts and a system that depicts hierarchical and associative relationships pertaining to an application domain. The possibility of reuse and data sharing permitted by ontology, along with the formal structure coupled with hierarchies of concepts and their inter-relationships offer the opportunity to draw complex inferences and reasoning. This rationale was the motivation to construct an ontology for Psoriasis Risk Assessment and Remedy (PRAR). This paper targets two issues: (i) Need for a medical database to derive Ontology (ii) Methodology for design of Semi-Automated Ontology Construction framework (SOCF) from pioneered data. Psoriasis is one of the most recurrent skin issues in India and the world at large and hence this paper targeted the need to generate a Psoriasis Remedy Database and automatically infer the relations between the Symptoms, Causes and Treatment through Semi-Automated Ontology Reasoning and Inference. The proposed system incorporated two phases: Formulation of a novel database for Psoriasis Risk Assessment Remedy (PRAR) (ii) Articulation of a novel framework for Psoriasis detection through computational modeling and Ontology Construction. The proposed methodology was tested on 112 samples from the authenticated UCI Machine Learning Repository. The ontology developed using the proposed SOCF mapped the risk factors and remedies for Psoriasis detection with 98.7% accuracy, this being reported for the first time.

Keywords: Psoriasis, OWL, XML, Mapping, Ontology, Reasoning, Inference.

1. Introduction

The field of medicine encounters the emergence of new diseases every day. The diagnosis of these increasing numbers of diseases is becoming difficult. The data collected during the diagnosis and treatment of these diseases is vast, unorganized and hence difficult to manage, thus raising the burden of medical practitioners to monitor and maintain all the information across diverse diseases. This revealed the critical need for an information system that could provide the required solution for the existing scenario by structuring and organizing the existing data. The concept of ontology is believed to address this issue. Ontology is defined as a formal explicit representation of conceptualization of a domain that provides a platform for the sharing and reuse of knowledge across heterogeneous platforms (Jian-xu chen et al . 2010). It contains semantic descriptions using concepts and relationship abstractions in a way that is readable by both humans and machine (Hadzic & Chang 2005). An ontology-based approach represents entities, ideas and events, along with their properties and relations, as a

form of knowledge representation. In the medical world, ontologies represent diseases, its associated symptoms, tests and treatment along with their properties and relationships. Ontological framework can be constructed for diseases of varied nature provided there is availability of data to support the framework. The main objectives of this paper is (i) to formulate a novel database for Psoriasis detection and report (PRAR), its associated symptoms and diseases.(ii) Propose a Semi-automated Ontology Construction Framework (SOCF) that can be applied to the published database (PRAR).

Psoriasis is a recurrent dermatology ailment that is more often revealed as systemic manifestations, especially arthritis manifestations (Gudjonsson 2007). An approximate 2 percent of adults suffer from this ailment, and it appears to be prevalent among men and women irrespective of their age. However, Psoriasis onset is most likely to infect people in the age groups ranging from 15 to 30 years (Gudjonsson 2007).The clinical course is capricious (Langley et al 2005)]. Personalized and tailored therapy can aid in reducing the

associated melancholy and enhance the patient's quality of life. Statistics suggest that close to one-third of patients with Psoriasis have a first-degree relative with the condition. Research suggests inheritance of the ailment from multiple factors (Langley et al 2005, Capon et.al.2004).With the unremitting growth of scientific information sources, the need for integrated access to these sources becomes ever more imperative. The aim of the Psoriasis Risk Assessment Remedy (PRAR) database is to allow access to multiple information sources in the medical domain through web interface.The prevalence, lack of appropriate data on Psoriasis, their risk factors and remedies was the rationale for this research. Hence the authors attempted to derive a novel Psoriasis Risk Assessment Remedy (PRAR) database and propose computational methods to enable semi-automated construction of ontology to map the derived data to Psoriasis detection.

Section 2 narrates the background work. Section 3 presents the materials used and the method for implementing the framework. Section 4 highlights the proposed ontology framework construction for Psoriasis risk Assessment Remedy (PRAR) database. Section 5 concludes the paper and also identifies avenues to further research in this area.

2. Related Work

Current developments in information technology and communication offer new opportunities in the implementation of high quality healthcare systems (Nicolae et.al.2015). Healthcare covers detailed processes of the diagnosis, treatment and prevention of disease, injury and other physical and mental impairments in humans (Nicolae et.al.2015).The healthcare industry in most countries are evolving at a rapid pace. The healthcare industry can be regarded as a place with rich information as they generate massive amounts of data including electronic medical records, administrative reports and other benchmark findings (Wickramasinghe et.al.2004). These technologies ensure the provision for better medical data processing, diagnosis that are more accurate and provide easier access to medical services (Saripalle 2004). The concept of ontology is befitting the representation of entities in the medical field. Hence Ontology based approach is adopted for

better knowledge representation of diseases, its causes, symptoms and treatment.

An ontology based information extraction process has been presented in (Anantharangachar et.al.2013). Here, ontology is used to guide the information extraction process from unstructured text. The extracted information is then stored into RDF (Resource Description Framework) triple form. In the field of delimited text files, only the mapping-based approaches have been applied thus far. (Ermilov et. al. 2013) proposed an approach based on mapping from tabular data to RDF. The mapping files were automatically generated. In order to prevent errors, the mapping files were allowed to be edited by the users. Eventually, the mapping files were used to map corresponding tabular data to RDF.

Apart from this, mapping based techniques have been applied in other formats as well for converting legacy data into RDF. The authors have listed different approaches based on new ontology, existing ontology, database reverse engineering, and domain schema ontology. Likewise, the authors (Lin et. al. 2013) have used database reverse engineering method for ontology extraction. Dadjoo et. al., in 2015 proposed an approach for constructing ontologies from relational databases. The authors appeared to have generated ontological concepts by extracting knowledge from the relational schema and database components such as constraints and triggers.

A review of previous research in this area has revealed the fact that there currently exists no database for investigating Psoriasis related information. Moreover no ontological framework has been proposed for Psoriasis diagnosis. Hence this research attempted to discover novel findings in Psoriasis diagnosis through Ontological framework.

3. Material and Methods

Ontology framework helps in structuring the vast data on causes, symptoms, test and treatment of diseases. The proposed ontology not only covers medical processes and background aspects, but also medical diagnosis and management of disease to provide common support to clinical practice on controlling Psoriasis. The framework can be extended to provide support for other diseases too. The following paragraph gives

information on the computing environment employed for implementation.

The ontology construction is carried out using Protégé 4.1 (http://protege.stanford.edu/). Protégé is an open access and free ontology editor application. Protégé has evolved, in collaboration with the University of Manchester, to represent ontology in the OWL - based on description logics. Protégé allows the support of OWL and support for exporting data. Protégé ontology involve a variety of other formats such as RDF/S, OWL and XML Schema. Besides, it has 69 plug-ins, that include ontology visualization (OntoViz), ontology alignment (PROMPT) and interfaces with rule engines and formalisms such as SWRL (Semantic Web Rule Language) Web based visualization and navigation of disease using ontology browser library. Ontology Browser behaves as the web interface, which allows the user to browse OWL Ontology and RDF (Linked Open Data) files. Ontology Browser functions in an Apache Tomcat 5.5 Server environment.

The following section describes the proposed ontology framework for Psoriasis Risk Assessment Remedy (PRAR) database in detail.

4. The Proposed Ontology Framework for Psoriasis Risk Assessment Remedy (PRAR) Database

The proposed system incorporated two phases: Formulation of a novel database for Psoriasis Risk Assessment Remedy (PRAR) (ii) Articulation of a novel framework for Psoriasis detection through computational modeling and Ontology Construction. The overall architecture of the Proposed Methodology is depicted in Figure 1.
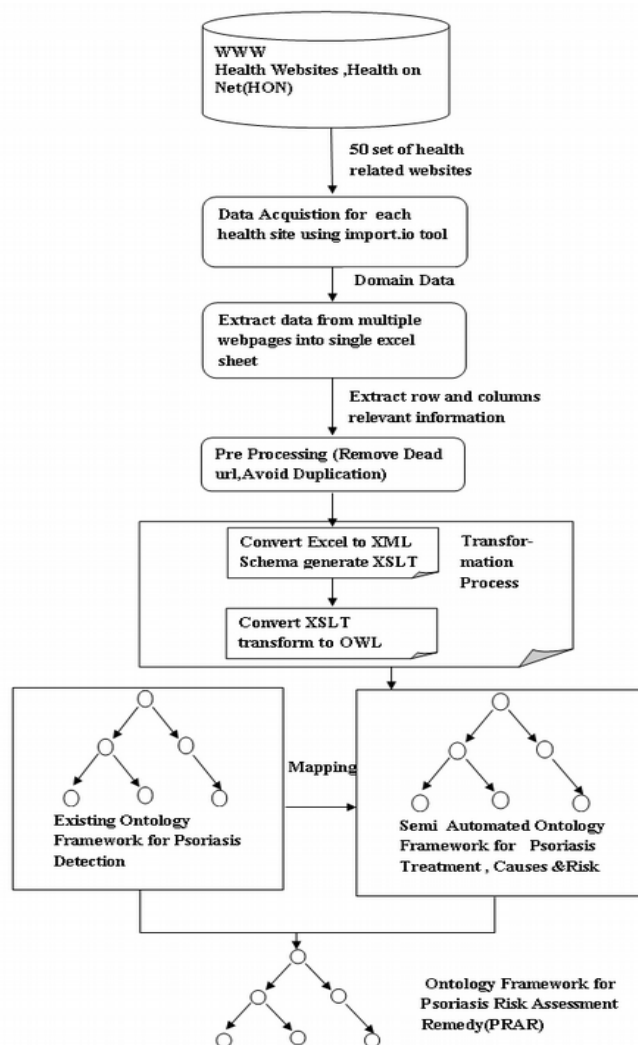
Figure 1. Architecture of the Proposed SOCF Methodology.

4.1 Data acquisition (health on net foundation)

The number of health information websites and online services is increasing day by day. It is known that the quality of these websites is very difficult to assess and evaluate (Mayer et al 2005). Its role is clearly to help users as quickly as possible towards access to required information instead of having to look and evaluate through all documents proposed by search engines. The Health On the Net Foundation (HON) is the leading organization in promoting and guiding the deployment of useful and reliable online medical-health information and was awarded the eEurope Award in 2004 in the eHealth category. HONcode is the most used ethical and trustworthy code for medical and health related information available on Internet (Boyer et al 1998). The data for this research was obtained from the database of certified websites as the quality of the available information is paramount, far more so than the quantity. A set of 50 health related websites were selected and the HON was manually assessed from that list.

4.2 Data acquisition of each health site using import.io tool

The same set of websites was crawled by using the import.io tool. It is a web based platform used to scrape data from any website (Import, https://www.import.io) and create a single searchable database containing information from several different sources. Figure 2 depicts the process of (i) Extracting rows and columns of data from a webpage into a data table (ii) Crawling a website and extracting data from multiple web pages into a single data table.

4.3 Data pre-processing

Data pre-processing is needed to remove noise i.e., unwanted characters or strings. In the data file, most of the attributes are quoted by the same string. Computation using such data generates erroneous results, which could stop the conversion process. The following pre-processing was incorporated to reduce such events. Figure 3 reveals the:

1. Removal of dead URLs.
2. Removal of duplication.
3. Removal of stopwords.

Figure 2. Extract data from multiple web pages into a single table

Figure 3 Data Pre-processing Tab

4.4 Transformation process

The transformation process from the relational database schema defined by tables into OWL ontologies was achieved by using XSLT. The basic generation process was based on XML schema. A relational database was represented by an XML document according to the XML schema. It was structured into rows and columns consisting of headers and records. Each entity in the column header was identified as Class Attribute and the description of the elements as sub classes. Considering the relational schema in Figure 4, the XML description of this relational schema was stored in the text file called schema.xml. This file was the input to the transformation process.

A set of manually defined mapping rules were used to generate an XSLT stylesheet to translate the XML schema into the domain ontology. This is a onetime translation when all XML documents share the same schema. A produced XSLT stylesheet can be used by any XSLT processor to automatically generate the desired ontology. The second step was to translate the source XML document to domain ontology. The elements in this document will be transformed in OWL using XSLT. The transformation relational schema to OWL was implemented as a set of XSLT stylesheets that take as input an XML description of relational schema.

4.4 Semi-automated ontology construction framework (SOCF)

The XML model is based on the labelled tree, where the meaning of the tag's nestings are interpreted by the program executed on it.

The OWL model is based on the subject-predicate-object structure from RDF/RDF-S (Resource Description Framework /Resource Description Framework Schema), where objects, their attributes and relations are naturally represented, and the semantics is specified (Decker et. al.2000, Mcguinness et.al.2004). Translation from XML to OWL is to interpret the tree structure of XML, and to represent the intended model in the subject-predicate-object structure.

Each and every attribute was assigned as either super-class/subclass. The hierarchical structure was framed with the superclass and subclass. In Psoriasis database, Semi Automated ontology construction from the extracted knowledge was represented in the relational database. The Semi Automated ontology construction framework contained seven classes that were extracted from XML and represented in OWL. Each and every class predicts the objects' structure through their attributes and the relations as represented in Figure 5. Reasoner is used to check the semantic consistency. The evaluation
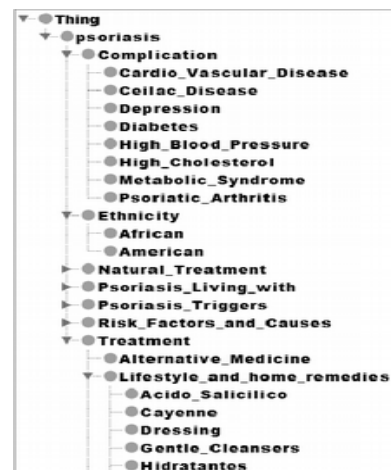

Figure 5. Semi Automated Ontology Construction Framework.

Figure 4. XML Schema for relational database

of the Psoriasis database ontology was done using the Reasoner pellet. The reasoner checks semantic consistency of the Psoriasis remedy database ontology and reports it as 'concepts that satisfy validity' or the 'measure of correctness in the concept hierarchy'. The reasoner asserts that the proposed ontology does not have any inconsistency (Hyosook et. al. 2012).

4.5 Existing ontology framework for psoriasis detection

The ontology construction of hierarchical structure consists of (Jian-xu chen.et.al.2010)

- a group of classes structured in a well-defined manner to represent the domain's salient classes,
- a group of slots related to the identified classes that describe their properties and relationships, and
- a cluster of instances of those classes.

In OWL, classes are interpreted as sets of sub classes. The hierarchical structure of Erythemato-squamous disease (from which Psoriasis is detected) is depicted in Figure 6. The next step was to construct the OWL structure generation that consisted of Category and Description. In Erythemato-squamous dataset, there were six classes, the category N_1 represented the Psoriasis class and it was inferred with the semi automated ontology framework for treatment, risk factor, causes and remedies.

It is concluded from this that once the ontology is constructed for a certain ailment, it is possible to detect that ailment from a set of other related ailments based on the attributes and their relationships. For eg. SOCF was constructed for Psoriasis. Erythemato-squamous data was input and from it the Psoriasis disease alone was inferred with all its associated risks and remedial measures.

4.6 Mapping two ontologies

The manually generated ontology for psoriasis detection comprising of structure generation from decision tree, rules generation and prediction were mapped with the proposed semi automated ontology construction framework (SOCF) as depicted in Figure 7. The Overall OWL Viz Mapping with existing ontology and PRAR is shown in Figure 8.
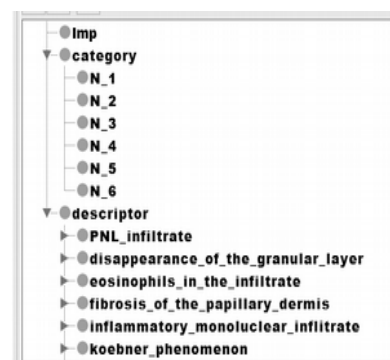
Figure 6. Hierarchy Structure of Erythemato-squamous Disease

SOCF (Semi Automated Ontology Construction Framework) classes were associated with the Category of N_1 as highlighted in Figure 8.
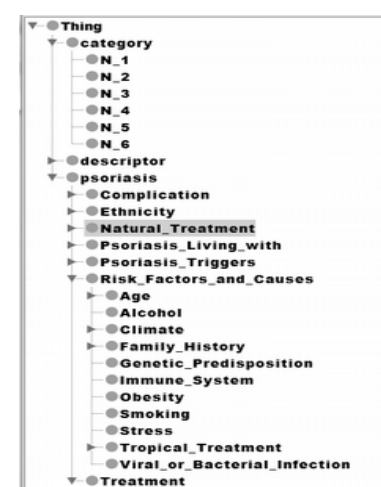
Figure 7. Ontology Framework for Psoriasis Risk Assessment Remedy (PRAR) Database

The ontoGraf shows the overall mapping output with test data as shown in Figure 9. The category attribute - Psoriasis, N_1 is represented by shaded ellipses. R1 and R2 are the individuals ie. the instances of the class Psoriasis. SOCF mapped the risk factors and remedies for Psoriasis detection accurately.

4.7 Experimental Results

The proposed Semi-Automated Ontology Construction framework (SOCF) for Psoriasis Detection was implemented on the Erythemato-squamous dataset downloaded from the UCI repository (UCI, http://archive.ics.uci.edu/ml /datasets). It is a benchmark dataset used for machine learning problems. The proposed SOCF approach was also applied on the Psoriasis categories. Some pre-processing techniques to delete unnecessary information such as dead URL and duplicated values were

Figure 8. The Overall OWL Viz Mapping with Existing ontology and Psoriasis Risk Assesment and Remedy

Figure 9. OntoGraf Visualization of SOCF with PRAR

done to get accurate data in order to obtain precise results. For the realization of this transformation, a suitable XML schema was developed. The 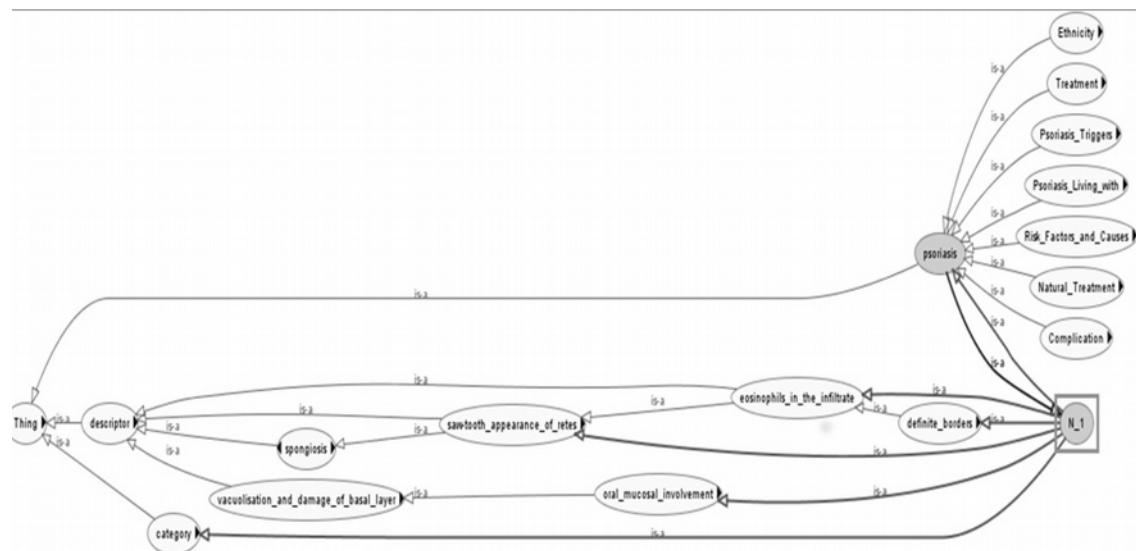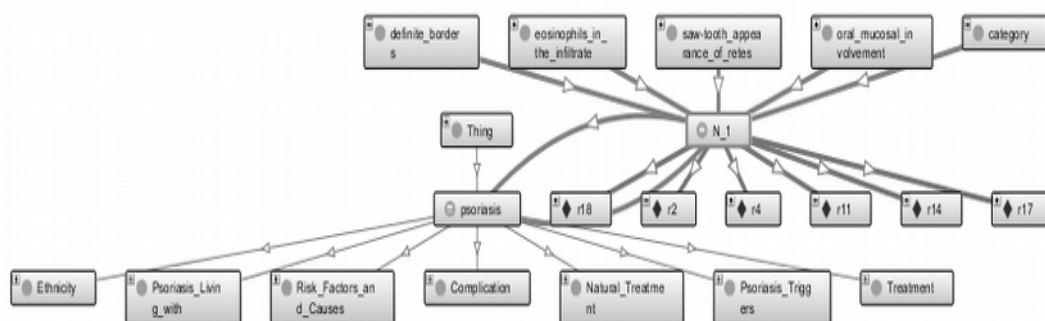process was implemented by executing the XSLT transformation over the XML instances document that in turn yielded the OWL output document containing the OWL classes and their properties. The proposed methodology was tested on 112 samples from the authenticated UCI Machine Learning Repository. The data contained 112 records, with each record having 34 features of which 12 were clinical and 22 were histopathological features. The proposed ontology mapped the risk factors and remedies for Psoriasis detection with 98.7% accurately.

5. Conclusion

Ontologies in medicine mainly concentrate on knowledge representation and re-organization of information viz. diseases, causes, symptoms, tests and treatment. The proposed work formulated an ontological framework for

Psoriasis Risk and Remedy database. This paper targeted the need to generate a Psoriasis Remedy Database and automatically infer the relations between the Symptoms, Causes and Treatment through Semi-Automated Ontology Reasoning and Inference.

Medical practitioners when dealing with Psoriasis disease can make use of the structured information available in the proposed ontological framework. The proposed methodology was tested on 112 samples from the authenticated UCI Machine Learning Repository. The ontology developed using the proposed SOCF mapped the risk factors and remedies for Psoriasis detection with 98.7% accuracy, this being reported for the first time.

REFERENCES

1. ANANTHARANGACHAR, R., S. RAMANI, S. RAJAGOPALAN, **Ontology Guided Information Extraction from Unstructured Text,** International Journal

of Web & Semantic Technology (IJWesT) vol. 4(1), January 2013, pp. 1302-1335.

2. BOYER, C., M. SELBY, J. R. SCHERRER, R. D. APPEL, **The Health on the Net Code of Conduct for Medical and Health Websites**. Computers in Biology and Medicine, vol. 28(5), 1998, pp. 603-610.

3. CAPON, F., R. C. TREMBATH, J. N. BARKER, **An Update on the Genetics of Psoriasis,** Dermatologic clinics, vol. 22(4), 2004, pp. 339-47.

4. DADJOO, M., E. KHEIRKHAH, **An Approach for Transforming of Relational Databases to OWL Ontology**, International Journal of Web & Semantic Technology (IJWesT), Vol. 6(1), January 2015.

5. DECKER, S., S. MELNIK, F. VAN HARMELEN, D. FENSEL, M. KLEIN, J. BROEKSTRA, I. HORROCKS, **The Semantic Web: The Roles of XML and RDF,** Internet Computing, IEEE, vol. 4(5), 2000, pp. 63-73.

6. ERMILOV, I., S. AUER, C. STADLER, **Csv2rdf: User-driven CSV to RDF Mass Conversion Framework,** Proceedings of the ISEM vol. 13 , 2013, pp. 04-06.

7. GUDJONSSON, J. E., J. T. ELDER, **Psoriasis: Epidemiology,** Clinics in Dermatology, vol. 25(6), 2007, pp. 535-546.

8. HADZIC, M., E. CHANG, **Ontology-based Multiagent Systems Support Human Disease Study and Control**, SOAS Frontiers in AI and Applications, vol. 135, 2005, pp.129-141.

9. JUNG, H., S. YOO, S. PARK, **Context Modelling Using Semantic Web Technologies**, Studies in Informatics and Control, ISSN 1220-1766, vol. 21(2), 2012, pp. 173-180.

10. CHEN, J.-X., S. L. SU, C. H. CHANG, J.-X. CHEN, S.-L. SU, C.-H. CHANG, **Diabetes Care Decision Support System,** Industrial and Information Systems (IIS), 2nd International Conference on. vol. (1). IEEE, 2010, pp. 323-326.

11. LANGLEY, R. G., G. G. KRUEGER, C. E. GRIFFITHS, **Psoriasis: Epidemiology, Clinical Features, and Quality of Life,** Annals of the Rheumatic Diseases, vol. 64(2), 2005, pp. ii18-ii23.

12. LIN, L., Z. XU, Y. DING, **OWL Ontology Extraction from Relational Databases via Database Reverse Engineering**, Journal of Software vol. 8, issue 11, 2013, pp. 2749-2760.

13. MAYER, M. A., A. LEIS, R. SARRIAS, P. RUÍZ, **Web Mèdica Acreditada Guidelines: Reliability and Quality of Health Information on Spanish-Language Websites,** Connecting Medical Informatics and Bioinformatics. Proc of MIE, 2005, pp. 1287-92.

14. MCGUINNESS, D. L., F. VAN HARMELEN, **OWL Web Ontology Language Overview**, W3C recommendation, vol. 10(10), 2004.

15. TODOR, N., M. BORZAN, L. E. CIULEANU, D. I. IANCU, V. A. GHENGHEA, T. E. CIULEANU, **Development and Evaluation of a Software-based Clinical Pharmacography System**, Studies in Informatics and Control, ISSN 1220-1766, vol. 24(4), 2015, pp. 427-438.

16. Open source software import.io: https://www.import.io.

17. Protégé , the Stanford Center for Biomedical Informatics Research at the Stanford University School of Medicine, http://protege.stanford.edu/

18. SARIPALLE, R., **Current Status of Ontologies in Biomedical and Clinical Informatics**, Univ. of Connecticut, 2004.

19. University of California, Irvine (UCI), Machine Learning Repository, http://archive.ics.uci.edu/ml/datasets

20. WICKRAMASINGHE, N, S. K. SHARMA, J. N. D. GUPTA, **Knowledge Management in Healthcare**, In Wickramasinghe N, Gupta JND, Sharma SK, ed., Creating knowledge-based healthcare organizations. Hershey, PA, USA: Idea Group Publishing; 2004, pp. 1-13.