The Nucleus **51**, No. 4 (2014) 434-438

www.thenucleuspak.org.pk

The Nucleus

ISSN 0029-5698 (Print)

ISSN 2306-6539 (Online)

Semantic Webs and Ontology-driven Biological Data Integration Methods for Bacteria

M.U.G. Khan, M. Idrees*, U. Hayat and A. Akram

Department of Computer Science & Engineering,University of Engineering & Technology, Lahore, Pakistan

ARTICLE INFO

Article history:

Received : 13 August, 2014

Revised : 22 January, 2015

Accepted : 12 February, 2015

Keywords :

Semantic Web

semantic querying

knowledge management

OWL

ABSTRACT

Integrating data from distant heterogeneous data sources is a big challenge for research community. By putting semantic web into practice, a new framework is being presented for knowledge engineering, querying and information sharing. In present paper, we presented an ontology based effort for the development of a knowledge that would allow for semantically querying bacteria knowledge base.

1. Introduction

It is a big challenge to combine varied and widely spread data in such a way that its consistency is not confronted because it is labeled another way in diverse resources, when there is biological relevancy for the same entity. For example, how much that entity is associated to any specific disease or involved in a certain process. There are many identical terms, acronyms and abbreviations in biomedicine language that refers to same entity, process or concept. Like glucose creating process is referred by many similar terms including 'glucose synthesis', 'glucose biosynthesis', 'glucose formation', 'glucose anabolism' and 'gluconeogenesis'. By using ontology a distinct identifier can be provided to describe information for a single entity and allows storing alternative names through using apposite metadata. Thus ontology is a skillful expression to biomedical entities, also synonyms, acronyms and abbreviations can be enhanced with ontologies. Further, ontologies facilitate the community to incorporate varied resources by allowing to reliablyidentifying an entity or group of entities centered on biological significance [1].

Although information is publicly available via World Wide Web, it is also put in database storage as well as described in publications but it is not capable to find information when a certain set of properties is provided to search engines. Reason is that biological information on web is not in machine understandable representation, as computers cannot sense words and interpret sentences to correctly propose the entities and their relevance to other entities that are stated in sentences. Semantic web is designed with key goal to add semantics to existing web, and this is achieved by designing ontologies that describes objects and relate them in a machine understandable way by using formal logic based representation. This constant effort is crucial in life sciences to expedite data integration from varied resources and semantic querying on knowledge bases [2].

Ontologies are already playing a vital role in dealing with medical entities, in discovery of grid [3] and semantic web. There is a shared portal of ontologies for biological resources which is called Open Biomedical Ontologies (OBO) and it includes popular Gene Ontology (GO) [4]. OBO controlled vocabularies and taxonomies are used for annotating biological resources by providing standardized vocabulary that helps making computer interpreted information available. OBO Foundry aided to redesign OBO ontologies to be mapped to Basic Formal Ontology (BFO) [5], an ontology to provide difference between entities and processes and basic relations can be used to link them [6]. Collectively they must offer a potent platform for describing and annotating knowledge for a specific domain and expose the opportunity to make varied granularity level queries that allow retrieving details from diverse resources. For example, there is a biochemical ontology that might state they enzymes are actually proteins that catalyze reactions. This information can support to query a knowledge base with terms proteins to fetch all data that is annotated as protein that catalyzes a reaction or as an enzyme. Even with efforts of OBO Foundry, OBO ontologies cannot be used that way as they lack obvious logical descriptions that are required to define membership of classes with respect to their properties. Like above example of protein cannot be queried from database using OBO ontology.

* Corresponding author : midrees10@gmail.com

Web Ontology Language (OWL) is accepted as an official knowledge representation recommendation for semantic web ontologies. A variant of it is OWL-DL, based on description logics (DL) that allows adding description for complex concepts by using simpler ones from first order logic reasoning. Reasoner, is a computer program that executes these reasoning tasks like ontology consistency. By using semantic web best practices incorporated with OWL-DL design, reasoning can be achieved for biological ontologies. Some already built biomedical ontologies are Foundational Model of Anatomy [7, 8] for identifying inconsistencies in ontologies, FungalWeb [9] for providing reasoning of important enzymes for yeast, BioPax OWL ontology [10] for pathway representation, are worth noting.

Cross-referencing is the key challenge in bioinformatics for identifying extensive biological entities that refer to same entity. To keep track of these identifiers is a gigantic problem and it becomes obligatory to build databases of database identifiers [11, 12]. LinkHub [13] offers an entry point for navigating these bewildering set of identifiers, whereas YeastHub [14] is a RDF based dataware house that allows an individual to add data as well as create queries amid of resources. Although they are flexible but they lack ontological knowledge, that result in indication for the equivalence of resources that are contributed by various users consequently prohibiting automatic discovery of semantically equivalent knowledge.

2. Related Work

Gathering data and analyzing it is utmost research phase in bio medical domain. Large volume of data is made available through wide variety of experimental methods in life sciences. This data is usually published and made available publicly on different share media with major contribution on World Wide Web. Although this data is extremely useful but it lacks competency, that leads to challenge of its integration and prospect research. This integration is required to allow biologists and researchers to understand widespread biological knowledge or distinguish any particular specie. As most of data sources are collected and maintained independently, so they are highly heterogeneous posing integration challenges [15]. Data integration methodology is usually categorized into two types:

1. Dataware house approach
2. Federated approach

Both techniques have certain limitations with performance overheads, heterogeneity and syntactic issues are worth noting. Semantic web data integration has emerged to overcome limitations of these methods.

Until now, semantic web, proposed by Tim Berners Lee in 2001, is already matured and broadly accepted in biological research. Semantic web uses Resource Description Framework (RDF) standards that describe resources as triplets consisting of a subject, a predicate and an object. For example, (Book, name, Bioinformatics Concepts) describes a resource *Book* whose *name* is *Bioinformatics Concepts.* In the meantime, OWL is further expressive than RDF by additionally enabling reasoning and inference in a domain of interest [16].

Clement Jonquet et al. [17] conducted a research for purpose of building a biological ontology recommender web service, aimed at facilitating data integration by use of ontologies for data annotation. With the challenge to figure out best suitable annotation for specific datasets, it used word-based documented metadata in a domain and suggested ontologies that were most suitable for annotating data. Ontologies were decided on base of three criteria naming coverage (most terms covering input text), connectivity (ontologies mapped to other ontologies) and size (number of concepts). Scores are then assigned to ontologies based on these. Sabou et al. [18] described deficiencies in currently available ontologies as, usually relationship amongst concepts is ignored and their meaning is ignored typically.

Simon Jupp et al. [19] presented an ontology based solution for chronic renal disease extended worldwide. By prompt identification of disease indications, early diagnosis is made possible that leads to quick underlying pathology. Major challenge faced to achieve this goal was the integration of tentative results collected from different resources. An approach was presented for developing a knowledge base that integrated heterogeneously gathered data on kidney and urine experiments to value benefits promised by semantic web for data integration. KUP was used as specialized ontology for the domain, whereas external database knowledge is integrated through use of RDF and SPARQL being used as querying language. This organization enabled to query across many proteins stated in urine, cumulative knowledge and convert them back into the genes' context stated in kidneys.

Stefan Schulz et al. introduced the concept of taxonomy in biological classification of entities into species. They discussed different techniques on representing biological taxa's by using standards of biomedical ontologies which are currently available, including OWL Description Logic (OWL DL) and Open Biomedical Ontologies Relation Ontology. Ambiguities related to species concept were overcome using approach to predict taxon data as biological organism's merits and an integrative ontology architecture. BioTop is the implementation of this approach [20].

Many annotations are being missed because of absent cross-references, said by Leon French et al in a research conducted for biomedical data integration. Wide set of sequence datasets are publicly available on internet with, awkwardly, lacking description and parameters. It is usually represented as natural language text which is the major obstacle in its searching, enquiring high throughput and data integration [21]. Lexical properties from Unified Medical Language System (UMLS) were used to extract concepts from text, which were in turn linked to classes in biomedical ontologies. Gene expression was made reachable by annotating it semantically with 89% precision achieved [22].

Acquisition of knowledge poses many challenges on information extraction from heterogeneous data sets that are distributed over and operated autonomously [23]. Various approaches and algorithms are devised to subdue these challenges through use of computer aided discoveries in respective domains. They use ontology based techniques that are customized relative to data that is to be integrated, in the interim also keeping in account agent based implementations in computational biology relationships [24].

More research was conducted by Robert Hoehndorf et al. on various ontological techniques done on bio-medical data. As ontologies are standardized ways for accessing any domain's knowledge to verify it for data inconsistencies and provide support in integration of data from heterogeneous sources. To achieve this goal, theories and techniques from many disciplines are taken into account for defining ontologies. These disciplines may include information management, knowledge representation, cognitive science, linguistics and philosophy, which require different strategies to be followed. These strategies depend on domain and success factor of ontology definition that provides quantifiable comparison of results [25].

Yi Liu et al. highlighted the problem of demanded profiling in research activity that are majorly of interest to funding agencies and researchers. Due to non-existent global classification of ontologies, systematic profiling is not achieved for various research activities. Yi Liu et al. explored research activities conducted by various funding agencies together with their domain of research and introduced annotation based on ontology to cope with this problem by grouping closely related domains based on priorities. They used disease ontology (DO) as their annotation technique to emanate profiling [26].

Moreover, National Centre for Biomedical Ontology (NCBO) was established a decade ago with goals to develop and maintain repository for biomedical ontologies, provide tools for creating ontologies and define web services for their use in bio-medical research. They also train the community in ontology driven approaches and define ways to follow best practices as well as inter-group collaborations. BioPortal is the central resource of NCBO which covers more than 270 ontologies worldwide and supports numerous web services to enable researchers to use these ontologies for annotation, data retrieval and lexical analysis on biomedical data [27].

SOBA, a sub component built on SmartWeb, was discussed in a paper by Paul Buitelaar et al. It is an ontology based component that collected and extracted information from a number of soccer web pages to build an automatic knowledge base that in turn can be used to search for textual information and questions for a specific domain. SOBA take in a tight connection between ontology, knowledge base and the information extracting module. It extracts data from heterogeneous sources like tabular structures, textual data and image captions in a semantically integrated way. It works by storing information that is extracted, in to a knowledge base which in turn is used for making linkages between newly discovered and existing entities [28].

There is an open source tool**.** Semantic web information Management with automated Reasoning Tool (SMART) which aims at providing a learning tool for life scientists to represent, integrating and querying heterogeneous bio-medical data spanned over distributed setting [29]. It uses AJAX, SVG and JSF technologies, RDF, OWL, SPARQL semantic web languages, triple stores (i.e. Jena) and DL reasoners (i.e. Pellet) for the automated reasoning. It provides features like composition of semantic queries which use DL reasoners, a GUI for providing query, mapping from DL to SPARQL and finally retrieval of inferences from RDF triple store. Enhancement in this SMART will empower more sophisticated information retrieval prospects in life sciences [30].

3. Methods

Proposed ontology based system is comprised majorly of three components, the ontology design, mapping of data and last is query answering.

3.1 Ontology Design

The ontology was designed using the best practices of semantic web by using OWL-DL. The goal was to cover all the entities of our data source, which includes functional and structural information, cross referencing of databases, biological processes, their interactions and literature references etc. These ideas were then mapped to OWL-DL to deduce objects, processes, roles and chronological regions [31, 32].

Basic entity relationships were unified in OWL-DL based Basic Relation Ontology (BRO), which is the organized hierarchically from “isRelatedTo” that is an elementary relation for presenting relation to groups of objects, partial objects and chronological relations. The domain of BRO was further mapped to notions of Basic Formal Ontology (BFO) subsequently comprising higher level ontology NULO, which is available at http://ontology.dumontierlab.com/nulo. NULO provides relations to BFO in a semantic manner.

To begin with, the classes were extracted using attributes from flat files then they were enhanced and improved to imitate knowledge of genome structure. For example, we had file containing experimental data about bacteria containing attributes bid, bname, datasource. From this file, we created classes: i) The Bacteria class and subclasses defining sub-types. ii) The DataSource class, containing the reference to original source. iii) The class Integration was added as subclass of BFO class for representing bid. Finally, definitions of classes were acquired from WordNet glossary and added to class via "comment" annotation property.

For the description of domain specific relations, object properties were defined. The "hasReference" relation or "isReferencedIn" relation provides kith and kin between Integration and its referenced entity. Next is the "hasSource" or "isSourceOf" relation that links entity to its original source and serves as a mean for data derivation. "HasStatus" or "isStatusOf" is a quality relation that describes status i.e., verified, uncharacterized or verified etc., for current frame. To end with, "hasOutcome" or "isOutcomeOf" relation describes the relation between entity and its outcome (like phenotype).

Numerous data properties were presented for accommodating information that is central to any particular entity. For example, date versioning, in which several values were to be maintained. Many features of chromosomes were related with a start and end co-ordinate that demarcate incessant area positioned on chromosome strand. One other property provides relationship between citation and a publication using "hasCitation" relation.

The ontology founded its grounds from NULO and suitable data classification and sufficient conditions for knowledge discovery were added to ontology to ensure its correctness and comprehension.

3.2 Data Mapping

It was quite a challenge to process source data, assuming that it was obtained from heterogeneous data sources with varied schemas and formats. We came through wide formats including Text, FASTA, NCBI, XML, and Excel files etc. Normalization and pruning of certain complex files was mandatory to bring up to a standard format. Another challenge was that many files were missing identifier (i.e., bid in our case) instead containing cross references. A parser was suggested to import and extract required information from data using attributes as keys.

3.3 Query Answering

With the advent of reasoning enabled applications and design of ontologies, the researchers are encouraged to mine particular information as well as allow discovering novel relations about their interest area. We goal at illustrating how a scientist can retrieve useful information of interest from our ontology driven knowledge base, from query posting to query granularity at various levels and among data sources with different identifiers.

4. Results and Discussion

Saccharomyces Genome Database (SGD) has assigned a unique identifier for each of its chromosomal feature. Other identifiers like gene names, aliases, their ORF names and cross references for all databases were assigned a namespace. Identifiers pointing to similar resources were made same by affirming the "sameAs" relation of OWL to SGD resources. OWL-DL reasoner will speculate these cross-references and will automatically resolve the problem of importing the data. A unique identifier is assigned to imported data in namespace to maintain relevance as well for linkage of resources using "has Object" relationship which in turn facilitates for querying data either by using object properties or namespace filtering.

Major step for design of ontology was the extraction of classes representing entities and determining relationships amid them. Moving forward class hierarchy was further improved and refined via domain concepts and integration was done by way of BFO ontology. This methodology can comprehend with the automation of extracting information and data, and extraction of relations from prototype of data sources. Data wrappers for our domain of interest are a leading step towards domain independent wrappers for population of ontologies through heterogeneous data sources.

There can be an overlap of ontologies in subsets of community driven OBO because they lack OWL semantics. Data integration from different sources requires that their individual statements to be considered equivalent. This can be achieved using "sameAs" OWL property that makes it equivalent to SGD. This allows the user to query the database via any equivalent identifiers returning the union of statements in result.

User friendly interface is constructed for facilitation of user to provide queries. Storing and retrieving ontological data is the foremost challenge. It is a sophisticated task to store voluminous databases along with their inferences in memory without complex underlying hardware. We will clearly require more sophisticated solutions.

5. Conclusion

In this paper we presented an approach for describing, integrating and querying biological data for bacteria using OWL-DL.We started with extension of BFO ontologies and incorporated the design of domain explicit ontology. We used basic relations for objects which helped for integration of identical resources from various data sources by using semantics of RDF and OWL-DL. This was the evolution of ontology based semantic framework for querying bacteria knowledge base. However, several challenges still remain to be become realization by providing the automation of ontology population, efficient storages and instinctual interfaces for users.

References

[1] D. Rubin, N. Shah and F. Natalya, Biomedical Ontologies: AFunctional Perspective, Briefings in Bioinformatics, **9** (2008) 75-90.

[2] C. Jonquet, M. A.Musen, and N. H. Shah, Building a biomedical ontology recommender web service, J Biomed Semantics **1** (2010) 1-18.

[3] G. Babnigg, C. Giometti, A Database of Unique Protein Sequence Identifiers for Proteome Studies**.** J.Proteomics **6 (**2006) 4514-4522.

[4] A. Silvescu, J. Reinoso, and V. Honavar, Ontology-Driven Information Extraction and Knowledge Acquisition from Heterogeneous, Distributed, Autonomous Biological Data, 1-10, Report submitted to Department of Computer Science 226 Atanasoff Hall, Iowa State University Ames, Iowa 50011-1040 Menlo Park, California. Simon, and E. Tomita (Eds.): ALT 2005, LNAI 3734, Springer-Verlag Berlin Heidelberg (2005) pp. 13–44.

[5] Y. Liu, A. Coulet, P. Pendu, and H. Shah, Using ontology-based annotation to profile disease research; Stanford Centre for Biomedical Informatics, 1265 Welch Road, Medical School Office Building, Room X215 Mail Code 5479, Stanford, CA 94305, USA (2012) doi:10.1186/2041-1480-4-S1-S8.

[6] A. Musen, F.Noy and H. Shah, Division of Biomedical Statistics and Informatics, Mayo Clinic, Rochester, Minnesota, USA, Department of Computer Science, University of Victoria, Canada, (2011) doi: 10.1136/amiajnl-2011-000631.

[7] A.L. Battista, N.V. Rosales, M.P. Michel and D. Umontier, SMART: A Web Based, Ontology-Driven, Semantic Web Query Answering Application, School of Computer Science, Department of Biology, Carleton University, Ottawa, Ontario, Canada. (2008) Technical Report: FBI-HH-B-262/05.

[8] E. Sirin, B. Parsia, B.C. Grau, A. Kalyanpur, and Y. Katz, Pellet: A Practical Owl-dl Reasoner**,** 3rd International Semantic Web Conference (ISWC2004) 2004) pp. 21-29.

[9] G. Heja, P. Varga, P. Pallinger and G. Surjan, Restructuring the Foundational Model of Anatomy**,** Stud Health Technol. Inform **124** (2006) 755-760.

[10] V. Haarslev, R. Möller and M. Wessel M, Querying the Semantic Web with Racer + nRQLin KI-04, Workshop on Applications on Description Logics (Ulm, 2004) doi:biomedcentral.com/1471-2105/8/S3/S4.

[11] I. Foster,C. Kesselman, J. Nick and S. Tuecke, The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration in Globus Project(2002.) doi = 10.1.1.14.8105

[12] J. Luciano, PAX of Mind for Pathway Researchers**.** J. Drug Discov. Today **10** (2005) 937-942.

[13] J. Christopher, S. Arash, S. Xiao, H. Volker and B. Greg, Semantic Web Infrastructure for Fungal Enzyme Biotechnologists**.** J. Web Semantic **4** (2006) 168-180.

[14] K. Wolstencroft, P. Lord, L. Tabernero, A. Brass, and R. Stevens, Protein Classification using Ontology Classification, Bioinformatics **22** (2006) e530-538.

[15] LR. Alan, Modularization of Domain Ontologies Implemented in Description logics and Related Formalisms including OWL,KC2003 (Sanibel Island, FL, USA, 2003), ACM Press.doi:ceur-ws.org/Vol-315/paper3.pdf

[16] U. Consortium, The Gene Ontology (GO) Project in 2006**.** Nucleic Acids Res **34** (2006) D322-326.

[17] L. French, S. Lane, T. Law, L. Xu and Paul Pavlidis, Application and Evaluation of Automated Semantic Annotation of Gene Expression Experiments. Bioinformatics Graduate Program, Department of Psychiatry and Centre for High-Throughput Biology, University of British Columbia, Vancouver, BC, Canada (2009) doi: 10.1093/bioinformatics/btp259

[18] M. Idrees and M. U. G. Khan, SMGCD: Metrics for Biological Sequence data**,** The Nucleus **51** (2014) 125-131.

[19] LinkHub. http://hub.gerteinlab.org.

[20] M.A. Remli, S. Deris and A. Abdullah, Ontology-based Semantic Data Integration for Lactococcus Lactis (L.lactis): Connecting Cross-References Heterogeneous Data Sources, 4th International Conference on Bioinformatics and Biomedical Technology IPCBEE **29** (2012) pp.100-104.

[21] P. Grenon, B. Smith and L. Goldberg, Biodynamic Ontology: Applying BFO in the Biomedical Domain**.** Stud Health Technol. Inform **102** (2004) 20-38.

[22] M. Sabou, V. Lopez, E. Motta and V. Uren, Ontology Selection: Ontology Evaluation on the Real Semantic Web. 4th International EON Workshop, Evaluation of Ontologies for the Web, EON'06, Proceedings of CEUR Workshop **179** (2010) doi:10.1186/2041-1480-1-S1-S1

[23] M. Aranguren, Ontology Design Patterns for the Formalization of Biological Ontologies, A Report submitted to the University of Manchester, Degree of Philosophy, Faculty of Engineering and Physical Sciences and Biological Sciences (2005) 1-67dissertation.

[24] R. Hoehndorf, M. Dumontier and G. Gkoutos, Evaluation of Research in Biomedical Ontologies, Briefings in Bioinformatics (2012) 1-17.

[25] N. Villanueva-Rosales, K. Osbahr and M. Dumontier, Towards a Semantic Knowledge Base for Yeast Biologists, Copyright is held by the Author/ Owner(s) May 8-12 2007, Banff, Canada. Dissertation.

[26] P. Buitelaar, P. Cimiano, S. Racioppa and M. Siegel, Ontology-based Information Extraction with SOBA, Saarbrücken, Kalrsruhe, Germany (2006), Medical Information Science Reference, IGI Global Data Mediation and Extraction at the 7th International Semantic Web Conference, Karlsruhe, Germany.

[27] T. Hussain and S. Asghar, Evaluation of Similarity Measuresfor Categoricaldata; The Nucleus **50** (2013) 387-394.

[28] S. Zhang, O. Bodenreider and C. Golbreich, Experience in Reasoning with the Foundational Model of Anatomy in OWL DL**.** Pac Symp Biocomput. (2006) pp. 200-211.

[29] SS. Dwight, M. Harris, K. Dolinski and C.A. Ball, J. Nucleic Acids Res. **30** (2002) 69-72.

[30] S. Jupp, J. Klein, J. Schanstra and R. Stevens, Developing a Kidney and Urinary Pathway Knowledge Base. J Biomed Semantics semantic Applications in Life Sciences **2** (Suppl 2) 9-10 July (2010), Boston, MA, USA.

[31] S. Schulz, H. Stenzhorn and M. Boeker, Report: The ontology of Biological Taxa. Institute of Medical Biometry and Medical Informatics, National University of Ireland, Galway, University Road, Galway, Ireland (2008) HG004028.

[32] U. Consortium, J. Nucleic Acids Res. **40** (2012) D71-D75.